SYSTEMS MODELING: APPROACHES AND APPLICATIONS - VOLUME II

EDITED BY: Alberto Jesus Martin, Ernesto Perez-Rueda and Daniel Garrido
PUBLISHED IN: Frontiers in Molecular Biosciences



frontiers Research Topics



frontiers

Frontiers eBook Copyright Statement

The copyright in the text of individual articles in this eBook is the property of their respective authors or their respective institutions or funders. The copyright in graphics and images within each article may be subject to copyright of other parties. In both cases this is subject to a license granted to Frontiers.

The compilation of articles constituting this eBook is the property of Frontiers.

Each article within this eBook, and the eBook itself, are published under the most recent version of the Creative Commons CC-BY licence. The version current at the date of publication of this eBook is CC-BY 4.0. If the CC-BY licence is updated, the licence granted by Frontiers is automatically updated to the new version.

When exercising any right under the CC-BY licence, Frontiers must be attributed as the original publisher of the article or eBook, as applicable.

Authors have the responsibility of ensuring that any graphics or other materials which are the property of others may be included in the CC-BY licence, but this should be checked before relying on the CC-BY licence to reproduce those materials. Any copyright notices relating to those materials must be complied with.

Copyright and source acknowledgement notices may not be removed and must be displayed in any copy, derivative work or partial copy which includes the elements in question.

All copyright, and all rights therein, are protected by national and international copyright laws. The above represents a summary only. For further information please read Frontiers' Conditions for Website Use and Copyright Statement, and the applicable CC-BY licence.

ISSN 1664-8714
ISBN 978-2-83250-712-4
DOI 10.3389/978-2-83250-712-4

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

SYSTEMS MODELING: APPROACHES AND APPLICATIONS - VOLUME II

Topic Editors:
Alberto Jesus Martin, Fundación Ciencia and Vida, Chile
Ernesto Perez-Rueda, Universidad Nacional Autónoma de México, Mexico
Daniel Garrido, Pontificia Universidad Católica de Chile, Chile

Citation: Martin, A. J., Perez-Rueda, E., Garrido, D., eds. (2022). Systems Modeling: Approaches and Applications - Volume II.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-712-4

Table of Contents

05 ***Editorial: Systems Modeling: Approaches and Applications–Volume II***
Daniel Garrido, Alberto J. M. Martín and Ernesto Pérez-Rueda

08 ***Understanding Metabolic Flux Behaviour in Whole-Cell Model Output***
Sophie Landon, Oliver Chalkley, Gus Breese, Claire Grierson and Lucia Marucci

21 ***A Novel Hybrid Logic-ODE Modeling Approach to Overcome Knowledge Gaps***
Gianluca Selvaggio, Serena Cristellon and Luca Marchetti

31 ***Modeling Progression of Single Cell Populations Through the Cell Cycle as a Sequence of Switches***
Andrei Zinovyev, Michail Sadovsky, Laurence Calzone, Aziz Fouché, Clarice S. Groeneveld, Alexander Chervov, Emmanuel Barillot and Alexander N. Gorban

50 ***Dynamic Control Balancing Cell Proliferation and Inflammation is Crucial for an Effective Immune Response to Malaria***
Anuj Gupta, Mary R. Galinski and Eberhard O. Voit

67 ***Three-Dimensional Model of Sub-Plasmalemmal Ca^{2+} Microdomains Evoked by T Cell Receptor/CD3 Complex Stimulation***
Diana Gil, Björn-Philipp Diercks, Andreas H. Guse and Geneviève Dupont

84 ***Optimal Experimental Design Based on Two-Dimensional Likelihood Profiles***
Tim Litwin, Jens Timmer and Clemens Kreutz

98 ***UPMaBoSS: A Novel Framework for Dynamic Cell Population Modeling***
Gautier Stoll, Aurélien Naldi, Vincent Noël, Eric Viara, Emmanuel Barillot, Guido Kroemer, Denis Thieffry and Laurence Calzone

109 ***Applications of Coarse-Grained Models in Metabolic Engineering***
Dieu Thi Doan, Manh Dat Hoang, Anna-Lena Heins and Andreas Kremling

123 ***GAL Regulon in the Yeast* S. cerevisiae *is Highly Evolvable* via *Acquisition in the Coding Regions of the Regulatory Elements of the Network***
Rajeshkannan, Anjali Mahilkar and Supreet Saini

137 ***Modeling and Optimization of a Molecular Biocontroller for the Regulation of Complex Metabolic Pathways***
Yadira Boada, Fernando N. Santos-Navarro, Jesús Picó and Alejandro Vignoni

151 ***Optimizing Dosage-Specific Treatments in a Multi-Scale Model of a Tumor Growth***
Miguel Ponce-de-Leon, Arnau Montagud, Charilaos Akasiadis, Janina Schreiber, Thaleia Ntiniakou and Alfonso Valencia

164 ***Cyclic Attractors Are Critical for Macrophage Differentiation, Heterogeneity, and Plasticity***
Manuel Azaid Ordaz-Arias, Laura Díaz-Alvarez, Joaquín Zúñiga, Mariana Esther Martinez-Sánchez and Yalbi Itzel Balderas-Martínez

176 ***Combination of Genome-Scale Models and Bioreactor Dynamics to Optimize the Production of Commodity Chemicals***
Jorge Lázaro, Giorgio Jansen, Yiheng Yang, Mario A. Torres-Acosta, Gary Lye, Stephen G. Oliver and Jorge Júlvez

190 ***Generalized Structural Kinetic Modeling: A Survey and Guide***
Jana C. Massing and Thilo Gross

209 ***Discrete Biochemical Systems Theory***
Eberhard O. Voit and Daniel V. Olivença

223 ***Predicting Metabolic Adaptation Under Dynamic Substrate Conditions Using a Resource-Dependent Kinetic Model: A Case Study Using* Saccharomyces cerevisiae**
K. J. A. Verhagen, S. A. Eerden, B. J. Sikkema and S. A. Wahl

235 ***Modeling Calcium Signaling in* S. cerevisiae *Highlights the Role and Regulation of the Calmodulin-Calcineurin Pathway in Response to Hypotonic Shock***
Simone Spolaor, Mattia Rovetta, Marco S. Nobile, Paolo Cazzaniga, Renata Tisi and Daniela Besozzi

254 ***An Epistatic Network Describes* oppA *and* glgB *as Relevant Genes for* Mycobacterium tuberculosis**
Ali-Berenice Posada-Reyes, Yalbi I. Balderas-Martínez, Santiago Ávila-Ríos, Pablo Vinuesa and Salvador Fonseca-Coronado

266 ***Mapping CAR T-Cell Design Space Using Agent-Based Models***
Alexis N. Prybutok, Jessica S. Yu, Joshua N. Leonard and Neda Bagheri

284 ***Generalized Property-Based Encoders and Digital Signal Processing Facilitate Predictive Tasks in Protein Engineering***
David Medina-Ortiz, Sebastian Contreras, Juan Amado-Hinojosa, Jorge Torres-Almonacid, Juan A. Asenjo, Marcelo Navarrete and Álvaro Olivera-Nappa

294 ***Time-Optimal Adaptation in Metabolic Network Models***
Markus A. Köbis, Alexander Bockmayr and Ralf Steuer

310 ***Dynamic Modelling of DNA Repair Pathway at the Molecular Level: A New Perspective***
Paola Lecca and Adaoha E. C. Ihekwaba-Ndibe

frontiers | Frontiers in Molecular Biosciences

TYPE Editorial
PUBLISHED 21 October 2022
DOI 10.3389/fmolb.2022.1048727

Check for updates

OPEN ACCESS

EDITED AND REVIEWED BY
Ugo Bastolla,
Spanish National Research Council (CSIC), Spain

*CORRESPONDENCE
Daniel Garrido,
dgarridoc@ing.puc.cl
Alberto J. M. Martín,
alberto.martin@uss.cl
Ernesto Pérez-Rueda,
ernesto.perez@iimas.unam.mx

SPECIALTY SECTION
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

RECEIVED 19 September 2022
ACCEPTED 10 October 2022
PUBLISHED 21 October 2022

CITATION
Garrido D, Martín AJM and Pérez-Rueda E (2022), Editorial: Systems modeling: Approaches and applications–Volume II. *Front. Mol. Biosci.* 9:1048727. doi: 10.3389/fmolb.2022.1048727

COPYRIGHT
© 2022 Garrido, Martín and Pérez-Rueda. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Editorial: Systems modeling: Approaches and applications–Volume II

Daniel Garrido[1]*, Alberto J. M. Martín[2]* and Ernesto Pérez-Rueda[3]*

[1]Department of Chemical and Bioprocess Engineering, School of Engineering, Pontificia Universidad Católica de Chile, Santiago, Chile, [2]Centro Científico y Tecnológico de Excelencia Ciencia & Vida, Fundación Ciencia & Vida, Escuela de Ingeniería, Facultad de Ingeniería, Arquitectura y Diseño, Universidad San Sebastián, Santiago, Chile, [3]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Unidad Académica Yucatán, Mérida, Mexico

KEYWORDS

systems modeling, biomedical, microbial biochemistry, metabolic engineering, theoretical approaches

Editorial on the Research Topic
Systems modeling: Approaches and applications–volume II

Introduction

The development of modeling tools has permitted an increased understanding of how components in different systems interact and behave. Thus, systems modeling has led to critical advances in several areas, such as medicine, biotechnology, and engineering. Applications include the study of ecological models, diseases and the impact of treatments, microorganism responses to specific environments, and the interactions between biomolecules.

The main goal of this Research Topic (*Systems Modeling: Approaches and Applications–Volume II*) was to bring together novel biological applications and studies on systems modeling. We were thrilled to witness the great interest in the field and the high number of manuscripts submitted. Broadly, works published in this section could be classified in four main categories: biomedicine, metabolic engineering, microbial biochemistry, and theoretical approaches and novel applications.

Among the biomedical articles of the number, Ponce-de-Leon et al., expanded a hybrid multi-scale model including time and space variables of fibroblast spheroids, which could be useful for addressing cancer cell resistance by including time, geometric and population variability. Prybutok et al. extended an agent-based modeling framework to design alternatives in Chimeric antigen receptor (CAR) T-cell therapy. This approach might accelerate the discovery of novel strategies

against solid tumors. Later, Ordaz-Arias et al. presented a regulatory network of macrophages, in order to study their plasticity, adaptability, and heterogeneity, and finding oscillations derived from the network structure. Zinovyev et al. developed a model of the cell cycle at the single cell level including internal dynamical cycles and switches. It predicted with great accuracy cell doubling times. Lecca and Ihekwaba-Ndibe et al. focused on DNA repair mechanisms, developing a mathematical model of the gene regulatory network of this biological process. Applications of the model include evaluation of the effect of certain mutations and control of participating genes. Gupta et al. studied the molecular mechanisms involved in the response of macaques to malaria using transcriptomics and metabolic modeling. Gupta et al. focused on pathogen detection pathways and inflammasome assembly, developing a comparative analysis identifying points of control for maintaining immune balance. Finally, Gil et al. developed a 3D model of calcium signaling pathways in T-cells to investigate how calcium microdomains occurred and included the role of ryanodine receptors in TCR/CD3 stimulation.

Microbial biochemistry articles include a study of Spolaor et al., who developed a mathematical model of the effect of hypotonic shock on calcium homeostasis and signaling pathways in *Saccharomyces cerevisiae*. The model included mechanosensitive channels, and provided an interpretation of regulatory processes in wild type and mutant yeasts. Verhagen et al. addressed resource limitations on optimal proteome allocations developing a resource-dependent kinetic model of *S. cerevisiae*. The model predicted proteome adaptations to multiple conditions with changing resources, and could be useful for industrial yeast applications. Rajeshkannan et al. presented a mathematical model of the GAL regulon in *S. cerevisiae*, which showed that binding affinities between regulatory proteins modulate gene expression at the single cell and population levels. Finally, Posada-Reyes et al. analyzed polymorphic interactions in the genomes and pangenome of *Mycobacterium tuberculosis*. They presented an epistatic network for this microorganism and identified targets of co-selection, contributing to our understanding of *M. tuberculosis* pathogenesis.

Metabolic engineering manuscripts included Landon et al., developing an analysis pipeline to interpret metabolic reaction fluxes, integrating machine learning, dimensionality reduction and network analysis. The work presented by Landon et al. focused on the *Mycoplasma genitalium* whole-cell model and the contribution of the model to understand gene knock-outs in a minimal genome. Doan et al. presented a coarse-grained mathematical model, comprising a micromolecular and a macromolecular component, aimed to represent a cell proteome during microbial growth in a bioprocess. Köbis et al. developed a constraint-based model that considers a time-optimal control problem, which allows to determine the fastest possible adaptation of a system to a cellular state. Boada et al. used multiobjective optimization for tuning gene circuits composed of a controller and a biosensor controlling metabolic pathways. This study might contribute to optimizing microbial cell processes and system robustness and stability. Finally, Lazaro et al. constructed a mathematical model that integrates two major steps in bioprocessing: single cell growth captured by a genome-scale metabolic model with bioreactor dynamics. This work used production of citralamate in *Escherichia coli* as case study, and might have important applications in biotechnological processes.

At last, other articles presented theoretical approaches, code and novel applications. Stoll et al. introduced a framework modeling dynamic population of interacting cells, based on probabilistic simulations and using TNF-induced cell death as case study. Litwin et al. addresses the task of determining model parameters in an ODE-based model, proposing a 2D likelihood approach to aid in optimal experimental design for parameter determination. Selvaggio et al. addressed the need of quantitative data required for calibrating model parameters. They proposed a hybrid model integrating ODE and logical formalities to describe biological complexity in layers and their communication. Medina-Ortiz et al. maximized the performance of predictive models in protein engineering, generalizing property-based encoders. This work contributes to predictive protein engineering without increasing model complexity. Massing et al. reviews generalized modeling, an approach to conventional dynamic modeling, highlighting recent advances and providing an application guide for this approach. Voit and Olivença focused on the Biochemical Systems Theory, an ODE-based approach for biochemical reaction analysis and simulation, expanding this theory to include stochasticity, discreteness and addressing time delays.

We consider that the field of systems modeling is in expansion, supported by the great quality and number of manuscripts included in this special issue. We appreciate the great interest and reception of the community, and hope that this special number is of interest for researchers in the field of computational biology, biochemistry, biomedicine, bioengineering and mathematics.

Author contributions

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

Funding

This work was supported by Dirección General de Asuntos del Personal Académico-Universidad Nacional Autónoma de México (IN-209620) and CONACYT (320012), Fondecyt 1190074 and 1181089, FONDEQUIP EQM190070 and Centro Ciencia & Vida, FB210008, Financiamiento Basal para Centros Científicos y Tecnológicos de Excelencia de ANID.

Acknowledgments

We thank Israel Sanchez and Manuel Lira for their technical support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Molecular Biosciences

TECHNOLOGY AND CODE
published: 17 December 2021
doi: 10.3389/fmolb.2021.732079

Understanding Metabolic Flux Behaviour in Whole-Cell Model Output

Sophie Landon[1,2], *Oliver Chalkley*[1,2,3], *Gus Breese*[2], *Claire Grierson*[1,4†] *and Lucia Marucci*[1,2,5†*]

[1]BrisSynBio, University of Bristol, Bristol, United Kingdom, [2]Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom, [3]Bristol Centre for Complexity Science, Department of Engineering Mathematics, University of Bristol, Bristol, United Kingdom, [4]School of Biological Sciences, University of Bristol, Bristol, United Kingdom, [5]School of Cellular and Molecular Medicine, University of Bristol, Bristol, United Kingdom

Whole-cell modelling is a newly expanding field that has many applications in lab experiment design and predictive drug testing. Although whole-cell model output contains a wealth of information, it is complex and high dimensional and thus hard to interpret. Here, we present an analysis pipeline that combines machine learning, dimensionality reduction, and network analysis to interpret and visualise metabolic reaction fluxes from a set of single gene knockouts simulated in the *Mycoplasma genitalium* whole-cell model. We found that the reaction behaviours show trends that correlate with phenotypic classes of the simulation output, highlighting particular cellular subsystems that malfunction after gene knockouts. From a graphical representation of the metabolic network, we saw that there is a set of reactions that can be used as markers of a phenotypic class, showing their importance within the network. Our analysis pipeline can support the understanding of the complexity of *in silico* cells without detailed knowledge of the constituent parts, which can help to understand the effects of gene knockouts and, as whole-cell models become more widely built and used, aid genome design.

Keywords: whole-cell modelling, machine learning, networks, snorkel, time series, weak learning

OPEN ACCESS

Edited by:
Ernesto Perez-Rueda, Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Alessio Franci, National Autonomous University of Mexico, Mexico
Paulo Burke, University of Colorado, United States

***Correspondence:**
Lucia Marucci
lucia.marucci@bristol.ac.uk

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *28 June 2021*
Accepted: *28 October 2021*
Published: *17 December 2021*

Citation:
Landon S, Chalkley O, Breese G, Grierson C and Marucci L (2021) Understanding Metabolic Flux Behaviour in Whole-Cell Model Output. Front. Mol. Biosci. 8:732079. doi: 10.3389/fmolb.2021.732079

INTRODUCTION

Recent years have seen a significant increase in the availability of high-throughput biological data (Gomez-Cabrero et al., 2014). The integration of data from methods that are becoming cheaper and more accessible (Wetterstrand, 2010) reveals interactions between cellular processes (Manzoni et al., 2018), aiding analysis (Zampieri et al., 2019). Leaps in the scale and capabilities of biological modelling give great scope for *in silico* data generation, and though mathematical models cannot fully replicate living cells, their output can help to understand biological mechanisms and inform experimental design to improve *in vivo* data collection. These models can formalise processes at a specific level (e.g., translation) or construct a trans-omic network of the relationship between different cellular processes (Yugi et al., 2019) and couple metabolism with gene expression (O'brien et al., 2013). Whole-cell models simulate every cellular process throughout the life cycle of a cell—only two are published, which model the life cycle of *Mycoplasma genitalium* (Karr et al., 2012) and *Escherichia coli* (Macklin et al., 2020). We focus on the *M. genitalium* model. This consists of 28 submodels that use multiple mathematical methods (linear programming and differential equations) to represent processes such as metabolism and cytokinesis, which integrate together at every time step.

The model is highly complex, is computationally expensive, and generates huge amounts of time series data relating to thousands of variables. Interpreting whole-cell model *in silico* data can be

difficult, but large-scale analysis is possible. Tools are required to automatically process and consolidate the output so they can be viewed and clarified, even by those with little computational expertise. Existing software tools that visualise whole-cell model output (Lee et al., 2013; Karr and Pochiraju, 2018) have limited capacity for processing large and varied datasets—they focus on visualisation of different output streams, so all analyses are done by eye, and there is no dimensionality reduction or statistical methodology.

A whole-cell model, with appropriate analysis software to process its output, could be a powerful predictive tool for gene editing. Genetic modifications can be trialled in a model before being physically made to save time and resources, and whole-cell models can be coupled with algorithms to predict genetic modifications intended to produce a chosen phenotype (Haimovich et al., 2015). Machine learning methods are suitable for whole-cell model analysis as they are data-driven, so they can identify correlations and classify data with few assumptions and little biological knowledge. Metabolism is one of the most widely modelled cellular subsystems; a stoichiometric matrix is used to create a constraint-based metabolic model (CBM), which can be used to predict steady-state fluxes (Bordbar et al., 2014). There have been applications of machine learning to CBMs, consolidated by Zampieri (Zampieri et al., 2019). Many have coupled CBMs with discriminative classifiers (Ho, 1995; Noble, 2006; Yegnanarayana, 2009), to predict or classify gene essentiality, drug side effects, and protein functions. Others have used unsupervised learning to explore patterns and pathways in metabolic systems (You et al., 2006). These methods of prediction and analysis can be scaled to whole-cell models. However, whole-cell model output is composed of time series—contrary to CBM output, which is steady-state rates—and the labelling of these types of data is becoming a barrier to large-scale machine learning. As computational power increases and new data analysis algorithms are developed, the availability of fully labelled datasets to train and validate models is a limiting factor, and so new methods are being formed to automatically label data.

Time series data come from all physical systems. Difficulty in interpreting it arises from the importance of ordering of different events, meaning that attributes of the data are dependent on each other in complex ways (Hannan, 2009). Of the various machine learning methods for time series classification, deep learning has emerged as the most reliable (Wang et al., 2017; Fawaz et al., 2019), although accuracies of each method vary with different datasets. There are also other factors that affect the performance of an algorithm, such as feature selection, feature engineering, and data pre-processing.

Many of these methods are supervised, meaning that they require labelled data in order to train a model. Historically, these labels would be manually generated by an expert to capture the ground truth of the problem, but labelling data manually is time-consuming and unfeasible for huge datasets. A solution to this problem is weak supervision, which uses weak labels (that do not express the ground truth) created from a model designed to map labels onto instances of the data (Zhou, 2018). Snorkel is a methodology that creates a generative model (a statistical model of the joint probability of a variable and target label) to automatically produce weak labels, after collating metrics from multiple manually defined labelling functions using features from the data (Ratner et al., 2019; Ratner et al., 2017).

Feature extraction is one of the most important aspects of building a machine learning model and can be the difference between failure and success (Domingos, 2012). It is also generally based on expert knowledge about the physical system (Barandas et al., 2020), as the most relevant features for analysis will vary depending on the objective of the machine learning model and the behaviour of the time series. The issue of time series analysis of whole-cell model generated metabolic flux is that there is very little experimental data for dynamic flux in bacterial cells, so the features that best define the flux behaviour are not intuitive. There has been previous work on dynamic metabolic fluxes, where reactions rates were calculated from derivatives of measured external metabolite concentration, or using dynamic metabolic flux analysis (DMFA) (Kuriya and Araki, 2020). For DMFA, a metabolic flux analysis process was used to minimise the sum of squared residuals between the actual and predicted flux rates. Then, the DMFA process was used to fit linear functions between consecutive time points. The methods were computationally inexpensive, due to the linear fit, and it was found that a lower number of time points produced a fit with smaller confidence intervals, suggesting that linear fits are suitable for approximating metabolic fluxes. Another method used dynamic flux balance analysis (dFBA) and polynomial fitting to find functions for reaction rates (Leighty and Antoniewicz, 2011). Polynomial functions were fitted to experimental data from metabolite concentrations, which were then differentiated to find functions for growth rate. These were used as boundaries for dFBA, enabling accurate simulations of reaction behaviour in time. Both of these methods deal with relatively smooth data, and estimation of fluxes from concentration derivatives also involves a smoothing process, which results in loss of information (Lequeux et al., 2010). As some of the flux behaviour we see from the *M. genitalium* model oscillates significantly in time (as in **Supplementary Figure S1**), to analyse this, we must extract features that can capture some of the variation. Analysis of oscillatory time series is relatively common, but this is usually within the context of understanding the physical system—for example, oscillatory time series decomposition has been carried on the phase dynamics of well-understood systems (Matsuda and Komaki, 2017).

It is important to consider that most machine learning algorithms are treated as black boxes, so results are created without context. For explanations of the functions of underlying structure in complex systems, network science can be used (Gosak et al., 2018). Network science is an area that has long been applied to the analysis of biological systems: protein interactions, metabolic reactions, and transcription regulation can be formalised as networks, leading to discoveries regarding properties of their interactions (Barabasi and Oltvai, 2004). Network structure has been used to predict metabolic functions and find pathways for metabolite flow (Stelling et al., 2002) and to find control loops within gene networks (Wong et al., 2012).

The complexity of genomic interactions, even in cells as small as *M. genitalium*, is such that there is not a clear path from the

TABLE 1 | Accuracies of Snorkel's weak labels for 10 manually labelled reactions.

	Accuracy
Aas4	99.6%
AceE	99.1%
Adk3	90.2%
Apts_Asp	95.8%
Apts_Trp	83.1%
ArcC	77.8%
DcdK	97.0%
Pyk_DADP	85.0%
Pyl_GDP	69.2%
TX_AROP22	94.3%

genome after knockouts to the end phenotype. Even with functional annotations, the genomic context of the genome (which will be several hundred genes after a single gene knockout) cannot be disregarded, as there may be redundancy in the genome, or unprecedented gene product interactions. The removed gene/s will not tell the full story, but zooming out to examine a large set of different genotypes through their metabolic fluxes can show us the trends across the full set of knockouts, providing a different angle than that of focusing on a single gene.

Here, we present a novel analysis pipeline that combines whole-cell model simulations of wild-type and gene knockout cells with time series classification and network analysis. The main steps include automatic labelling of metabolic fluxes as normal or abnormal (where normality refers to the behaviour of a reaction flux from a knockout simulation with respect to the behaviour of that reaction in a wild-type simulation), dimensionality reduction of the reactions for visualisation, and network analysis of the reactions. This analysis—looking at intermediate steps that connect genotype to phenotype—aims to increase our understanding of cellular processes and provides foundations for *in silico* genome design.

MATERIALS AND METHODS

Description of the Data

We began with two sets of data—one to train the machine learning models and one to apply them and analyse the output. The simulations were generated from running the *M. genitalium* whole-cell model on a supercomputer cluster, with each gene singly knocked out. The model requires 8 GB of RAM for each simulation and was run on BlueGem, a 900-core supercomputer at the University of Bristol, using MATLAB R2013b. It is available at https://github.com/CovertLab/WholeCell. The raw metabolic flux time series was then converted to Pandas DataFrames and stored in a pickle format to save space. The training set consisted of time series of reaction fluxes for three repetitions of every possible single knockout from the *M. genitalium* model, of which there are 359, plus 200 wild-type simulations. Each time series is 50,000 s in total, and we used the time series of 279 reactions from each simulation. There was 1,270 simulations in total. The dataset that we applied to the analysis consisted of 10 repetitions of all of the single gene knockouts, with the same reaction time series, and so this dataset has 3,411 simulations in total. One knockout, MG_469, consistently caused the model to crash and the simulations to terminate, and a few simulations did not complete due to errors on the supercomputer cluster. The metabolic flux data are about 200 Mb per simulation after processing, so the training dataset (three repetitions of each single knockout) is ≈200 Gb, and the analysis dataset ≈700 Gb. More repetitions of each knockout would make for a more accurate dataset, but due to the size of the data, we were limited by storage space.

Labelling

Snorkel is a system that takes input data points and manually defined labelling functions and collates these into a generative model that outputs probabilistic labels for the data. The labelling functions will produce noisy labels, which are then used as weak supervision for a stronger predictive function by combining three measures—the labelling propensity (whether the data point has been assigned a label or not), the accuracy of each label, and the correlations of the multiple labelling functions. The label matrix generated from these measures is then used to define an exponential distribution that can predict probabilistic training labels. The normality of 10 reactions was manually labelled by visual inspection of the time series, comparing features of the plots such as smoothness and linearity with wild-type time series from the same reactions (Correll et al., 2012; Correll and Heer, 2017), and used to validate Snorkel's weak labels, the accuracies of which are shown in **Table 1**. The algorithm was implemented using the Snorkel library in Python.

The manual labelling was done based on the phenotypic classes defined by the original publication of the *M. genitalium* model, which used the production capacity of various features from the model output to classify a simulation (Karr et al., 2012). The combinations of these features that contribute to a particular class are detailed in **Table 2**, and the simulations used in the analysis dataset were all labelled by manual inspection of the model output.

TABLE 2 | Manual labels of phenotypic classes (shown on the left-hand column) and their corresponding combinations of substance production (the column headings).

	DNA	RNA	Protein	Growth	Division
Metabolic	×	×	×	×	×
RNA	✓	×	×	×	×
Protein	✓	✓	×	×	×
Slow growing	✓	✓	✓	✓	×
DNA	×	✓	✓	✓	×
Septum	✓	✓	✓	✓	×
Non-essential	✓	✓	✓	✓	✓

Note. A cross means that there is no active production of that substance in the case of DNA, RNA, and protein; and in the case of growth and division, these things do not occur. A tick means that opposite—so, for example, in a simulation classified as "non-essential," we see production of DNA, RNA, and protein, as well as both growth and division; and in a simulation classified as "metabolic," we see none of these things. In the case of the "slow growing" phenotype, division begins at the end of the simulation but does not complete.

Training and Tuning the Neural Networks

Once the data are fully labelled, a standard discriminative model can be trained for classification. In this case, we chose to use a neural network, implemented with the Python library tensorflow (version 2.0.0-rc0). With the use of the data labelled by the generative model, a neural network was trained for each reaction. Each neural network had four hidden layers and used a softmax activator function and Adam optimiser. Different combinations of hyperparameters (epoch size, batch size, and number of nodes in a layer) were tested, so that an optimal combination could be used for each network to find the highest accuracy. Generally, the combination of hyperparameters can have a significant effect on the neural network output, so these factors are important. Epoch size refers to the number of rounds of back-propagation performed by the network, batch size means the number of training data samples input before the model updates, and number of nodes refers to number of nodes of the network in each hidden layer. Epoch size will leave the data underfitted if too small and overfitted if too large; batch size is generally optimised for processing time (in that larger batch sizes will train the network faster, whereas a smaller batch size may help the weights converge faster); and the number of nodes is usually taken to be some number between the amount of input nodes and the amount of output nodes. There is no set method for selecting hyperparameters for neural nets, and it is frequently taken to be a trial-and-error process (Sarle, 1994). We tuned our neural networks *via* a brute-force approach, where different parameters within a set range were trialled to increase the accuracy of the network. Epoch size was kept relatively low; as after some testing, many of the neural networks converged to accuracies $>95\%$ after only five epochs, and so we tested epoch values of 5, 10, and 15. Batch sizes of 50, 100, and 150 used, and node numbers of 750, 1,500, and 2,250 were tried, where we selected the network with hyperparameters that gave the highest accuracy. The reactions from neural networks that gave accuracy of less than 70% were removed, leaving 267 reactions and neural networks with a mean accuracy of 93.6%. K-fold cross-validation was performed to check if overfitting was an issue, using the sklearn Python library (version 0.21.3), with 10-fold. The accuracies across the folds are shown in **Supplementary Figure S2** and averaged across the folds for each reaction. As the averaged accuracies across the folds do not differ significantly from accuracies recorded, we conclude that the data have not been overfitted.

Network Formation and Features

After the neural networks were trained and fluxes classified across the dataset, we turned to network analysis. With the stoichiometric matrix for the metabolism, **S**, taken from the *M. genitalium* model knowledge base, we reduced it to its binary format (as we were focusing on the topology of the metabolic network rather than the exact stoichiometry) to form a metabolic adjacency matrix **A** from the relationship

$$\mathbf{A} = \mathbf{S}^{t}\mathbf{S}, \tag{1}$$

which can create a widely used graphical representation of a metabolic network, where the reactions form nodes of the graph, and the substrates form edges that connect them (Palsson, 2006). We were able to find a set of driver nodes (the set of nodes that must be controlled in order to fully control the network) using the maximal_matching function in Python's NetworkX library (version 2.4). This function takes an undirected graph and greedily finds a matching by iterating over pairs of edges in the graph to see whether the node that connects them is in the matching. The pathways associated with the driver nodes were found *via* the Enzyme Commission numbers from the supplementary material of the *M. genitalium* model (Karr et al., 2012), where the Python library bioservices was used to look up the pathways for each EC number from Kyoto Encyclopedia of Genes and Genomes (KEGG).

The metabolic sub-networks were plotted in python-igraph where, for each class across the dataset, the affected reactions are shown as a sub-network with a colour gradient corresponding to how frequently that reaction behaves abnormally. A threshold for "noisy" reactions was found from wild-type simulations, where an exponential distribution was fitted to the frequencies of reactions classified as behaving abnormally by the neural networks. For a wild-type simulation, in theory, all reactions should be classified as normal, but as the *M. genitalium* model is stochastic, there can be a range of different behaviours, depending on the initial conditions of the simulation and other random processes (e.g., radiation and DNA damage). The interval under which 95% of the data were contained was found, and this value was selected as a rate parameter, which was used as the threshold of significance for whether a reaction was considered to be behaving abnormally consistently.

We then performed principal component analysis (PCA) using the SciPy library (version 1.3.1) to reduce the data to two dimensions and plot the data on a scatter plot using Seaborn (version 0.9.0). After the reduction, 84% of the variance in the full dimensions of the data was conserved, so there was no significant information loss after this operation. After having found the driver nodes, we trained a linear support vector machine (SVM) for the normality of each one to separate the data points on the PCA plot, selecting those that could divide the data with $>95\%$ accuracy. For the SVM, we used the sklearn Python library (version 0.21.3).

RESULTS

A schematic of our pipeline is shown in **Figure 1**, with the main steps of weak labelling, neural network classification, and network analysis shown. We began with two datasets: one for training and testing the neural network and one for analysis of single knockouts. The training dataset contained three repetitions of all 359 possible single gene knockouts, plus 200 wild-type simulations, giving 1,270 simulations in total. Each simulation had 279 dynamic reactions out of the total 645 (over half of the reactions were consistently at steady state throughout the cell life cycle, which does not require a complex classifier to identify), with up to 50,000 timesteps. Although the exact steady-state values may vary across simulations, we focused specifically on the reactions that have behaviours that change in time, assuming that they are more likely to show the most sensitive components of the metabolism. Given that metabolic networks are formulated with

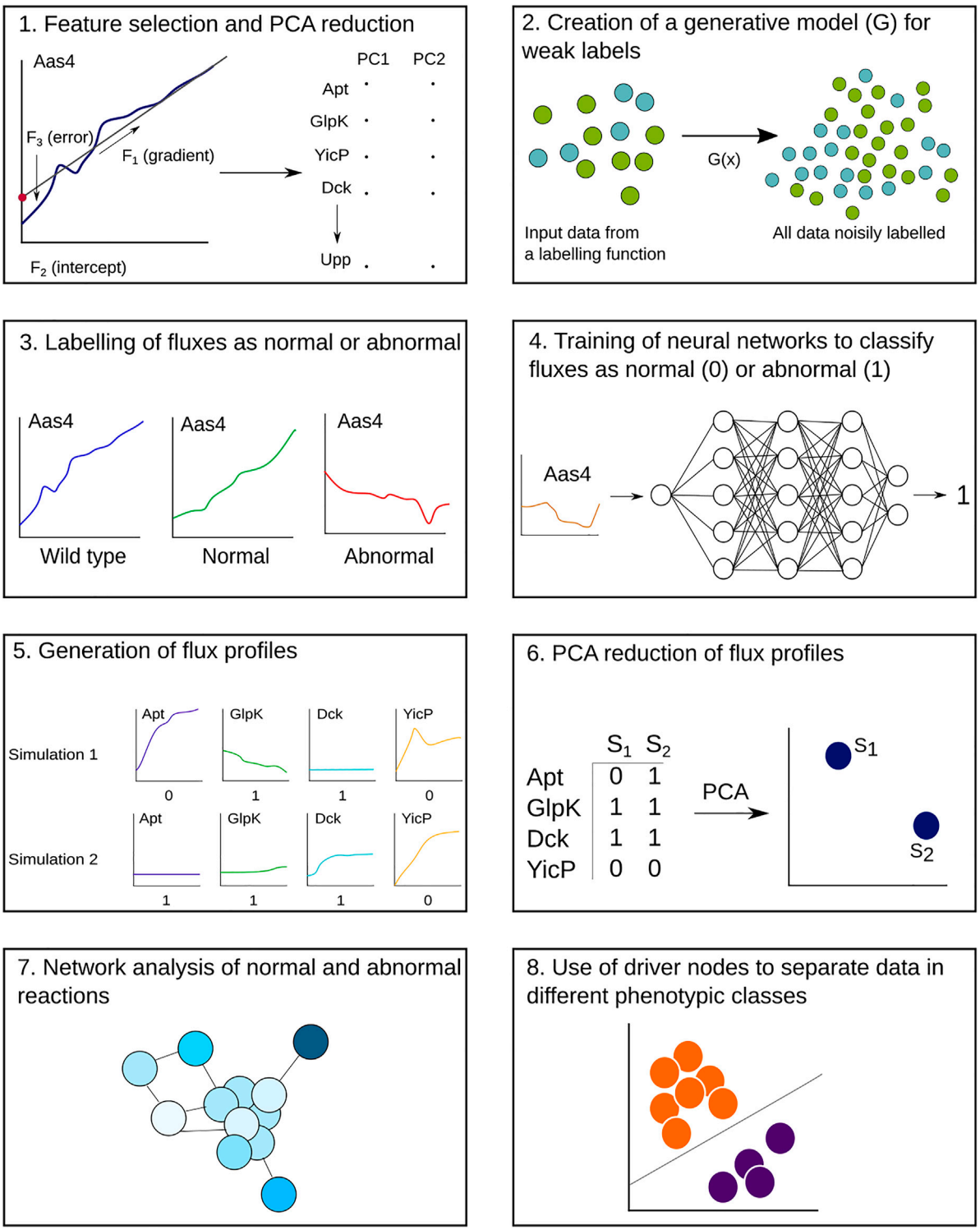


FIGURE 1 | Step-by-step workflow of the analysis pipeline, beginning with the metabolic fluxes from the whole-cell model output. Steps 1–4 are applied to a training dataset of gene knockout simulations, where the end result is a trained neural network for each reaction. (1) For each reaction flux time series, four features are extracted and reduced to two dimensions through principal component analysis (PCA). (2, 3) The extrema of these data are used to define boundaries for normal and abnormal behaviours, which are then used to create a generative function to map labels onto the reactions. (4) Neural networks are trained using these labelled data to classify reactions as normal or abnormal. Steps 5–8 are applied to a separate analysis dataset of gene knockout simulations. (5) The neural networks are used to classify the analysis dataset and create a flux profile for each simulation. (6) The flux profiles are reduced to two dimensions and plotted. (7, 8) Network analysis of the reactions reveals nodes that control the metabolic network and correlate with different phenotypes after gene knockouts.

steady-state behaviours in mind, reactions that deviate from this seemed to be the most interesting to analyse, with regard to understanding the cell phenotype. The analysis dataset consisted of 10 repetitions of the 359 gene knockouts, with the same number of reactions and timesteps, totalling 3,411 simulations. There are some gaps in the dataset, as some files were corrupted, and one knockout consistently caused the model to crash.

The *M. genitalium* whole-cell model has drastically varying fluxes through different reactions (see **Supplementary Figure S1**). Furthermore, it is not always clear how the removal of a particular gene will affect cellular processes or cell viability. For each reaction, we presume there is a range of normal behaviours over which the cell can produce all necessary compounds for division, and dynamics outside of that range result in negative effects (e.g., build-up or depletion of certain metabolites) that affect the rest of metabolism and disrupt other processes, potentially causing cell death. The normality of reaction fluxes in a simulation can be used to understand the effects of gene knockouts through the cell cycle, and how metabolism is affected. This can help with predicting and explaining the effects of gene knockouts and looking at patterns across different simulations. We visualised the reaction flux behaviour across our entire dataset, and we looked at the topology of the metabolic network (in particular, how the network can be controlled by input nodes) to help explain the role of different reactions.

Implementation of Snorkel for Weak Labelling

Manual labelling was impractical with such a large dataset, so we implemented Snorkel, which has previously been shown to perform as accurately as hand labelling (Ratner et al., 2019). There are other methods of weak supervision available, but they use either inaccurate labels (which still require a manually labelled dataset) or locate incorrect labels within a previously labelled dataset Northcutt et al. (2019). Inaccurate labels are those that are known to be incorrect, and imprecise labels are those that contain some high level information about the data, but do not show the ground truth. Snorkel is the main approach that uses imprecise labels for time series (Robinson et al., 2020), as other approaches have used imprecise labels for semantic similarity in words, which is not applicable to time series (Saunshi et al., 2019).

Snorkel requires manually defined labelling functions, which are an important heuristic for the basis of the methodology. The underlying patterns are used to form probabilistic labels, so together they need to capture some approximation of ground truth. In this case, we created labelling functions by amalgamating four key features extracted from each reaction flux time series. There is very little information in the literature about what normal behaviours for metabolic fluxes should look like, so we must make an assessment of the most important features from time series inspection.

As Snorkel is designed to work with noisy and sometimes conflicting labels, we used a simple method to define the labelling functions. A linear regression function was fitted to each time series; and the intercept, gradient, coefficient of determination (R^2), and mean squared error were found (**Figure 2**). These captured the variation observed and shown in **Supplementary Figure S1**: smoothness/oscillation in the mean squared error, the increasing or decreasing nature in the gradient, and the linearity in the coefficient of determination. These were features that we chose based on manual inspection of the reaction behaviour, with the intent of describing the important aspects of the time series, so in choosing them we aimed to capture the most relevant information. Fitting non-linear functions to the data may have provided more accurate labelling functions, but due to the complexity and variety of the time series, this would have required a many visual analyses and likely a broad set of different non-linear functions.

The results were reduced through PCA (where 82.9% of the variance was conserved across the reactions), leaving a two-dimensional space over which boundaries of different thresholds could be drawn, which was much simpler and faster to visualise and compute in two dimensions than it would have been before the dimensionality reduction (**Figure 2**). Using two dimensions allowed us to easily verify visually the efficacy of this labelling method while approximately dividing the data for the weak labelling. Loosely, the boundaries were defined by the extrema of the wild-type simulations, which were taken to be the edges of normal behaviours for each reaction (**Figure 2**). Any simulations outside these boundaries were classified as abnormal. Other shapes could also be used at this stage.

Three different boundaries were defined for different labelling schemes, as different confidence thresholds performed better or worse depending on the reaction. Boundaries at the extrema and then at 99% and at 95% were selected as the three labelling functions after comparison of their performance and then combined to form the generative model. We then implemented Snorkel, leaving us with 1,270 weakly labelled time series for each reaction. Ten reactions were manually labelled as normal or abnormal to test the accuracy of Snorkel's labels, where characteristics like smoothness or the increasing or decreasing nature of the time series were used as comparison features to decide whether the behaviour of a reaction was normal or abnormal. The majority of the Snorkel labels gave over 90% accuracy, with the lowest at 69.2% (see the *Materials and Methods* and **Table 1**).

Training of Neural Networks and Flux Profiling

The Snorkel results were used to train a neural network for each reaction, as artificial neural networks are some of the most effective classification algorithms (Caruana and Niculescu-Mizil, 2006; Raczko and Zagajewski, 2017). Neural networks consist of layers of nodes, representing artificial neurons with assigned weighted connections. The weights are adjusted through rounds of backpropagation or epochs until they predict correct classes for different types of input (Kröse et al., 1993).

Once trained and assessed for accuracy using k-fold cross-validation to verify that they had not been overfitted (see the *Materials and Methods* section and **Supplementary Figure S2**), the neural networks were used to classify the normality of reactions for the analysis dataset. From this, we generated a "flux

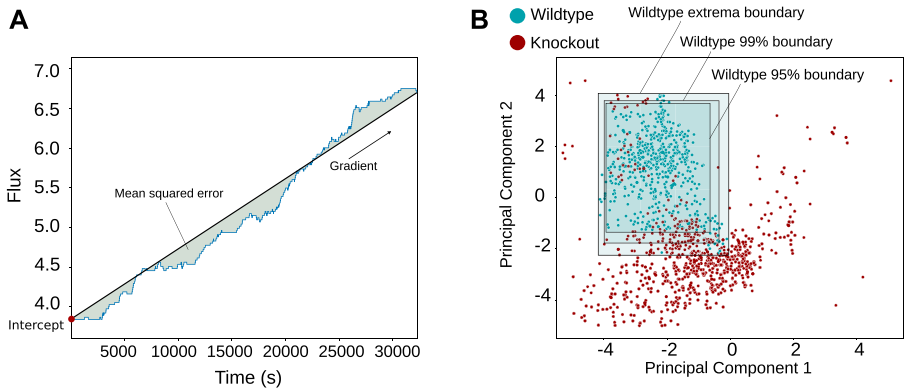


FIGURE 2 | Method of feature extraction and normality classification shown graphically. The features shown in plot **(A)** were taken, and principal component analysis (PCA) was applied for each reaction flux in each simulation across the dataset to create plot **(B)**, retaining 82.9% of the variance. The wild-type simulations are shown in blue, and then three boundaries are shown (extrema, 99% confidence intervals, and 95% confidence intervals), which are used to form three labelling schemes.

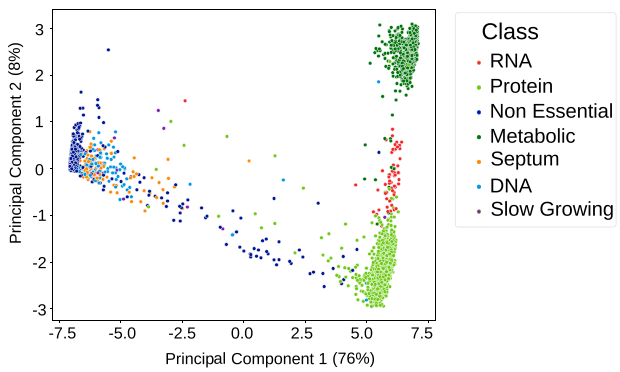


FIGURE 3 | Principal component analysis (PCA) plot of the flux profiles (binary strings of normal vs. abnormal classifications for each flux in a simulation) from 3,411 gene knockout simulations, reduced to two dimensions (while retaining 84.5% of the variance) and then shown in different colours that correspond to manual labels. The classes are defined by the presumed root cause of lack of cell division; or if the cell divides, the simulation is classified as non-essential.

profile" for each simulation: a binary string for each reaction within that simulation, where 0 means normal behaviours and 1 means abnormal. Reactions for neural networks with less than 70% accuracy were removed (of which there were 12 in total), leaving 267 reactions with a mean accuracy of 93.6%. We applied PCA to reduce the flux profiles to two dimensions while retaining most of the variance and visualised, as shown in **Figure 3**. Each point is the flux profile of a simulation, and the principal components correspond to the reduced dimensions of the reaction flux profiles. As PCA preserves global and pairwise distances between all data points, unlike other dimensionality reduction processes that focus on local distance (such as t-Distributed Stochastic Neighbor Embedding (Van der Maaten and Hinton, 2008) and Uniform Manifold Approximation and Projection (McInnes et al., 2018)), this enables us to see not only the relationship between data points but the relationship between the different clusters, leading to clearer interpretability.

The analysis dataset simulations were previously hand-labelled by phenotype according to differences in cell behaviours of the simulation output. The labelling classes were non-essential, DNA disruption, RNA disruption, metabolic disruption, protein disruption, or septum disruption (Rees-Garbutt et al., 2020)—see the *Materials and Methods* for details. The non-essential class is defined by whether the cell divides or not, in keeping with current definitions of gene essentiality (Zhang and Zhang, 2008), and the other classes are defined by what is indicated by the output data to be the root cause of cell death.

In **Figure 3**, several clusters of flux profiles are visible. To validate their significance, we coloured the flux profile points according to manual labels of the phenotype that has occurred after the knockout; it can be seen (**Figure 3**) that these clusters correspond to the manually defined phenotypic classes. This suggests (intuitively) that different sets of reactions behave abnormally for each different class of phenotype, with different scales in the proportion of reactions affected, which will separate the different classes in the PCA space. We expect the majority of reactions in a simulation labelled as non-essential to be classified as normal and the non-essential simulations to be clustered together in the PCA space, as their flux profiles will be similar. Then, for simulations with greater disruption (e.g., the metabolic phenotypic class, where there is no growth and no DNA, RNA, or protein is created (see the *Materials and Methods* section), where many reactions are behaving abnormally), these will be placed much further away from the non-essential cluster.

Analysis and Biological Context Within the Metabolic Network

The clustering analysis is useful to show the big picture across the entire dataset but does not suggest much biological insight that could be applied to lab experiments. In order to make sense of the data in a way that can be used in an experiment, we need to understand these results at the scale of groups of genes or reactions. To ascribe biological meaning to trends seen across

TABLE 3 | List of all of the driver nodes, whether they can linearly separate different phenotypic classes in the PCA space, and their associated pathways (if available).

Driver	Pathways	Linearly separable
TX_CO2	Glycolysis, TCA cycle, pyruvate metabolism, carbon metabolism	Yes
TX_COA	Glycolysis, TCA cycle, pyruvate metabolism, carbon metabolism, pantothenate and CoA biosynthesis, methane metabolism	Yes
TXPYDX	Vitamin B6 pathway	Yes
TX_ACAL	Pentose phosphate pathway	No
TX_CAP	Purine metabolism, carbon metabolism	No
TX_DDCA	Glycerolipid metabolism	n/a
TX_FOR	One carbon pool by folate, carbon metabolism	n/a
TX_H2O2	n/a	Yes
TX_HDCA	Glycerolipid metabolism	Yes
TX_HDCEA	Glycerolipid metabolism	Yes
TX_LIPOATE	n/a	Yes
TX_NAC	Nicotinate and nicotinamide metabolism	Yes
TX_O2	Purine metabolism, pyrimidine metabolism	Yes
TX_OA	Pyruvate metabolism, carbon metabolism, methane metabolism	No
TX_OCDCA	Glycerolipid metabolism	Yes
TX_OCDCEA	Glycerolipid metabolism	Yes
TX_RIBFLV	Riboflavin metabolism, biosynthesis of secondary metabolites	Yes
TX_THF	One carbon pool by folate, folate biosynthesis	Yes
TX_TTDCA	Glycerolipid metabolism	n/a
TX_TTDCEA	Glycerolipid metabolism	n/a
Upp	Pyrimidine metabolism	Yes

Note. PCA, principal component analysis; TCA, tricarboxylic acid.

the dataset, we analysed the topology of the metabolic network, as this is a representation of the relationships between different reactions, and so we can see how it is affected by reactions behaving abnormally after knockouts. It has been shown that the modularity of the *E. coli* metabolic network corresponds to metabolic functions (Ravasz et al., 2002), and so, from a graphical perspective, we aimed to explain some of the biology behind the phenotypic classes and the flux profiles. The *M. genitalium* metabolic network is significantly smaller than many bacterial metabolisms (645 reactions vs., e.g., 2,382 in *E. coli* (Feist et al., 2007)), due to its genome size—however, analysis is not trivial. We used a graphical representation of the network, where each reaction is a node, and substrates that connect reactions are edges, as in the stoichiometric matrix of the metabolism in the knowledge base of the *M. genitalium* model. We visualised the reactions affected across each class in individual graphs, shown in **Supplementary Figure S3**.

There are multiple ways to gauge the importance of a node within a network. Most commonly used are centrality measures (Freeman, 1977), but for dynamic networks, we can focus on the control of the network *via* the nodes. From the graphical representation, we used a maximal matching algorithm to find driver nodes. Driver nodes are the set of nodes in the network that need to be managed in order to have full control over the system, which can be found for both directed and undirected networks (Liu et al., 2011; Nacher et al., 2019)—therefore, in terms of input into the metabolism and flow through the metabolic pathways, their behaviours affect other reactions downstream, and they could be indicators of phenotypes after gene knockouts. The driver nodes of the network are shown and named in **Supplementary Figure S4**. For each driver node, the pathways associated with that reaction were found from KEGG (Kanehisa and Goto, 2002) or (if there was no annotation for that reaction) the pathways associated with reactions that were one degree away from the driver, as shown in **Table 3**.

Metabolic networks are known to be robust (Smart et al., 2008; Holme, 2011), so many reactions can be individually removed without causing adverse effects. However, within *M. genitalium* metabolism, very few metabolites are organically synthesised (Dybvig and Voelker, 1996). Transport reactions for essential substrates such as amino acids are far more important than they might be in a larger cell that has the capabilities to synthesise these things itself. Within the metabolic network for the most widely used constraint-based *E. coli* model [iAF1260 (Feist et al., 2007)], 75% of the driver nodes are transport reactions, compared with 95% in the *M. genitalium* metabolic network.

We found several driver nodes that can be individually used as features to divide the data into separate classes (referred to in the text using their reaction identifiers from the model). For all driver nodes, we modelled a linear SVM across the 2D data of the analysis dataset. We then selected those that could separate the data into normal vs. abnormal behaviours with over 95% accuracy as good and simple indicators of metabolic behaviours, shown in **Figure 4**. Of the driver nodes, 83% could linearly separate the data with greater than 95% accuracy (listed in **Supplementary Table S4**), compared with only 60% of the non-driver nodes, demonstrating their significance. Additionally, we can use individual driver nodes to mark phenotypic classes—normal behaviours for TX_NAC, the reaction that transports nicotinamide into the cell, correlate strongly with the simulations classified as non-essential, with a phi coefficient [a measure of correlation between binary variables (Ekström, 2011)] of 92%. Behaviours of TX_RIBFLV can split the dataset into the classes where we see growth (non-essential, septum, and DNA phenotypes) and the classes where there is no growth (metabolic, RNA, and protein phenotypes) with a phi coefficient of 95%. Equally, we can see that abnormal behaviours for Upp

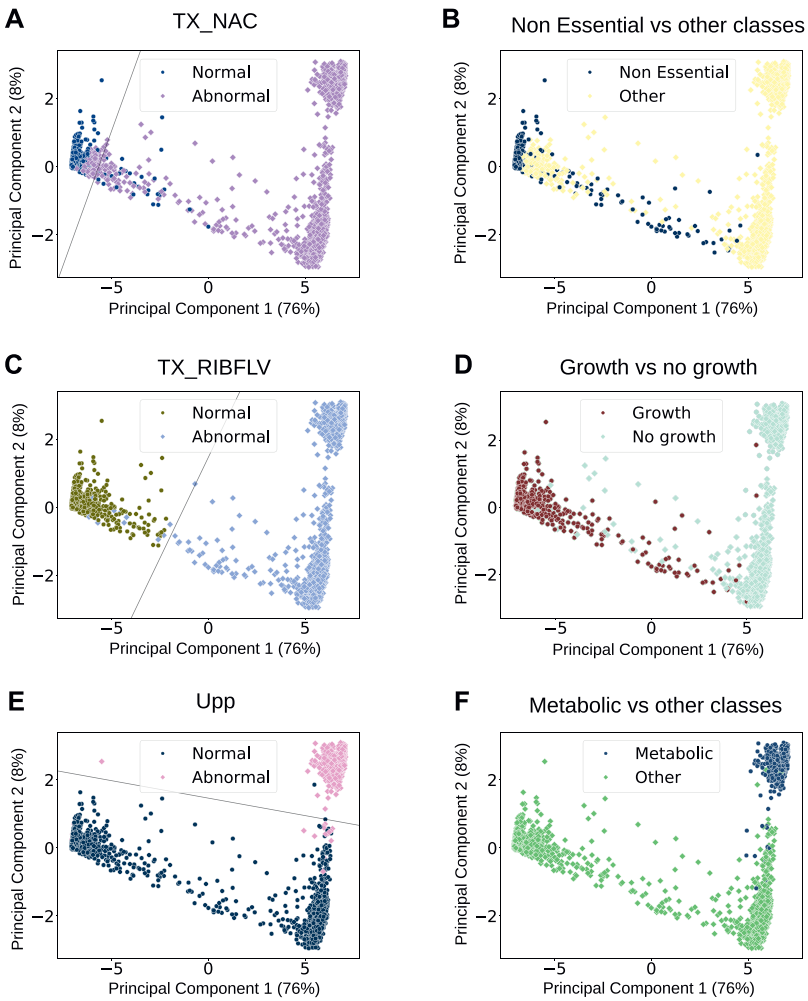


FIGURE 4 | Scatter plots of all flux profiles reduced to a 2D feature space. On the left [plots **(A,C, E)**], each point is labelled with the behaviour (normal or abnormal) of a single reaction that is a driver node. The lines shown decision boundaries for support vector machines (SVMs); models that form a hyperplane to linearly separate different classes of data, where in this case the classes will be flux profiles where the specified reaction behaves normally, and flux profiles where the specified reaction behaves abnormally. The reactions are referred to using their identifiers that are used in the model. On the right [plots **(B, D, F)**], each point shows the manual label of the phenotypic classes that correlate with the behaviour of the reactions on the left—the simulations that show normal behaviours of TX_NAC have a 92% correlation with those that are manually classified as non-essential; those that show normal behaviours of TX_RIBFLV show a 95% correlation with those that are manually classified as non-essential, DNA, or septum phenotypes (which all show cell growth), and those that show abnormal behaviours for Upp show a 97% correlation with those that are manually classified as metabolic.

(dephosphorylation of uracil) are strongly indicative of a metabolic phenotype and can be used as a feature to separate metabolic disruption phenotypes from other types of phenotype, with a phi coefficient of 97%. Overall, the driver node analysis showed that it is possible to identify important reactions within the network that correlate with certain cell behaviours, meaning that we can focus on these to understand the end phenotype rather than the entire set of reactions.

DISCUSSION

We have shown multiple analysis methods that take a large high-dimensional dataset and distill it into visualisations that are easy to interpret. The pipeline of weak labelling followed by neural network classification is applicable to any system that outputs time series, although the features used for the initial labelled schema may have to be changed, according to what the researcher intends to look for, and the type of time series that is being analysed. As discussed previously, it is particularly useful for where "normal" behaviours for a system is not well defined, and the mechanisms that underlie the system cannot be distilled into a form that is understood. We have shown that it is applicable for black-box models, but it could also be used for data from complex physical systems where we do not understand the fundamental structure, such as meteorological phenomena. Additionally, the driver node analysis is applicable to any system where there is input,

output, and internal structure, as it can highlight the most important parts of a high-dimensional system.

The processing of complex data is imperative to understand whole-cell model output, and this method demonstrates how the behaviours of specific reactions can be used as a marker of a particular phenotypic class and their importance to the corresponding cellular process.

Understanding the effects of single gene knockouts is a deceptively difficult task, as the domino effect of gene removal can cause large changes in the behaviour of a cell through its life cycle. Visualising and analysing thousands of time series is a challenge faced by many branches of research. These two problems come together in the context of whole-cell models. Using Snorkel and neural networks, we have been able to classify metabolic fluxes as normal or abnormal and visualise them in two dimensions, meaning that the dataset separates into groups that can be interpreted. Whole-cell model data must be understood in the context of controllable biological mechanisms to be relevant to genome design: in order to use knowledge gained from modelling in real cells, we must understand the internal operations as well as the output. The flux behaviour across different gene knockouts, and in particular the driver nodes, can show the links between genotype and phenotype, plus unprecedented effects that a gene may have on reactions seemingly unrelated to its functional annotation, on a scale that is only possible in a whole-cell model. As this analysis gives an overview of the entire metabolism, we can approach the problem of understanding gene knockouts in a way that includes the genomic context of the remaining genes and the behaviour of their associated reactions, rather than examining the phenotype with regard to the single gene that has been removed.

The driver nodes can also give insight into the essentiality of *Mycoplasma* functions. Most of the driver reactions are not associated with annotated genes, as many transporter proteins are putative—however, given that *M. genitalium* synthesises very few compounds and gains most from its surrounding media, this is an important knowledge gap. The external media for *Mycoplasma* culture is generally undefined rich media, so knowledge of exactly which of the media components are essential for growth would be valuable for lab use and simplify *Mycoplasma* production (Gaspari et al., 2020). This may also help with linking un-annotated genes with modelled functions, leading to better understanding of the *M. genitalium* genome. For example, an essential protein in JCVI-syn3A [one of the first synthetic organisms; designed to function as a minimal cell (Breuer et al., 2019)] has recently been classified as a riboflavin transporter protein, showing that vitamin transport is an essential function for a minimal organism (Zhang et al., 2021). As *M. genitalium* does not synthesise riboflavin, this suggests that one of its un-annotated genes must be a riboflavin transporter. As more wet lab work is done with *M. genitalium*, it will be interesting to compare it to the model results and the importance of different driver nodes. The essentiality of similar transport reactions could also be looked at in other organisms, as these results may be applicable to other *Mycoplasma*s.

For genome design, there has long been an idea of "modularity" in cells, at different scales and abstractions (Papin et al., 2004). Cellular subsystems that use a unique set of molecules and rules to perform a function such as DNA replication or glycolysis use chemical specificity to keep their processes separate from other functional modules (Hartwell et al., 1999). It has been proposed recently that the future of genome design may be in minimal cells, combined with different functional modules to create cells for specific purposes (Gibson, 2014). This would require a detailed understanding of not only how a genome maps to its phenotype and how the genes themselves can form functional modules but also concerning the ways in which these modules interact. This is one of the main advantages of using a whole-cell model rather than a constraint-based model—from observing the behaviour of reactions, we can see how other mechanisms in the cell (e.g., DNA production) are affected, which we would not in a constraint-based model.

The metabolism submodel in the *M. genitalium* model is a central hub of activity and an integral stepping stone for substance transfer between cellular processes. Although internal mechanisms and local rules for the model were gathered from experimental data and are biologically valid, the complexity that arises from so many parameters being integrated together means that the model has to be treated as a black box. Analysing the behaviour of the model could ultimately lead to better biological understanding of the connections between cellular processes. If the way that two processes are coupled together *in silico* in the whole-cell model yields output that matches experimental data, this can help to develop insight into how these processes are linked in a real cell. This could aid genome design, where insights from modelling can rationally guide *in vitro* experiments and gene editing (Landon et al., 2019; Rees-Garbutt et al., 2020, Rees-Garbutt et al., 2021).

We can see from **Figure 3** that the knockouts that cause DNA and septum disruptions cause similar behaviours in the flux profiles to non-essential gene knockouts, likely because most of their reaction behaviours were classified as normal. **Supplementary Figure S3** shows that fewer than 10 reactions were consistently affected across the simulations within these phenotypic classes, so we can infer that these reactions might be the bridge between the metabolism process and the DNA replication or cytokinesis process. Limitations of the *M. genitalium* model mean that the results presented here do not include multiple cell divisions, and it is possible that more widespread effects on metabolism would be revealed in future work with more generations.

The interactions between the metabolism and the other phenotypic classes (protein and RNA) are less simple, as there are significantly more reactions that are consistently behaving abnormally. This is not surprising, as there are two main functions for a cell to perform: growth and replication. Growth occurs consistently through the cell cycle and requires constant synthesis and degradation of different proteins and RNAs. There is also a temporal element, as cascades of

reactions that form different proteins may need to occur in a specific order. Any disruption to an aspect of this process during the life cycle will filter down to other processes, whereas if DNA replication is disrupted, it is primarily cell division that will be halted. In future studies, it would be interesting to see if dividing the proteins into functional groups and pathways for further analysis leads to a better understanding of their roles and how they interact with each other.

It is hard to draw solid conclusions about cell behaviours, as *M. genitalium* is an organism where not all of the genes are classified, and the data that the model was built upon are from many different sources and organisms. In terms of the network analysis, there are some reactions that have been observed in *M. genitalium* but do not have known enzymes to catalyse them, which leaves gaps within the model. There may be unexpected and unusual behaviours that are not captured in the training data as well, leading to misclassifications; for example, the reactions that performed badly in the neural network classifications may be sensitive to small changes in the metabolic network, meaning that their behaviours are inconsistent and unpredictable. However, it is useful to flag these reactions and, in the future, to use different approaches to understand their behaviours. There is also the possibility that, after applying the machine learning processes, the results show more about the internal features of the model itself than the actual biology, which is a good starting point for lab work.

As whole-cell models become more widely used, analysis software will become more important. The most recent whole-cell model is of *E. coli* (Macklin et al., 2020), which is a better-understood organism than *M. genitalium*, with significantly more data available to validate and add to it, so this is an important development for the field. However, the complexity of models will increase hugely with the size of the genome of the organism, and as *E. coli* has an order of magnitude more genes than *M. genitalium* (Blattner et al., 1997), analysis tools that can provide data processing and dimensionality reduction will be even more important for enhancing understanding and ultimately genome design.

DATA AVAILABILITY STATEMENT

The datasets generated and analysed for this study can be found in the University of Bristol data repository https://doi.org/10.5523/bris.3u1v7dy42fk332watjl81m13y0. The code used for this publication can be found at https://doi.org/10.5523/bris.879w4p1r8iy32ef9vcwftq9ts.

AUTHOR CONTRIBUTIONS

CG, LM, OC, GB, and SL were involved in the ideation. OC designed the methodology to automatically classify metabolic reaction fluxes as normal and abnormal. GB manually labelled data, created the generative model, implemented Snorkel, and created neural networks as a proof of concept with help from OC. SL trained neural networks and classified the analysis dataset, carried out the network analysis, created the figures, and wrote the paper. CG, LM, and OC were involved in editing and feedback on paper. CG and LM supervised all the research activities.

FUNDING

LM is supported by the Medical Research Council grant MR/N021444/1, by the Engineering and Physical Sciences Research Council (grants EP/R041695/1 and EP/S01876X/1), and by the EU Horizon 2020 research project COSY-BIO (grant 766840). OC, LM, and CG are supported by a BrisSynBio, a BBSRC/EPSRC Synthetic Biology Research Centre (BB/L01386X/1), flexi-fund grant. OC is supported by the Bristol Centre for Complexity Sciences (BCCS) Centre for Doctoral Training (CDT) EP/I013717/1. SL is supported by EPSRC Future Opportunity Scholarships.

ACKNOWLEDGMENTS

We would like to thank the Advanced Computing Research Centre (ACRC) and BrisSynBio, a BBSRC/EPSRC Synthetic Biology Research Centre, at the University of Bristol, for access to the BlueCrystal and Bluegem supercomputers. Special thanks to the HPC and RDSF teams of the ACRC, particularly Christopher Woods, Simon Burbidge, Matt Williams, and Damian Steer, for their help with BlueCrystal, BlueGem, data storage, and publication. We would like to thank Thomas Gorochowski (University of Bristol) for useful feedback on the manuscript. We would like to thank Joshua Rees-Garbutt for generating the single-gene KO simulations we analysed.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.732079/full#supplementary-material

REFERENCES

Barabási, A.-L., and Oltvai, Z. N. (2004). Network Biology: Understanding the Cell's Functional Organization. *Nat. Rev. Genet.* 5, 101–113. doi:10.1038/nrg1272

Barandas, M., Folgado, D., Fernandes, L., Santos, S., Abreu, M., Bota, P., et al. (2020). Tsfel: Time Series Feature Extraction Library. *SoftwareX* 11, 100456. doi:10.1016/j.softx.2020.100456

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., et al. (1997). The Complete Genome Sequence of escherichia Coli K-12. *science* 277, 1453–1462. doi:10.1126/science.277.5331.1453

Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014). Constraint-based Models Predict Metabolic and Associated Cellular Functions. *Nat. Rev. Genet.* 15, 107–120. doi:10.1038/nrg3643

Breuer, M., Earnest, T. M., Merryman, C., Wise, K. S., Sun, L., Lynott, M. R., et al. (2019). Essential Metabolism for a Minimal Cell. *Elife* 8, e36842. doi:10.7554/eLife.36842

Caruana, R., and Niculescu-Mizil, A. (2006). "An Empirical Comparison of Supervised Learning Algorithms," in Proceedings of the 23rd international conference on Machine learning, 161–168. doi:10.1145/1143844.1143865

Correll, M., Albers, D., Franconeri, S., and Gleicher, M. (2012). "Comparing Averages in Time Series Data," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1095–1104. doi:10.1145/2207676.2208556

Correll, M., and Heer, J. (2017). "Regression by Eye: Estimating Trends in Bivariate Visualizations," in Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 1387–1396.

Domingos, P. (2012). A Few Useful Things to Know about Machine Learning. *Commun. ACM* 55, 78–87. doi:10.1145/2347736.2347755

Dybvig, K., and Voelker, L. L. (1996). Molecular Biology of Mycoplasmas. *Annu. Rev. Microbiol.* 50, 25–57. doi:10.1146/annurev.micro.50.1.25

Ekström, J. (2011). *The Phi-Coefficient, the Tetrachoric Correlation Coefficient, and the pearson-yule Debate*.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., et al. (2007). A Genome-scale Metabolic Reconstruction for *Escherichia coli* K-12 MG1655 that Accounts for 1260 ORFs and Thermodynamic Information. *Mol. Syst. Biol.* 3, 121. doi:10.1038/msb4100155

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry* 40, 35–41. doi:10.2307/3033543

Gaspari, E., Malachowski, A., Garcia-Morales, L., Burgos, R., Serrano, L., Martins Dos Santos, V. A. P., et al. (2020). Model-driven Design Allows Growth of Mycoplasma Pneumoniae on Serum-free media. *NPJ Syst. Biol. Appl.* 6, 33–11. doi:10.1038/s41540-020-00153-7

Gibson, D. G. (2014). Programming Biological Operating Systems: Genome Design, Assembly and Activation. *Nat. Methods* 11, 521–526. doi:10.1038/nmeth.2894

Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., et al. (2014). *Data Integration in the Era of Omics: Current and Future Challenges*.

Gosak, M., Markovič, R., Dolenšek, J., Slak Rupnik, M., Marhl, M., Stožer, A., et al. (2018). Network Science of Biological Systems at Different Scales: a Review. *Phys. Life Rev.* 24, 118–135. doi:10.1016/j.plrev.2017.11.003

Haimovich, A. D., Muir, P., and Isaacs, F. J. (2015). Genomes by Design. *Nat. Rev. Genet.* 16, 501–516. doi:10.1038/nrg3956

Hannan, E. J. (2009). *Multiple Time Series*, Vol. 38. John Wiley & Sons.

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From Molecular to Modular Cell Biology. *Nature* 402, C47–C52. doi:10.1038/35011540

Ho, T. K. (1995). "Random Decision Forests," in Proceedings of 3rd international conference on document analysis and recognition (IEEE), 278–282.1.

Holme, P. (2011). Metabolic Robustness and Network Modularity: a Model Study. *PloS one* 6, e16605. doi:10.1371/journal.pone.0016605

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., and Muller, P.-A. (2019). Deep Learning for Time Series Classification: a Review. *Data Min Knowl Disc* 33, 917–963. doi:10.1007/s10618-019-00619-1

Kanehisa, M., and Goto, S. (2002). *Novartis Foundation Symposium*. Wiley Online Library, 91–100.The Kegg Database.

Karr, J., and Pochiraju, S. (2018). *Wholecell.org*. Available at: http://www.wholecell.org/tools/". Online(accessed 03 13, 20).

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, B., Jr, et al. (2012). A Whole-Cell Computational Model Predicts Phenotype from Genotype. *Cell* 150, 389–401. doi:10.1016/j.cell.2012.05.044

Kröse, B., Krose, B., van der Smagt, P., and Smagt, P. (1993). *An Introduction to Neural Networks*.

Kuriya, Y., and Araki, M. (2020). Dynamic Flux Balance Analysis to Evaluate the Strain Production Performance on Shikimic Acid Production in escherichia Coli. *Metabolites* 10, 198. doi:10.3390/metabo10050198

Landon, S., Rees-Garbutt, J., Marucci, L., and Grierson, C. (2019). Genome-driven Cell Engineering Review: *In Vivo* and In Silico Metabolic and Genome Engineering. *Essays Biochem.* 63, 267–284. doi:10.1042/ebc20180045

Lee, R., Karr, J. R., and Covert, M. W. (2013). Wholecellviz: Data Visualization for Whole-Cell Models. *BMC bioinformatics* 14, 253. doi:10.1186/1471-2105-14-253

Leighty, R. W., and Antoniewicz, M. R. (2011). Dynamic Metabolic Flux Analysis (Dmfa): a Framework for Determining Fluxes at Metabolic Non-steady State. *Metab. Eng.* 13, 745–755. doi:10.1016/j.ymben.2011.09.010

Lequeux, G., Beauprez, J., Maertens, J., Van Horen, E., Soetaert, W., Vandamme, E., et al. (2010). Dynamic Metabolic Flux Analysis Demonstrated on Cultures where the Limiting Substrate Is Changed from Carbon to Nitrogen and Vice Versa. *J. Biomed. Biotechnol.* 2010, 621645. doi:10.1155/2010/621645

Liu, Y.-Y., Slotine, J.-J., and Barabási, A.-L. (2011). Controllability of Complex Networks. *nature* 473, 167–173. doi:10.1038/nature10011

Macklin, D. N., Ahn-Horst, T. A., Choi, H., Ruggero, N. A., Carrera, J., Mason, J. C., et al. (2020). Simultaneous cross-evaluation of heterogeneous e. coli datasets via mechanistic simulation. *Science* 369, eaav3751. doi:10.1126/science.aav3751

Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., et al. (2018). Genome, Transcriptome and Proteome: the Rise of Omics Data and Their Integration in Biomedical Sciences. *Brief. Bioinformatics* 19, 286–302. doi:10.1093/bib/bbw114

Matsuda, T., and Komaki, F. (2017). Time Series Decomposition into Oscillation Components and Phase Estimation. *Neural Comput.* 29, 332–367. doi:10.1162/neco_a_00916

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426.

Nacher, J. C., Ishitsuka, M., Miyazaki, S., and Akutsu, T. (2019). Finding and Analysing the Minimum Set of Driver Nodes Required to Control Multilayer Networks. *Sci. Rep.* 9, 576–612. doi:10.1038/s41598-018-37046-z

Noble, W. S. (2006). What Is a Support Vector Machine? *Nat. Biotechnol.* 24, 1565–1567. doi:10.1038/nbt1206-1565

Northcutt, C. G., Jiang, L., and Chuang, I. L. (2019). Confident Learning: Estimating Uncertainty in Dataset Labels. arXiv preprint arXiv:1911.00068.

O'Brien, E. J., Lerman, J. A., Chang, R. L., Hyduke, D. R., and Palsson, B. Ø. (2013). Genome-scale Models of Metabolism and Gene Expression Extend and Refine Growth Phenotype Prediction. *Mol. Syst. Biol.* 9, 693. doi:10.1038/msb.2013.52

Palsson, B. Ø. (2006). *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press.

Papin, J., Reed, J., and Palsson, B. (2004). Hierarchical Thinking in Network Biology: the Unbiased Modularization of Biochemical Networks. *Trends Biochemical Sciences* 29, 641–647. doi:10.1016/j.tibs.2004.10.001

Raczko, E., and Zagajewski, B. (2017). Comparison of Support Vector Machine, Random forest and Neural Network Classifiers for Tree Species Classification on Airborne Hyperspectral apex Images. *Eur. J. Remote Sensing* 50, 144–154. doi:10.1080/22797254.2017.1299557

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. (2019). Snorkel: Rapid Training Data Creation with Weak Supervision. *VLDB J.* 29 (2), 709–730. doi:10.1007/s00778-019-00552-1

Ratner, A. J., Bach, S. H., Ehrenberg, H. R., and Ré, C. (2017). "Snorkel: Fast Training Set Generation for Information Extraction," in Proceedings of the 2017 ACM international conference on management of data, 1683–1686.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Baraba´si, A.-L. (2002). Hierarchical Organization of Modularity in Metabolic Networks. *science* 297, 1551–1555. doi:10.1126/science.1073374

Rees-Garbutt, J., Chalkley, O., Landon, S., Purcell, O., Marucci, L., and Grierson, C. (2020). Designing Minimal Genomes Using Whole-Cell Models. *Nat. Commun.* 11, 836–912. doi:10.1038/s41467-020-14545-0

Rees-Garbutt, J., Rightmyer, J., Chalkley, O., Marucci, L., and Grierson, C. (2021). Testing Theoretical Minimal Genomes Using Whole-Cell Models. *ACS Synth. Biol.* 10, 1598–1604. doi:10.1021/acssynbio.0c00515

Robinson, J., Jegelka, S., and Sra, S. (2020). "Strength from Weakness: Fast Learning Using Weak Supervision," in International Conference on Machine Learning (PMLR), 8127–8136.

Sarle, W. S. (1994). *Neural Networks and Statistical Models*.

Saunshi, N., Plevrakis, O., Arora, S., Khodak, M., and Khandeparkar, H. (2019). "A Theoretical Analysis of Contrastive Unsupervised Representation Learning," in International Conference on Machine Learning, Long Beach, CA (PMLR), 5628–5637.

Smart, A. G., Amaral, L. A. N., and Ottino, J. M. (2008). Cascading Failure and Robustness in Metabolic Networks. *Proc. Natl. Acad. Sci.* 105, 13223–13228. doi:10.1073/pnas.0803571105

Stelling, J., Klamt, S., Bettenbrock, K., Schuster, S., and Gilles, E. D. (2002). Metabolic Network Structure Determines Key Aspects of Functionality and Regulation. *Nature* 420, 190–193. doi:10.1038/nature01166

Van der Maaten, L., and Hinton, G. (2008). Visualizing Data Using T-Sne. *J. machine Learn. Res.* 9.

Wang, Z., Yan, W., and Oates, T. (2017). "Time Series Classification from Scratch with Deep Neural Networks: A strong Baseline," in 2017 International joint conference on neural networks (IJCNN) (IEEE), 1578–1585. doi:10.1109/ijcnn.2017.7966039

Wetterstrand, K. (2010). Dna Sequencing Costs: Data from the Nhgri Genome Sequencing Program (Gsp). "Available at: www.genome.gov/sequencingcostsdata". Online: accessed: 13-03-20.

Wong, E., Baur, B., Quader, S., and Huang, C.-H. (2012). Biological Network Motif Detection: Principles and Practice. *Brief. Bioinformatics* 13, 202–215. doi:10.1093/bib/bbr033

Yegnanarayana, B. (2009). *Artificial Neural Networks*. New Delhi: PHI Learning Pvt. Ltd.

You, C. H., Holder, L. B., and Cook, D. J. (2006), Application of Graph-Based Data Mining to Metabolic Pathways. in Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06) IEEE, 169–173. doi:10.1109/icdmw.2006.31

Yugi, K., Ohno, S., Krycer, J. R., James, D. E., and Kuroda, S. (2019). Rate-oriented Trans-omics: Integration of Multiple Omic Data on the Basis of Reaction Kinetics. *Curr. Opin. Syst. Biol.* doi:10.1016/j.coisb.2019.04.005

Zampieri, G., Vijayakumar, S., Yaneske, E., and Angione, C. (2019). Machine and Deep Learning Meet Genome-Scale Metabolic Modeling. *Plos Comput. Biol.* 15, e1007084. doi:10.1371/journal.pcbi.1007084

Zhang, C.-T., and Zhang, R. (2008), Gene Essentiality Analysis Based on Deg, a Database of Essential Genes. *Microbial Gene Essentiality: Protocols and Bioinformatics*. Springer, 391–400. doi:10.1007/978-1-59745-321-9_27

Zhang, C., Zheng, W., Cheng, M., Omenn, G. S., Freddolino, P. L., and Zhang, Y. (2021). Functions of Essential Genes and a Scale-free Protein Interaction Network Revealed by Structure-Based Function and Interaction Prediction for a Minimal Genome. *J. Proteome Res.* 20, 1178–1189. doi:10.1021/acs.jproteome.0c00359

Zhou, Z.-H. (2018). A Brief Introduction to Weakly Supervised Learning. *Natl. Sci. Rev.* 5, 44–53. doi:10.1093/nsr/nwx106

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Landon, Chalkley, Breese, Grierson and Marucci. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 20 December 2021
doi: 10.3389/fmolb.2021.760077

A Novel Hybrid Logic-ODE Modeling Approach to Overcome Knowledge Gaps

Gianluca Selvaggio[1†], *Serena Cristellon*[1,2†] *and Luca Marchetti*[1,3*]

[1]*Piazza Manifattura, Fondazione The Microsoft Research-University of Trento Centre for Computational and Systems Biology (COSBI), Rovereto, Italy,* [2]*Department of Mathematics, University of Trento, Trento, Italy,* [3]*Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, Trento, Italy*

Mathematical modeling allows using different formalisms to describe, investigate, and understand biological processes. However, despite the advent of high-throughput experimental techniques, quantitative information is still a challenge when looking for data to calibrate model parameters. Furthermore, quantitative formalisms must cope with stiffness and tractability problems, more so if used to describe multicellular systems. On the other hand, qualitative models may lack the proper granularity to describe the underlying kinetic processes. We propose a hybrid modeling approach that integrates ordinary differential equations and logical formalism to describe distinct biological layers and their communication. We focused on a multicellular system as a case study by applying the hybrid formalism to the well-known Delta-Notch signaling pathway. We used a differential equation model to describe the intracellular pathways while the cell–cell interactions were defined by logic rules. The hybrid approach herein employed allows us to combine the pros of different modeling techniques by overcoming the lack of quantitative information with a qualitative description that discretizes activation and inhibition processes, thus avoiding complexity.

Keywords: hybrid modeling, logic modeling, ordinary differential equations (ODEs), computational systems biology, simulation algorithms

OPEN ACCESS

Edited by:
Ernesto Perez-Rueda, Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Francesco Delfino, INSERM U1054 Centre de Biochimie Structurale de Montpellier, France
Luis Diambra, National University of La Plata, Argentina

***Correspondence:**
Luca Marchetti
marchetti@cosbi.eu
luca.marchetti@unitn.it

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *17 August 2021*
Accepted: *09 November 2021*
Published: *20 December 2021*

Citation:
Selvaggio G, Cristellon S and Marchetti L (2021) A Novel Hybrid Logic-ODE Modeling Approach to Overcome Knowledge Gaps. Front. Mol. Biosci. 8:760077. doi: 10.3389/fmolb.2021.760077

INTRODUCTION

Computational models have become a cornerstone of modern biology as a tool for data interpretation and serving in parallel with experimental techniques to disentangle process complexity (Markowetz, 2017). Depending on the available information and the addressed questions, we can model the processes with different approaches, in function of model granularity and abstraction level (Ideker and Lauffenburger, 2003). Biological processes for which are available high-throughput omics data can be described using interaction networks (Danaher et al., 2014; Hawe et al., 2019); these can be integrated with other experimental evidences (such as knock-out experiments) to provide directed graphs (Gross et al., 2019). By increasing the biological knowledge of the processes involved, providing the sign of the interactions, we can obtain a regulatory graph that together with a set of rules for each component defines a logic model (Abou-Jaoudé et al., 2016).

The logic formalism is the simplest way to model interactions among entities, in a parameter-free fashion, and has been used since the 1970s to qualitatively describe biological pathways (Kauffman,

1969), and intracellular and intercellular signaling networks (Gonzalez et al., 2008; Morris et al., 2010) up to collective cell behaviour (Varela et al., 2018). In the case that the pathways and their components are thoroughly characterized, ordinary and partial differential equations (ODEs and PDEs) can be used instead (Aldridge et al., 2006). If the number of molecules involved in the processes does not meet the Continuum hypothesis requirements, then the stochastic methods allow to overcome the problem (Simoni et al., 2019).

Selecting the appropriate mathematical formalism to describe the biology is thus a trade-off between the a priori available knowledge (e.g., parameters, concentrations etc.) and the required granularity to address the biological problem.

Another important aspect to consider in systems biology is the interplay between different biological layers, as most models focus on a single scale. Progresses have been made in implementing hierarchical representations to study how local variations may affect the dynamics at other levels. An example of this approach is the work of Uluseker et al. (2018) in which the authors built an ODE model that integrates in a holistic framework the glucose homeostasis together with other regulatory hormones at different levels: gut, liver, and adipose tissue. However, models that span over different levels (e.g., from subcellular to tissue) are difficult to parametrize and implement. Model combination (Palsson et al., 2013) can be a solution but it is a challenging task due to the non-modular implementation of these mathematical frameworks, and goes beyond simple coupling of the equations.

The technical difficulties of model integration, arising from the different modeling formalisms, have been tackled by recasting the mathematical descriptions to a single approach. Ryll et al. (2014), with their model of hormonal regulation of glucose homeostasis, proposed a strategy to integrate a logic model of signaling network with an ODE model of metabolic processes: the Boolean representation was converted into a set of logic-based ODEs. The integration required a calibration step in which the added parameters and the missing ones of the kinetic model were fitted to experimental data.

As outlined in Uluseker et al. (2018), biological phenomena interlay different abstraction levels, where interconnected modules form complex collective behaviour. Hierarchical models, as mentioned previously, are used to provide a structured holistic representation of complex biological systems. Single models communicate, through feedbacks, at the systemic level producing the macroscopic behavior. Often, for these interactions, only limited knowledge or qualitative measurements are available and thus a complete ODE description is hindered. Here, we propose an integrative approach that leverages on different formalisms (ODE and logic), with a fine-grained ODE representation of the bottom layer to properly describe the variable dynamics, and the logic formalism to represent in a coarse-grained fashion the regulative interactions. This approach does not require a model re-parametrization or recasting to a common description thus enabling model reuse.

Although the modeling approach is the focus of the present work, we decided to convey our strategy by presenting a case study: the Delta-Notch signaling pathway.

Delta-Notch signaling is among the most conserved pathways in tissue development based on the negative-feedback loop between the two elements (Artavanis-Tsakonas et al., 1999). Upon Delta ligand binding to the Notch receptor of another cell, a response is triggered leading the receiving cell to repress Delta, governing fate selection. Several examples of Delta-Notch salt-and-pepper patterning are present in nature like in *Drosophila* (Renaud and Simpson, 2001; De Joussineau et al., 2003), as well as in the mouse inner ear (Hartman et al., 2010), and mouse and zebrafish retina (Del Bene et al., 2008).

One of the first models investigating the Delta-Notch signaling pathway, developed by Collier et al. (1996), used the ODE formalism to qualitatively describe the dynamics of active Notch and Delta between adjacent cells. The input of a generic cell was modeled as the average of all neighbor Deltas. The intracellular Notch activation and consequent fate decision were described using a phenomenological Hill function. Agrawal et al. (2009) adopted a fine-grained approach describing with an ODE system the intracellular processes of cleavage, transcription, translation, transport, and degradation, after Notch activation. The work focused on the analysis of the single cell fate decision, rather than the pattern formation, highlighting the possibility of a phenotypic switch from bistable system to oscillatory, by tuning a single parameter. Mjolsness et al. (1991) and Marnellos and Mjolsness (1998) modeled neuroblasts and sensory organ precursor cell differentiation in *Drosophila*, as nodes in a recurrent neural network. Cells are represented as discrete entities, which can interact with neighbors. A minimal two-gene network was used, allowing interaction with other gene products from within the same cell or from neighboring ones. Varela et al. (2018) developed a 2D logical model of lateral inhibition, using the software Epilog. Each cell of the discrete tissue contains a two-component logic model that responds to the input coming from the neighboring cells. Both Varela et al. and Marnellos et al. approaches allow to simulate pattern formation at the tissue level without in depth knowledge of the kinetic parameters or species concentrations. Also, agent-based models (ABM) have been employed to address how complex behaviors arise from the cell–cell interactions (Reynolds et al., 2019) or cell–environment interaction (Yu and Bagheri, 2020). In particular, Reynolds et al. (2019) developed an ABM that recreates Delta-Notch patterns using for each agent a set of rules that define the increment of each species, thus providing a more abstract view.

The hybrid strategy we propose defines a semi-quantitative framework optimal to simulate tissue level dynamics with fixed interacting cells. The implementation, at the lower level, of an ODE-based model generates a quantitative time that allows to better appreciate the grid evolution. Complex behavior and spatial effects can be implemented including in the logical rules more than the first line of neighbors and by changing the geometry of the grid (i.e., cylindrical, toroidal etc.).

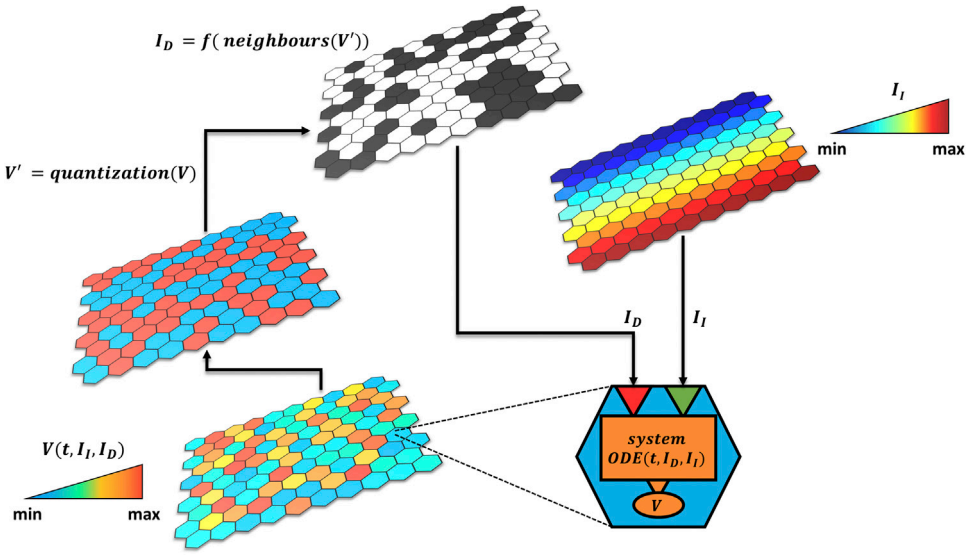


FIGURE 1 | Hybrid model approach. The tissue is composed by a grid of quantitative modules described by an ODE system depending on two types of inputs I_D and I_I, respectively dependent and independent from the other modules. The former is usually cast as a function of the internal variable (V) of neighboring cells. Thresholding the Vs of all cells will generate a logical matrix (V'), which is then used to compute I_D by applying a logical rule f. I_I identifies instead those inputs that are only dependent from the spatial position.

Materials and Methods

The intracellular signaling cascade model parameterization is provided in **Supplementary Material S1**. All computations were implemented and performed in MATLAB (R2019b); simulations were performed with an Intel Core i7-8700T processor, CPU 2.40 GHz and installed RAM 16.00 GB. Numerical integration of the ODE system was made using the ODE solver ode15s.

RESULTS

Hybrid Modeling Approach

In this work, we propose a hybrid modeling approach to describe complex biological phenomena where the lack of kinetic parameters, species concentrations, or mechanistic knowledge hinders a complete ODE description.

The model is built hierarchically in a bottom-up fashion (**Figure 1**). At the lower level, there is a matrix of quantitative single modules (i.e., signaling pathways, cells) described by a system of ordinary differential equations characterized by a set of variables. Each module can receive two types of inputs: independent (I_I) or dependent (I_D) from the other modules. The latter is a logical variable (Boolean or multivalued) which describes the interactions between the single modules or environmental feedbacks (i.e., pathway crosstalk, extracellular signaling) and encodes through a logical rule the contribution of the neighboring modules' output variables (V). The independent input, I_I, is used instead to portrait those signals that are only position dependent, as diffusive molecules or other environmental cues.

The hybrid approach we propose thus puts into communication the two layers (intracellular and tissue) and can be simulated as illustrated by the pseudo-code of **Figure 2**.

While I_I is uniquely defined by the cell position, I_D is locally defined by the variable V of the neighboring modules. To define each module's I_D, we first threshold V for all the modules, generating the logic matrix V'; we compute I_D by applying the logical rule f on V', considering a specific neighborhood of the module of the grid in exam. Given the planar representation and the finite number of grid elements, border cells may have a lower number of neighbors when compared with the others; it is thus important to define boundary conditions to mitigate artifact effects. Possible strategies that we also apply in the context of the case study, are cylindrical conditions, where two borders of the grid are put in contact by a single fold of the grid or toroidal conditions, where a double fold of the grid put in pairwise contact the grid borders. The modules are then integrated until the variable V crosses the quantization threshold ($t_{threshold}$).

The process is then repeated until the break condition is met; this can be grid equilibrium, maximum allowed simulation time, or other ad hoc constrains.

CASE STUDY: DELTA-NOTCH

Delta-Notch is a highly conserved cell–cell communication pathway present in most animals; it allows cells to select different fates based upon the neighborhood consensus. In *Drosophila*, during the neuronal development phase, cellular differentiation gives rise to salt-and-pepper patterns with cells either reaching neuronal fate or not (Campos-Ortega, 1995; Bertrand et al., 2002). The phenomenon of lateral inhibition among adjacent cells, mediated by Delta-Notch signaling pathway, has a major role in this kind of pattern formation. Depending on the interconnectivity of the network, different

Inputs: N ODE-modules with output variable V and independent input (I_I), Logical rule (f), break condition, $V_{threshold}$

Output: modules variables time series

$while(not(break\ condition))$

$V' \leftarrow V > V_{threshold}$

$I_D \leftarrow f(neighbours(V'))$

$Integrate\ N\ modules\ until\ t_{threshold}, where\ V(i) > V_{threshold}\ \ i \in \{1:N\}$

end

FIGURE 2 | Hybrid model simulation pseudo-code, an extended version of the code is available in **Supplementary Material S1**.

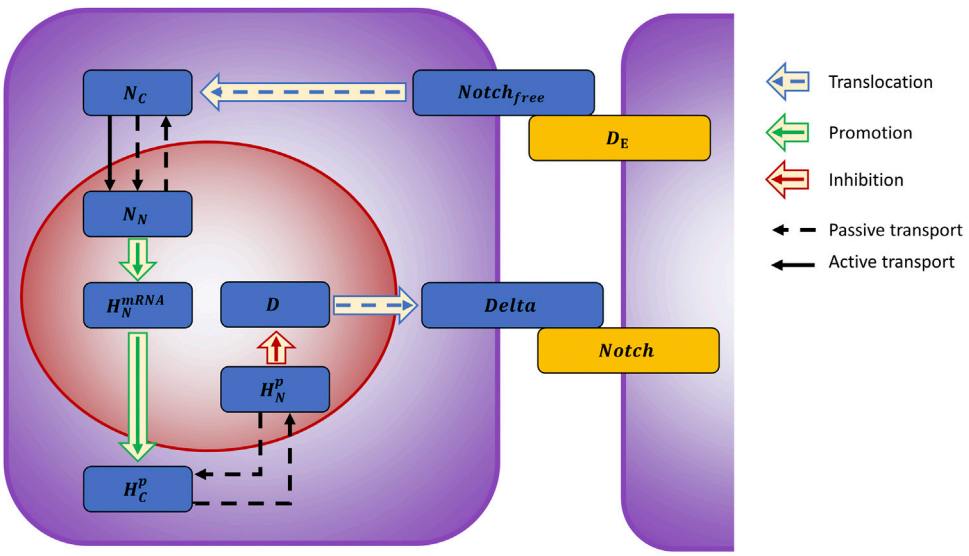


FIGURE 3 | Graphical representation of the model describing the Delta-Notch pathway. Upon binding to the Delta ligand (D_E) from another cell (yellow rectangle) the Notch receptor releases its intracellular domain in the cytoplasm (N_C). After entering the nucleus, either due to passive or active transport, Notch (N_N) acts as transcription factor inducing the expression of Hes mRNA (H_N^{mRNA}) and its translation into Hes protein (H_C^p). H_C^p migrates into the nucleus (H_N^p) where it acts as repressor of Delta expression (D) leading to a decrease in Delta protein at the membrane. Coupling this system with its counterpart in a neighbor cell will lead to having a Delta$^+$ cell and a Delta$^-$ cell, respectively expressing or not the protein. The model consists of two cell compartments: nuclear (red) and cytoplasmatic (violet).

patterns can arise, and since cells are observed to extend protrusions, even non-adjacent cells can interact (Renaud and Simpson, 2001; Hadjivasiliou et al., 2016).

In the following sections, we will describe how we developed and integrated the intracellular ODE model and the intercellular logic model.

Intracellular Signaling Cascade

We developed a mathematical model describing the Notch intracellular pathway, building upon the computational work of Agrawal et al. (2009) (for details see **Supplementary Material S1**). A set of ordinary differential equations quantitatively trace the different components involved in the processes, following the Delta-Notch binding on the outer membrane (**Figure 3**).

The cell is modeled with two compartments: cytoplasm (volume V_C) and nucleus (volume V_N). Every cell receives as external input the ligand Delta (D_E) from other cells; we considered it binary: either present ($D_E = D_E^{max}$) or not ($D_E = 0$). Upon binding of D_E with free Notch ($Notch_{free}$), Notch intracellular domain is cleaved and released in the cytoplasm (N_C) with a rate k_{cl}. Assuming that Notch expression is maintained constant (N_{tot}) the free amount of Notch on the membrane as function of the other species is:

$$Notch_{free} = N_{tot} - N_C - N_N \cdot \frac{V_N}{V_C}$$

N_C has a molecular weight of 110 kDa, thus it can permeate the nuclear envelope in two modalities: passive and active transport. The former is due to concentration gradient ($N_N - N_C$) and

occurs at rate k_{tr}^{P}. Active transport is modeled with a first-order kinetic (see Supplementary material) with transport rate $\mathrm{k_{tr}^{A}}$. The differential equations describing N_C and N_N are:

$$\frac{dN_C}{dt} = \mathrm{k_{cl}} \cdot Notch_{free} \cdot D_E + \frac{k_{tr}^{P}}{V_C} \cdot (N_N - N_C) - \frac{\mathrm{k_{tr}^{A}}}{V_C} \cdot N_C - \mu_N \cdot N_C$$

$$\frac{dN_N}{dt} = \frac{k_{tr}^{P}}{V_N} \cdot (N_C - N_N) + \frac{\mathrm{k_{tr}^{A}}}{V_N} \cdot N_C - \mu_N \cdot N_N$$

H_N^{mRNA} represents the concentration of Hes-mRNA in the nucleus, while H_C^P and H_N^P are Hes protein concentrations in the cytoplasm and in the nucleus, respectively. We described their dynamics as follows:

$$\frac{dH_N^{mRNA}}{dt} = \frac{k_{tH} \cdot N_N^2}{N_N^2 + K_H^2} - \mu_{H^{mRNA}} \cdot H_N^{mRNA},$$

$$\frac{dH_C^P}{dt} = k_{H^P} \cdot H_N^{mRNA} \cdot \frac{V_N}{V_C} + \frac{k_{tr}^{P}}{V_C} \cdot (H_N^P - H_C^P) - \mu_{H^P} \cdot H_C^P,$$

$$\frac{dH_N^P}{dt} = -\frac{k_{tr}^{P}}{V_N} \cdot (H_N^P - H_C^P) - \mu_H^P \cdot H_N^P$$

Transcription of Hes-mRNA was modeled with an activation Hill function $\frac{k_{tH} \cdot N_N^2}{N_N^2 + K_H^2}$, with coefficient equal to 2, and a maximal transcription rate k_{tH}. Hes-mRNA nuclear export was assumed to occur instantaneously and translation into Hes protein occurs at rate k_{H^P}. Hes1 has a molecular weight of ~30 kDa, hence we model nuclear permeation only due to concentration gradient $(H_N^P - H_C^P)$, with rate k_{tr}^{P}.

Delta transcription is inhibited by H_N^P, thus we modeled it with a repression Hill function $\frac{k_{tD} \cdot K_D^2}{K_D^2 + H_{pN}^2}$, with a maximal transcription rate k_{tD}. To simplify our system, we considered Delta mRNA as the final read out, implying that the translation process and protein maturation will simply add a delay to the transcriptional response.

$$\frac{dD}{dt} = \frac{k_{tD} \cdot K_D^2}{K_D^2 + H_{pN}^2} - \mu_D \cdot \mathrm{D}.$$

All components face degradation, with rates μ_N, $\mu_{H^{mRNA}}$, μ_H^P, and μ_D, where subscripts indicate the respective species.

Intercellular Signal Communication

The logic formalism was used to describe cell–cell communications during the lateral inhibition process. The epithelium was represented by a two-dimensional grid of cells (C_i), characterized by the intracellular concentration of Delta (D) binarized accordingly to a fixed threshold (Th):

$$C_i = \begin{cases} 1 & \text{if } D \geq Th \\ 0 & \text{if } D < Th. \end{cases}$$

Delta positive cells influence the neighbors' fate, activating their Notch pathway. The presence of Delta ligand (D_E) for the cell C_i is a logical function of N neighboring cells.

$$D_E(C_1, \ldots, C_N) = \begin{cases} D_E^{\max} & \text{if } \sum_{j=1}^{N} C_j^{dist} \geq \vartheta, \\ 0 & \text{otherwise.} \end{cases}$$

where ϑ represents the minimum number of the N neighboring cells required to be Delta positive, and *dist* indicates the integer distance at which a cell is considered neighbour.

Model Integration and Simulations

To simulate the pattern development, we integrated the two models. The tissue is represented by a two-dimensional grid of hexagonal cells, each containing the fine-grained kinetic model of **Figure 3**. Cell–cell communication is instead implemented with logical rules.

To bridge the two representations, for each cell the input of the ODE model (D_E) is defined according to the state of the neighboring cells via a logical rule. To replicate the salt-and-pepper pattern observed in biological contexts like the *Drosophila* neuronal development, we used the following rule:

$$D_E = \begin{cases} D_E^{\max} & \text{if } \sum_{i=1}^{6} C_i^1 \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

This implies that a cell receives as input $D_E = D_E^{max}$ if at least one of the six most proximal neighbors ($dist = 1$) is delta positive. After initializing the variables of each cell (**Figures 4A**), the simulation follows the steps described previously until the equilibrium of the grid is reached. To simulate the pattern emergence, we started from random initial conditions, and we selected independently for each cell the species concentrations from their biological range.

Interestingly, in our simulation when starting with random initial conditions (**Figures 4A**) the grid evolves first toward a naïve state, with no fate decided (**Figures 4B**), then it converges to the salt-and-pepper pattern accordingly to the selected rule.

Moreover, as illustrated in **Figure 5**, since the single ODE systems are responsible for the evolution of the epithelial grid, it is possible to observe the dynamics of each module. This approach thus provides multiple levels of information in function of the investigation objectives, allowing to pass from the collective cell behavior to the single intracellular dynamics.

Different rules for D_E can include more than one circle of neighbors (distance greater than one), giving rise to a variety of different patterns (**Figure 6**).

In a cell, several signaling pathways concur to fate decision, in addition to Delta-Notch (I_D) we can consider the positional input Wnt (I_I). This soluble protein can affect Notch signaling during fate decisions by diffusing in the tissue and activating a concentration-dependent inhibition of the Notch intracellular domain transcriptional activity (Collu et al., 2012). Wnt was integrated with a fixed concentration dependent on the cell position on the grid (**Figure 7**). This test case also provides an example of our modeling approach considering an external input not dependent on neighbors.

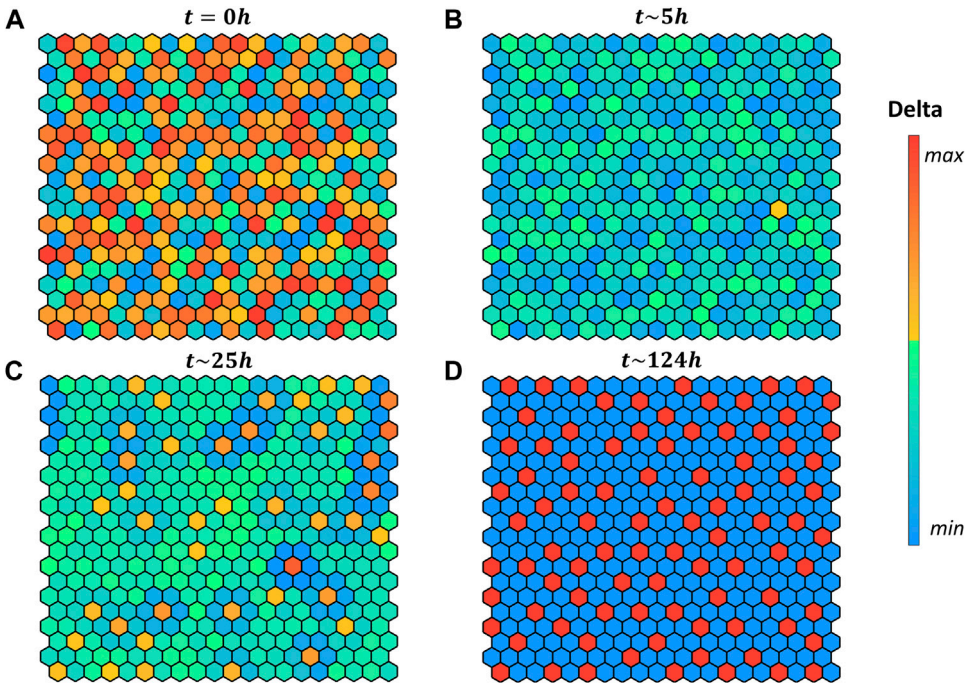


FIGURE 4 | Pattern evolution of a 20 × 20 grid with cylindrical boundary conditions (lateral borders are in contact) using the logical rule $D_E = D_E^{max}$ if $\sum_{i=1}^{6} C_i^1 \geq 1$, encoding that the input is present if at least one of the six neighbors at distance one are Delta positive. Colors indicate the intracellular delta concentration according to the color bar. **(A)** Initial values of the variables are set randomly within the biological boundaries. **(B)** All cells in the grid become Delta negative. **(C)** Delta-positive cells start to emerge and affect their neighbour's fate. **(D)** Equilibrium grid is reached. The hours reported above each grid indicates the quantitative time obtained from the ODE systems.

DISCUSSION

The description of a multicellular system is always a trade-off between complexity and tractability. Using a fine-grained approach as ODEs (or PDEs) provides a detailed representation of the system, which can be quantitatively used to understand the underpinning mechanism at the core of the biological processes. However, to be able to simulate such models, a large number of parameters are required, eventually leading to parametrization issues. Moreover, moving from a single cell model to a multicellular one will generate a rapid increase in the differential equation number, thus opening the door to integrability and running time problems. On the other hand, parameter-free approaches (such as logic modeling), although able to overcome information gaps, allow only for a qualitative representation of the system behavior. This kind of interpretation can be sometimes insufficient to provide useful insights on biological problems or help to analyze experimental evidence. Here, we introduce a new hybrid approach that leverages on these two formalisms to produce a semi-quantitative representation of a multicellular/tissue environment. Our approach can be also employed to describe interactions among pathways, as bridging signaling cascades by representing the kinase activity as a Boolean variable, or between organs as describing with an ODE system the glucose metabolism and with a logical variable the insulin presence.

To showcase the hybrid formalism, we selected Delta-Notch signaling pathway (Artavanis-Tsakonas et al., 1999) and the consequent cell fate selection in an epithelium (Renaud and Simpson, 2001; De Joussineau et al., 2003). As previously mentioned, different approaches and different granularities have been used to investigate this problem: from coarse-grained intercellular models as the one of Collier et al. (1996) to fine detailed intracellular models as the one of Agrawal et al. (2009). The latter provided an in-depth quantitative description of the Notch signaling pathway, which can be used to investigate the change in phenotypic behavior of the network (from bistable to oscillatory) through sensitivity analysis. However, representing cell–cell communication without embedding the model into a multicellular system may oversimplify important dynamics. Furthermore, fine-grained representation cannot be indefinitely scaled up by simply adding other ODEs for each cell of the epithelium because this would eventually lead to stiffness and numerical instability. The progressively granularity reduction can help to overcome these issues although it provides just a qualitative representation of the system behavior that focuses on the pattern generation: qualitative ODEs (Collier et al., 1996), agent-based systems (Reynolds et al., 2019), and logical models (Varela et al., 2018). These approaches allow to simulate pattern formation at the tissue level but, due to their nature, they either lack quantitative time or information about the species concentrations. To bridge these two levels, we suggest, analogously to hierarchical models (Uluseker et al., 2018), to connect multiple single quantitative modules through the logic formalism. We selected an ODE formalism to describe the intracellular kinetics of the different species which allows

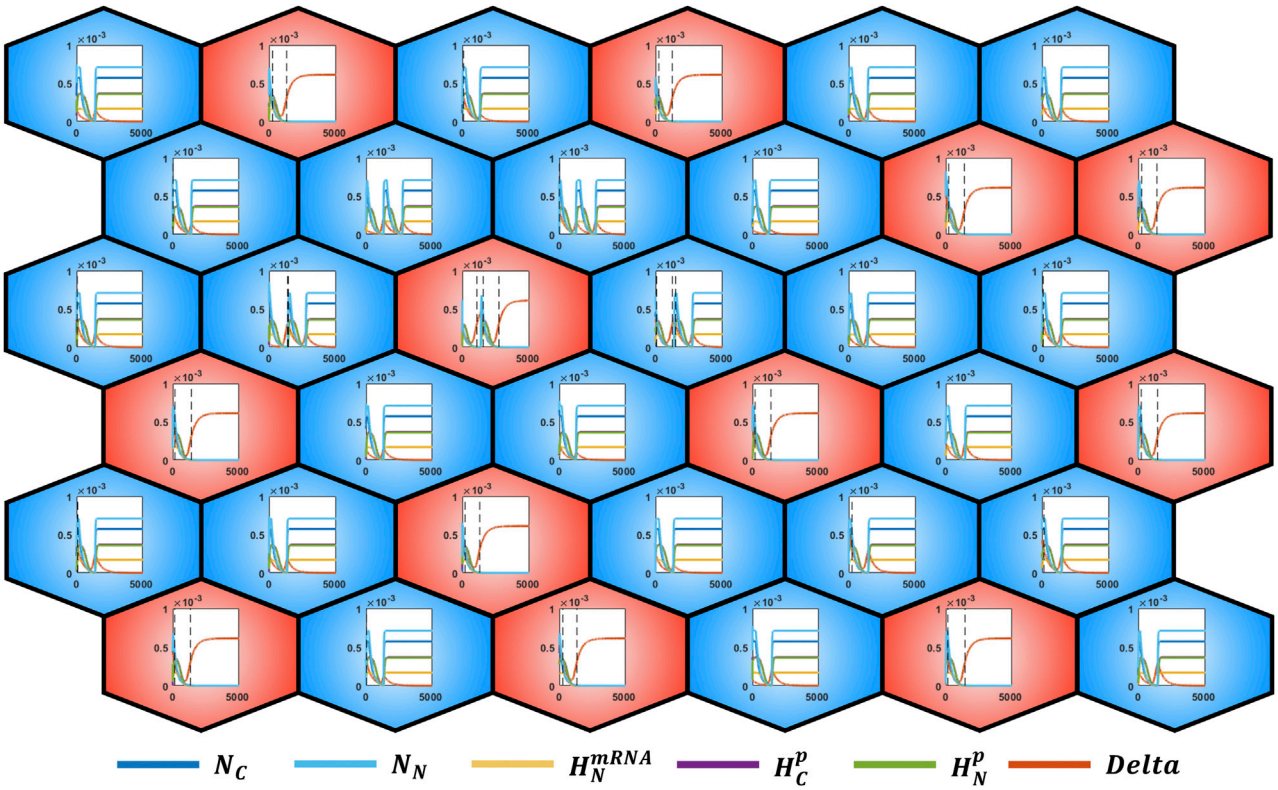


FIGURE 5 | alt-and-pepper pattern of a 6 × 6 grid with cylindrical boundary conditions obtained by applying the rule $D_E = D_E^{max}$ if $\sum_{i=1}^{6} C_i^1 \geq 2$. Each cell contains a plot of the computed intracellular variables dynamics; black dashed vertical lines indicate when the cell has generated an event by crossing the Delta threshold. The color of the cell indicates if the dominant variable is Delta or Notch (in agreement with the intracellular dynamics).

to quantitatively trace the system variables for each cell and provide a quantitative time. The rationale to use ODEs for the intracellular pathway hinges on the amount of data and the availability of experimental techniques apt to investigate missing gaps at this level, while cell–cell interaction and tissue dynamics are harder to explore and quantitatively characterize. As represented in **Figure 1**, each cell can receive two types of input: I_D or I_I; the former being used to encode cell–cell communication. The system is simulated following the pseudo-code of **Figure 2** until the break condition is met.

In our formalism, the tissue is composed of a grid of cells, each uniquely identified by their position and neighbors. We used a structure made by a 2D single cell layer of hexagonal cells, and the number of neighboring cells varies depending on the distance we consider (6 at distance 1, 12 at distance two etc.). The boundary condition of the grid, important for the evolution, was assumed to be cylindrical (folding the epithelium along the vertical axis) enabling the study of periodic patterns over a larger domain.

The hybrid strategy we propose tries to overcome the conundrum of providing a detailed enough description of the problem while keeping the model complexity under control, each formalism fulfilling a different purpose. The ODE system, being a modular quantitative representation of the intracellular cascade, can be expanded or substituted without major requirements (beside parameter calibration). Furthermore, multicellular logical models of Delta-Notch, as the case study presented by Varela et al. (2018), lack quantitative time and are simulated with a synchronous/asynchronous update of the grid. In our approach, the ODE system provides a quantitative time to the tissue system based on which the cellular grid is updated, allowing for a closer biological interpretation of the resulting dynamics. The internal species dynamics, stored along the simulation, can be instrumental to evaluate the biological processes at different scales.

The logic layer connecting the different single modules is used to describe qualitatively the receptor binding processes between adjacent cells. In addition, it is possible to encode biological information in the system using multilevel inputs; this, together with the logic rule complexity, can account for the receptor binding properties, although at a descriptive level. It is also possible to combine logical dependent inputs, I_D, with continuous independent input, I_I. This possibility, in the particular case of the Delta-Notch, can be used to model the intestinal crypt, where the soluble factor WNT controls part of the Notch intracellular cascade according to its concentration modulating the cell stemness (Demitrack and Samuelson, 2016). The hybrid strategy can thus be applied to the crypt system, considering Delta as a dependent input coming from the neighboring cells and WNT as continuous input with a concentration gradient (**Figure 7**).

The hybrid approach we propose in this work, despite the qualitative representation of some model components, can be

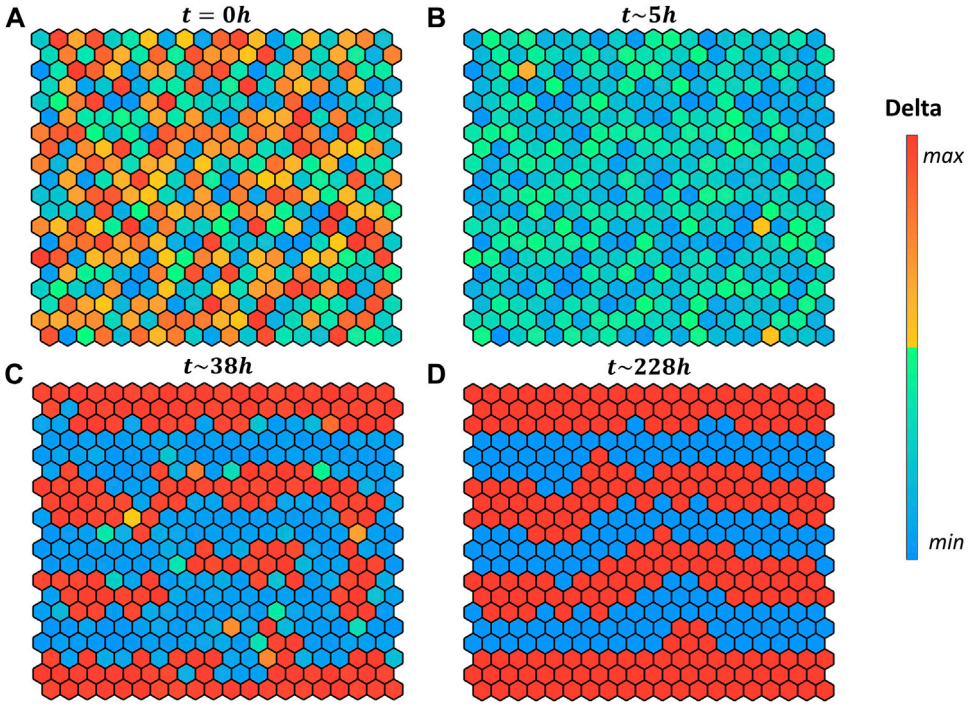


FIGURE 6 | Pattern evolution of a 20 × 20 grid with cylindrical boundary conditions (lateral borders are in contact) using the logical rule $D_E = D_E^{max}$ if $\sum_{i=1}^{18} C_i^3 \geq 10$, encoding that the input is present if at least ten neighbors out of the 18 at distance three are Delta positive. Colors indicate the intracellular Delta concentration according to the color bar. **(A)** Initial values of the variables are set randomly within the biological boundaries. **(B)** All cells in the grid become Delta negative. **(C)** Delta-positive cells start to emerge and affect their neighbor's fate. **(D)** Equilibrium grid is reached. The hours reported above each grid indicates the quantitative time obtained from the ODE systems.

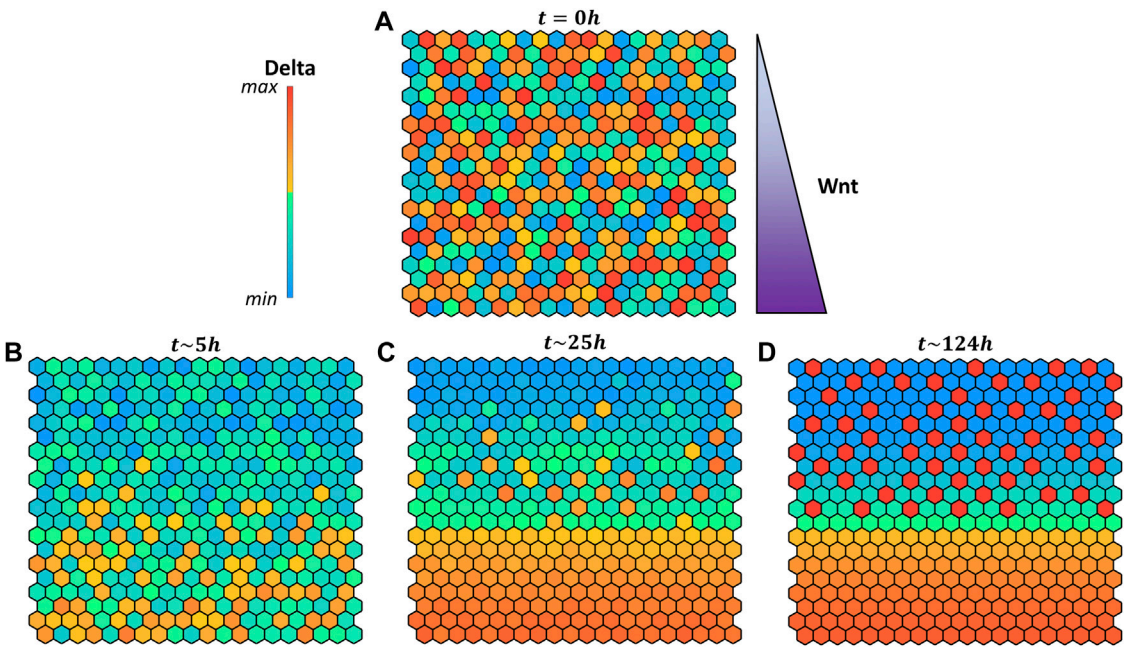


FIGURE 7 | Pattern evolution of a 20 × 20 grid with cylindrical boundary conditions (lateral borders are in contact) using the logical rule $D_E = D_E^{max}$ if $\sum_{i=1}^{6} C_i^6 \geq 1$, encoding that the input is present if at least one of the six neighbors at distance one are Delta positive. A positional gradient simulates Wnt concentration over the grid. Colors indicate the intracellular Delta concentration according to the color bar. **(A)** Initial values of the variables are set randomly within the biological boundaries. **(B)** Cells in the upper part of the grid become Delta negative, while on the bottom Wnt inhibition starts to manifest its effect. **(C)** A clear separation between differentiated (Delta + or Delta –) and undifferentiated cells (Wnt) emerges. **(D)** Equilibrium grid is reached. The hours reported above each grid indicates the quantitative time obtained from the ODE systems.

used to investigate areas in which there are still uncertainties on the underlying mechanism or a lack of the system characterization. This can be used to pave the road toward a more modular representation of biological problems, progressively expanding the current models by replacing the logic parts with more quantitative modules as soon as the necessary information are available.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

GS and SC developed the hybrid modeling approach and computed all the results. LM provided overall guidance of the project. All authors contributed to article preparation and approved the final version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.760077/full#supplementary-material

REFERENCES

Abou-Jaoudé, W., Traynard, P., Monteiro, P. T., Saez-Rodriguez, J., Helikar, T., Thieffry, D., et al. (2016). Logical Modeling and Dynamical Analysis of Cellular Networks. *Front. Genet.* 7, 94. doi:10.3389/fgene.2016.00094

Agrawal, S., Archer, C., and Schaffer, D. V. (2009). Computational Models of the Notch Network Elucidate Mechanisms of Context-dependent Signaling. *Plos Comput. Biol.* 5, e1000390. doi:10.1371/journal.pcbi.1000390

Aldridge, B. B., Burke, J. M., Lauffenburger, D. A., and Sorger, P. K. (2006). Physicochemical Modelling of Cell Signalling Pathways. *Nat. Cel Biol.* 8, 1195–1203. doi:10.1038/ncb1497

Artavanis-Tsakonas, S., Rand, M. D., and Lake, R. J. (1999). Notch Signaling: Cell Fate Control and Signal Integration in Development. *Science* 284, 770–776. doi:10.1126/SCIENCE.284.5415.770

Bertrand, N., Castro, D. S., and Guillemot, F. (2002). Proneural Genes and the Specification of Neural Cell Types. *Nat. Rev. Neurosci.* 3, 517–530. doi:10.1038/nrn874

Campos-Ortega, J. A. (1995). Genetic Mechanisms of Early Neurogenesis in Drosophila Melanogaster. *Mol. Neurobiol.* 10, 75–89. doi:10.1007/BF02740668

Collier, J. R., Monk, N. A. M., Maini, P. K., and Lewis, J. H. (1996). Pattern Formation by Lateral Inhibition with Feedback: a Mathematical Model of Delta-Notch Intercellular Signalling. *J. Theor. Biol.* 183, 429–446. doi:10.1006/jtbi.1996.0233

Collu, G. M., Hidalgo-Sastre, A., Acar, A., Bayston, L., Gildea, C., Leverentz, M. K., et al. (2012). Dishevelled Limits Notch Signalling through Inhibition of CSL. *Development* 139, 4405–4415. doi:10.1242/dev.081885

Danaher, P., Wang, P., and Witten, D. M. (2014). The Joint Graphical Lasso for Inverse Covariance Estimation across Multiple Classes. *J. R. Stat. Soc. B* 76, 373–397. doi:10.1111/rssb.12033

De Joussineau, C., Soulé, J., Martin, M., Anguille, C., Montcourrier, P., and Alexandre, D. (2003). Delta-promoted Filopodia Mediate Long-Range Lateral Inhibition in Drosophila. *Nature* 426, 555–559. doi:10.1038/nature02157

Del Bene, F., Wehman, A. M., Link, B. A., and Baier, H. (2008). Regulation of Neurogenesis by Interkinetic Nuclear Migration through an Apical-Basal Notch Gradient. *Cell* 134, 1055–1065. doi:10.1016/j.cell.2008.07.017

Demitrack, E. S., and Samuelson, L. C. (2016). Notch Regulation of Gastrointestinal Stem Cells. *J. Physiol.* 594, 4791–4803. doi:10.1113/JP271667

Gonzalez, A., Chaouiya, C., and Thieffry, D. (2008). Logical Modelling of the Role of the Hh Pathway in the Patterning of the Drosophila wing Disc. *Bioinformatics* 24, i234–i240. doi:10.1093/bioinformatics/btn266

Gross, T., Wongchenko, M. J., Yan, Y., and Blüthgen, N. (2019). Robust Network Inference Using Response Logic. *Bioinformatics* 35, i634–i642. doi:10.1093/bioinformatics/btz326

Hadjivasiliou, Z., Hunter, G. L., and Baum, B. (2016). A New Mechanism for Spatial Pattern Formation via Lateral and Protrusion-Mediated Lateral Signalling. *J. R. Soc. Interf.* 13, 20160484. doi:10.1098/rsif.2016.0484

Hartman, B. H., Reh, T. A., and Bermingham-McDonogh, O. (2010). Notch Signaling Specifies Prosensory Domains via Lateral Induction in the Developing Mammalian Inner Ear. *Proc. Natl. Acad. Sci.* 107, 15792–15797. doi:10.1073/pnas.1002827107

Hawe, J. S., Theis, F. J., and Heinig, M. (2019). Inferring Interaction Networks from Multi-Omics Data. *Front. Genet.* 10, 535. doi:10.3389/fgene.2019.00535

Ideker, T., and Lauffenburger, D. (2003). Building with a Scaffold: Emerging Strategies for High- to Low-Level Cellular Modeling. *Trends Biotechnol.* 21, 255–262. doi:10.1016/S0167-7799(03)00115-X

Kauffman, S. A. (1969). Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets. *J. Theor. Biol.* 22, 437–467. doi:10.1016/0022-5193(69)90015-0

Markowetz, F. (2017). All Biology Is Computational Biology. *PLOS Biol.* 15, e2002050. doi:10.1371/journal.pbio.2002050

Marnellos, G., and Mjolsness, E. (1998). A Gene Network Approach to Modeling Early Neurogenesis in Drosophila. *Pac. Symp. Biocomput.*, 30–41. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9697169 (Accessed January 27, 2021).

Mjolsness, E., Sharp, D. H., and Reinitz, J. (1991). A Connectionist Model of Development. *J. Theor. Biol.* 152, 429–453. doi:10.1016/S0022-5193(05)80391-1

Morris, M. K., Saez-Rodriguez, J., Sorger, P. K., and Lauffenburger, D. A. (2010). Logic-based Models for the Analysis of Cell Signaling Networks. *Biochemistry* 49, 3216–3224. doi:10.1021/bi902202q

Palsson, S., Hickling, T. P., Bradshaw-Pierce, E. L., Zager, M., Jooss, K., O'Brien, P. J., et al. (2013). The Development of a Fully-Integrated Immune Response Model (FIRM) Simulator of the Immune Response through Integration of Multiple Subset Models. *BMC Syst. Biol.* 7, 95. doi:10.1186/1752-0509-7-95

Renaud, O., and Simpson, P. (2001). Scabrous Modifies Epithelial Cell Adhesion and Extends the Range of Lateral Signalling during Development of the Spaced Bristle Pattern in Drosophila. *Develop. Biol.* 240, 361–376. doi:10.1006/dbio.2001.0482

Reynolds, E. R., Himmelwright, R., Sanginiti, C., and Pfaffmann, J. O. (2019). An Agent-Based Model of the Notch Signaling Pathway Elucidates Three Levels of Complexity in the Determination of Developmental Patterning. *BMC Syst. Biol.* 13 (13), 1–16. doi:10.1186/S12918-018-0672-9

Ryll, A., Bucher, J., Bonin, A., Bongard, S., Gonçalves, E., Saez-Rodriguez, J., et al. (2014). A Model Integration Approach Linking Signalling and Gene-Regulatory Logic with Kinetic Metabolic Models. *BioSystems* 124, 26–38. doi:10.1016/j.biosystems.2014.07.002

Simoni, G., Reali, F., Priami, C., and Marchetti, L. (2019). Stochastic Simulation Algorithms for Computational Systems Biology: Exact, Approximate, and Hybrid Methods. *Wires Syst. Biol. Med.* 11, e1459. doi:10.1002/wsbm.1459

Uluseker, C., Simoni, G., Marchetti, L., Dauriz, M., Matone, A., and Priami, C. (2018). A Closed-Loop Multi-Level Model of Glucose Homeostasis. *PLoS One* 13, e0190627. doi:10.1371/journal.pone.0190627

Varela, P. L., Ramos, C. V., Monteiro, P. T., and Chaouiya, C. (2018). EpiLog: A Software for the Logical Modelling of Epithelial Dynamics. *F1000Res* 7, 1145. doi:10.12688/f1000research.15613.1

Yu, J. S., and Bagheri, N. (2020). Agent-Based Models Predict Emergent Behavior of Heterogeneous Cell Populations in Dynamic Microenvironments. *Front. Bioeng. Biotechnol.* 8, 249. doi:10.3389/FBIOE.2020.00249

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2021 Selvaggio, Cristellon and Marchetti. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 01 February 2022
doi: 10.3389/fmolb.2021.793912

Modeling Progression of Single Cell Populations Through the Cell Cycle as a Sequence of Switches

Andrei Zinovyev[1,2,3]*, *Michail Sadovsky*[4,5,6,7], *Laurence Calzone*[1,2,3], *Aziz Fouché*[1,2,3], *Clarice S. Groeneveld*[8,9], *Alexander Chervov*[1,2,3], *Emmanuel Barillot*[1,2,3] and *Alexander N. Gorban*[7,10]

[1]Institut Curie, PSL Research University, Paris, France, [2]INSERM, Paris, France, [3]MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, Paris, France, [4]Institute of Computational Modeling (RAS), Krasnoyarsk, Russia, [5]Laboratory of Medical Cybernetics, V.F.Voino-Yasenetsky Krasnoyarsk State Medical University, Krasnoyarsk, Russia, [6]Federal Research and Clinic Center of FMBA of Russia, Krasnoyarsk, Russia, [7]Laboratory of Advanced Methods for High-Dimensional Data Analysis, Lobachevsky University, Nizhniy Novgorod, Russia, [8]Cartes d'Identité des Tumeurs (CIT) Program, Ligue Nationale Contre le Cancer, Paris, France, [9]Oncologie Moleculaire, UMR144, Institut Curie, Paris, France, [10]Department of Mathematics, University of Leicester, Leicester, United Kingdom

Cell cycle is a biological process underlying the existence and propagation of life in time and space. It has been an object for mathematical modeling for long, with several alternative mechanistic modeling principles suggested, describing in more or less details the known molecular mechanisms. Recently, cell cycle has been investigated at single cell level in snapshots of unsynchronized cell populations, exploiting the new methods for transcriptomic and proteomic molecular profiling. This raises a need for simplified semi-phenomenological cell cycle models, in order to formalize the processes underlying the cell cycle, at a higher abstracted level. Here we suggest a modeling framework, recapitulating the most important properties of the cell cycle as a limit trajectory of a dynamical process characterized by several internal states with switches between them. In the simplest form, this leads to a limit cycle trajectory, composed by linear segments in logarithmic coordinates describing some extensive (depending on system size) cell properties. We prove a theorem connecting the effective embedding dimensionality of the cell cycle trajectory with the number of its linear segments. We also develop a simplified kinetic model with piecewise-constant kinetic rates describing the dynamics of lumps of genes involved in S-phase and G2/M phases. We show how the developed cell cycle models can be applied to analyze the available single cell datasets and simulate certain properties of the observed cell cycle trajectories. Based on our model, we can predict with good accuracy the cell line doubling time from the length of cell cycle trajectory.

Keywords: cell cycle, mathematical modeling, molecular switches, transcription epoch, single cell data

OPEN ACCESS

Edited by:
Alberto Jesus Martin, Universidad Mayor, Chile

Reviewed by:
Anand Banerjee, Virginia Tech, United States
Paola Lecca, Free University of Bozen-Bolzano, Italy

***Correspondence:**
Andrei Zinovyev
andrei.zinovyev@curie.fr

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *12 October 2021*
Accepted: *15 December 2021*
Published: *01 February 2022*

Citation:
Zinovyev A, Sadovsky M, Calzone L, Fouché A, Groeneveld CS, Chervov A, Barillot E and Gorban AN (2022) Modeling Progression of Single Cell Populations Through the Cell Cycle as a Sequence of Switches. Front. Mol. Biosci. 8:793912. doi: 10.3389/fmolb.2021.793912

1 INTRODUCTION

Progression through the cell cycle represents a complex dynamical process, regulated at multiple levels including transcriptome and proteome. The major components of it have been characterized (Hunt, 1991; Hunt et al., 2011), and a complex molecular machinery has been revealed (Tyson, 1991). Nevertheless, many aspects of cell cycle functioning remain to be elucidated (Giotti et al., 2019).

Progression through the cell cycle can be seen as a trajectory in multidimensional space of all possible cellular states, similar to other processes such as differentiation or ageing. However, this trajectory is characterized by special properties because it represents a periodic process. From an oversimplified perspective, at the end of this trajectory, a cell splits into two daughter cells twice as small, where each daughter cell has a state identical to the initial state of its parent. This requirement imposes certain constraints on the geometry and underlying mechanisms of the cell cycle trajectory (CCT), which could be reproduced with a mathematical model.

Cell cycle process has been a subject of mathematical modeling for many decades (Chen et al., 2004a; Ingolia and Murray, 2004; Sible and Tyson, 2007). Most of the existing models focused on reproducing the regulatory logics at the level of protein expression, protein-protein interactions and post-translational modifications. Multiple modeling formalisms have been used such as chemical kinetics (Tyson, 1991; Chen et al., 2004b), logical modeling (Fauré et al., 2006; Deritei et al., 2019), Petri nets (Kotani, 2002), or approaches based on tropical algebra (Noel et al., 2012; Radulescu et al., 2012). A hybrid approach, combining discrete, governed by Boolean dynamics, and continuous, governed by chemical kinetics, variables was suggested to model cell cycle (Singhania et al., 2011; Noël et al., 2013). The mathematical description of the cell cycle transcriptional dynamics has not yet been thoroughly addressed, though.

High-throughput omics measurements gave rise to a number of molecular studies with the objective to characterize each cell cycle phase in terms of their associated molecular changes, i.e., sets of specifically expressed genes (Dominguez et al., 2016; Giotti et al., 2019). The appearance of single cell technologies reinforced the interest towards the description of the molecular organization of the cell cycle for several reasons. First, it allows the visualization of the cell cycle trajectory explicitly without synchronizing individual cells, which can be problematic, especially *in vivo*. Then, recent single cell transcriptomic and proteomic studies provide molecular description of progression through the cell cycle in a continuous fashion. Such representation attempts to delineate the cell cycle phase borders and also characterizes each cell for its precise progression position within each phase (Leng et al., 2015; Liu et al., 2017; Hsiao et al., 2020; Mahdessian et al., 2021).

A thorough understanding of cell cycle functioning is of utmost importance for cancer research, where the deviation from the normal cell cycle progression is expected. A number of questions can be raised: What is the normal pattern of the events comprising a cell cycle, and to what extent does it vary in normal physiology? What deviations from a normal cell cycle are characteristic for a tumor cell? What processes trigger these changes and are they specific to a cancer type? and many others.

Some mathematical models of the cell cycle try to tackle these questions. For example, agent-based or cellular automaton cell cycle models focus on the optimization of cancer drug delivery (Altinok et al., 2007), competition of fast and slow cell cycles within a tumor under treatment (Tzamali et al., 2020), or cell confluence and elongation of the G1 phase (Bernard et al., 2019). However, most of the existing models remain limited to describe the behavior of cell cycle during tumorigenesis at full complexity because of the existing discrepancy between the nature of the available molecular data and the level of the details of these models. Thus, the most comprehensive data source currently available is at the level of transcriptomic changes in single cells, while the existing modeling efforts focus on protein players. The data reveal the role of hundreds of genes and proteins in cell cycle dynamics, while the models include a tiny fraction of this number. Therefore, we believe that the development of mathematical models matching the scale and the nature of the abundant available data is still highly needed. In particular, even a simple mechanistic model of cell cycle transcriptome dynamics, capturing its main features, is lacking in the field. It appears that using dynamical variables representing relatively large lumps of genes (e.g., all genes involved in DNA replication) might be a useful coarse-grained approach to model cellular transcriptomes, which is one motivation of this study.

Single cell studies of cell cycle trajectories in snapshots of actively proliferating cells represent a unique opportunity to formulate the most general principles of cell cycle functioning. A recent study has introduced the principle of minimizing transcriptomic acceleration (Schwabe et al., 2020), which suggests that the transcriptomic cell cycle trajectory represents a spiral, or, after neglecting the relatively slow drift unrelated to cell cycle progression, a shape close to "a flat circle". This type of trajectories was indeed phenomenologically observed in the HeLa cell line profiled with scRNASeq technology, after deconvoluting the transcriptomic dynamics connected to the cell cycle from other sources of transcriptional heterogeneity. In particular, the absence of cell cycle-related transcriptional epochs was deduced from this model.

In the current study, we suggest an extended and different point of view on the properties of transcriptomic cell cycle trajectory which, in our opinion, in some cases better matches its observed properties in various cellular systems, when sufficiently good quality data can be collected. We propose a formal model of CCT as a sequence of epochs of growth during each of which the trajectory is approximately linear in the space of logarithmic coordinates. Therefore, CCT can be modeled as a piecewise linear trajectory in the space of logarithms of some extensive cell properties, followed by a shift at the vector with coordinates $-\log 2$ which represents the cell division event. This model explicitly assumes existence of well-defined transcriptional epochs in CCT.

Movement along linear trajectory in the space of logarithms of the values of some cellular properties means that along the trajectory any two such properties x_i, x_j are connected through a power law dependence $x_i = \alpha x_j^{\beta}$, $\alpha, \beta = \text{const}$. Such dependencies are known as allometric in many fields of biology (Holford and Anderson, 2017; Packard, 2017; White et al., 2019; Pretzsch, 2020; Zhou et al., 2021). Some approaches in mathematical chemistry and theoretical biology, dealing with systems in stable non-equilibrium, exploit the linear relations between chemical potentials which can be expressed as logarithms of species concentrations (Bauer, 1935; Gorban, 2018).

Particular cases of allometric dependencies are when all the quantities grow linearly with physical time, or when all the quantities follow exponential growth or decay $x_i = b_i \exp(a_i t)$. The model of movement along piecewise linear trajectories with an event of cell division represents the simplest scenario, easy to

simulate and analyse theoretically. Nevertheless, the most important conclusions derived from this analysis will stay valid for the trajectories that do not deviate too much from linearity.

Using the model of piecewise linear growth with division, we formulate a fundamental statement about correspondence between the number of linear segments in the cell cycle trajectory m, which corresponds to a number of the most important states of the cell cycle-related transcriptional machinery, and its effective embedding dimension n. The first part of the statement, $m \geq n$, can be described as a strict theorem with formal proof, whereas the second part, $m \leq n$, can be formulated as a feasible hypothesis, that can be validated using available data. The correspondence $m = n$ suggests that the embedding dimensionality of the transcriptomic cell cycle trajectory is larger than 2, since the number of segments we can observe can be as high as four or five. This allows us to state that the shape of the cell cycle trajectory is essentially not flat.

The type of models discussed here was partly introduced by *Shkolnik* (a pseudonym for a collective authorship), including authors of this manuscript (Shkolnik, 1989). Here, we significantly extend the previous effort and adapt it to the description of the cell cycle trajectory in single cell datasets.

In order to connect the geometric properties of cell cycle trajectory to interpretable mechanistic parameters, we extended the model of piecewise linear growth in logarithmic coordinates, to a simple kinetic model with rates depending on time as piecewise constant functions. In this case, some of the segments of the trajectory become nonlinear but remain smooth and do not deviate from linearity too far. Therefore, the suggested model is conceptually similar to previously suggested hybrid discrete-continuous models, but conceptualizes them, addresses the transcriptional dynamics and can be fit to multiple available scRNASeq datasets (Singhania et al., 2011; Noël et al., 2013).

The suggested cell cycle modeling framework and the representation of the cell cycle progression as a system of switches allows us to 1) determine which genes play the most important role in each transcriptional epoch, in a concrete system under study, 2) compare the genes related to the same transcriptional epoch between two biological systems or conditions, 3) predict the ratios between physical time durations of the transcriptional epochs, 4) predict the effect of shortening of certain transcriptional epochs on the shape of the cell cycle trajectory and transcriptional dynamics of the related groups of genes, and 5) predict the doubling time of proliferating cell populations from the length of the cell cycle trajectory observed in single cell scRNASeq data. The suggested framework can be exploited to study the cell cycle in various systems, from cell lines to tumors.

2 METHODS AND MATERIALS

2.1 Single Cell Data Used in This Study

We made a screening of available single cell sequencing of cancer cell lines in order to identify datasets with sufficient number of good quality single cell transcriptomic profiles and in which the principal source of transcriptomic heterogeneity would be progression through the cell cycle. We identified publicly available scRNASeq data on CHLA9 Ewing sarcoma cell line, produced with 10x Genomics sequencing technology (Miller et al., 2020), which contained more than 4,000 cells with total number of unique molecular identifiers (UMIs) varying from 10,000 to 50,000 per cell, after applying the standard quality criteria and filtering cells containing a large fraction (>20%) of reads in mitochondrial genes. For this dataset, we reanalyzed the raw sequencing data using Kallisto mapper (Bray et al., 2016) resulting in a loom file that could be used for obtaining the gene expression levels and for quantifying RNA velocity vectors (La Manno et al., 2018).

In addition, we used a recently published collection of 200 scRNASeq profiles of cancer cell lines from Cancer Cell Line Encyclopedia (CCLE) collection (Kinker et al., 2020). We also analyzed several scRNASeq datasets by downloading them directly from Gene Expression Omnibus (GEO).

The estimation of cell line doubling times, when available were obtained from Cellosaurus database (Bairoch, 2018).

2.2 Definition of Cell Cycle Genes

We systematically tested several existing definitions of cell cycle gene sets and verified that our results remain qualitatively invariant even if the choice of cell cycle gene set can vary. In our experiments, we used the following cell cycle gene set definitions:

- Standard "Regev's set": markers of S- and G2/M cell cycle phases used in scanpy tutorials (Tirosh et al., 2016)
- Set of cell cycle genes annotated in Reactome pathway database (Jassal et al., 2020)
- Set of top-contributing genes, extracted from application of independent component analysis (ICA) to the dataset under study, from those components whose top-contributing genes were strongly associated with the cell cycle. In particular, similar to our previous work (Aynaud et al., 2020), two independent components were significantly enriched with the markers of S- and G2/M cell cycle phases in all single cell cell line datasets we analyzed.

Cell cycle phase scores were computed as an average expression of marker genes for the corresponding cell cycle phase in log scale, which roughly corresponds to the geometric mean of the raw count measures.

2.3 Pooling Reads From Neighbouring Cells for Compensating the Technical Drop-Out

We found out that the cell cycle trajectories appear less noisy and more tractable by trajectory inference methods when standard pooling approach was applied to the raw count data, using an initial estimate of cell-to-cell proximity. More precisely, we used the initial standard data normalization and dimensionality reduction in order to compute the distances between cells and construct the initial kNN graph, which was used to pool row reads from a cell and all its k nearest neighbours. In our experiments, we used $k = 10$ and $n = 30$ components for reducing the data

dimensionality during normalization. Pooled read counts were used for final normalization, but the initial total read counts per cell measure were kept for visualization and further analysis.

2.4 Cell Cycle Trajectory-Based Single Cell Data Normalization

Total number of reads in a cell represents a strong signal in proliferating cell populations. By itself, it is an extensive value such that it should be divided (approximately) by half in the process of cell division. In our modeling approach, we needed a description of the cell state in terms of extensive values of gene expression levels measured such that they would be also divided approximately by two on average after the moment of cell division. Therefore, the widely used global library size normalization did not suit our purposes, since after global library size normalization, cell division does not lead to halving the total number of reads.

At the same time we observed that without any library size normalization, the cells presumably located at similar stages of cell cycle progression could be characterized by a wide range of total number of reads, probably caused by technical variability factors. Therefore, library size normalization was required but not at the global cell population level. We hypothesized that the total number of reads should increase in the course of cell cycle progression on average such that the cells characterized by similar value of pseudotime along the cell cycle trajectory could be normalized to the same local library size. As usual, this poses a chicken-or-egg problem because for reconstructing the cell cycle trajectory one needs normalized data, and for normalization of the library size one needs a reconstructed trajectory. This problem is similar to those approaches which use normalization locally conditioned on clusters in single cell datasets (Azizi et al., 2018).

We used a simplified two-stage approach for library size normalization which preserved both the geometric structure of CCT and the trend of increasing the total number of reads along CCT.

1. The row count data have been normalized to the global median number of counts and ln(x+1)-transformed, using standard functions of scanpy. 10,000 most variable genes have been selected, the dimensionality was reduced to 30 by PCA. In the reduced space, a kNN graph has been computed using the standard Euclidean distance for $k = 10$. This graph was used for pooling reads from neighbor cells as described above.
2. For such initially normalized dataset, we computed closed cell cycle trajectory in the subspace of cell cycle genes, by fitting a principal closed curve, using the Python implementation of ElPiGraph (Albergante et al., 2020). The data points were partitioned according to the proximity to the nodes of the elastic principal curve.
3. In each partition, we analyzed the distribution of the total number of reads across cells. We performed correction of cell-to-node assignment by splitting an anomalously wide partition between two neighboring partitions. The anomalously wide partition corresponded to the moment of cell division since it contained both cells at the very end of cell cycle progression with the largest number of reads and cells just after cell division event containing the minimal number of reads. Splitting this distribution allowed us to distinguish cells just before and just after the cell division into distinct partitions.
4. The median total number of counts in each resulting corrected partition was computed. The median values of the total number of reads in the cells of each partition have been smoothed by univariate spline or a piecewise-linear function of pseudotime, taking into account the cyclic boundaries of the trajectory.
5. Each cell's library size was normalized to the smoothed local median value of the total number of reads.
6. The newly normalized pseudocount data matrix passed through the same pre-processing as described in 1), namely a) Pooling reads from neighbour cells using the kNN graph obtained with trajectory-based normalized data, b) ln(x+1) transformation, selecting most variable 10 ,000 genes.

The cell cycle trajectory-based normalization procedure is illustrated in the Jupyter notebook at https://github.com/auranic/CellCycleTrajectory_SegmentModel, which can be easily reused for other cell lines.

2.5 Computing the Cell Cycle Trajectory and Quantifying Pseudotime

We used the ElPiGraph Python package to fit elastic principal curves or closed elastic principal curves (principal circles) to single cell data distributions (Albergante et al., 2020). ElPiGraph was applied in the data space defined by the set of 10 ,000 most variable genes or by the cell cycle-related genes, after dimensionality reduction by PCA (first 30 principal components were retained). In order to compute open elastic principal curve with q nodes, first a closed curve was fit with $q/2$ nodes, then a node with the least number of data points projected onto it was removed from the principal graph, and this configuration was used as an initialization to compute the elastic principal graph without branching and having q nodes.

The pseudotime s_i for a data point x_i was computed as a continuous geodesic distance measured from the root node to the projection of x_i onto the principal curve, quantified in the units of the number of edges. Therefore, the value of the pseudotime was in the range $[0, q-1]$, where q is the number of nodes. The root of the principal curve was chosen as one of its ends, such that the value of the initial total number of reads would increase as a function of pseudotime.

2.6 Curvature Analysis of the Cell Cycle Trajectory

In order to compute the Riemannian curvature of the principal curve defined by the position of its nodes in the multi-dimensional space $y_i \in R^n$, $i = 1, \ldots, q$, the node coordinates were first represented as n functions of the natural parameter (pseudotime) s, $y_i^k = y_i^k(s_i), i = 1, \ldots, q, k = 1, \ldots, n$. The value s_i for each node was taken as a number of edges of the elastic

principal curve connecting the node i to the root node. Each set of numbers $y_i^k(s_i), i = 1, \ldots, q$ was interpolated by a cubic univariate spline $y^k(s)$. In each node i of the curve the curvature was evaluated as $R_i = \sum_{k=1}^{n} (\frac{d^2 y^k(s)}{ds^2}|_{s=s_i})^2$.

2.7 Estimating the Effective Dimensionality of a Set of Vectors

In order to estimate the effective dimensionality of CCT, we used scikit-dimension Python package (Bac et al., 2021). We used linear estimators of global intrinsic dimensionality, based on application of PCA and various approaches to select the significant number of eigenvalues from the scree plot.

In order to compute the effective rank of a rectangular matrix, we looked at the distribution of its singular values, and selected such a number of them that the ratio between the largest and the smallest number would not exceed 10, such that the reduced matrix is well-conditioned.

3 RESULTS

3.1 Example of a Cell Cycle Trajectory Extracted From Single Cell Data

The current study is motivated by the observation that after appropriate pre-processing of single cell RNA-Seq data (see Methods), one can observe the cell cycle trajectory (**Figure 1**) which can be approximated by a piecewise linear curve, with a gap between the beginning and the end of the trajectory corresponding to the cell division moment.

Here we use the example of Ewing sarcoma cell line CHLA9 sequenced at single cell level using the Chromium 10x technology (Miller et al., 2020). The distinguishing feature of this dataset was that it contained a significant number of proliferating cells with single cell transcriptomes of good quality (more than 4,000 cells with the total number of Unique Molecular Identifiers (UMIs) between 10 ,000 and 50 ,000). Also, the proliferation signal in this dataset seems to explain the largest fraction of transcriptomic heterogeneity, since in the plane of the first two principal components one can clearly observe the cyclic trajectory. In other cell line single cell datasets, the proliferative signal can be masked by other sources of transcriptomic heterogeneity, requiring special procedures of data treatment to reveal it (Aynaud et al., 2020; Liang et al., 2020; Schwabe et al., 2020).

The scRNA-Seq data have been normalized in order to preserve the pattern of dynamics of the total number of counts (UMIs) along the CCT, see Methods section. The normalized gene expression levels are represented at the logarithmic scale, following the standard practice. The multi-dimensional distribution of single cell transcriptomic profiles projected into the space of the first 30 principal components has been approximated by a principal curve (see Methods). The curvature of the principal curve has been estimated using the standard formulas of differential geometry, which revealed the existence of curvature peaks, and reflecting the rapid turning points of the trajectory. We hypothesized that these turning points correspond to the large-scale changes in the transcriptional programs of the cell cycle process. The pattern of momentary velocities of the transcriptomic changes, estimated with RNA velocity, was compatible with this hypothesis (**Figure 1A**).

The pseudo-temporal dynamics of the known cell cycle-related genes confirmed that the trajectory curvature peaks delineate biologically meaningful transcriptional epochs. The epoch 0-A-B can be understood as an early G1 phase of the cell cycle, B-C as significantly overlapping with late G1-and S-phases, and C-D as overlapping with S- and G2-phases. The epoch D-E can presumably reflect the relatively short M phase (mitosis). Analysis of pseudotemporal gene expression dynamics inferred for this cell cycle trajectory shows that known cell cycle genes such as different cyclin types or E2F transcription factors have behaviour compatible with our interpretation (**Figure 1C**). We denote the identified transcriptional epochs as T1, T1s, T2s and Tm.

The switches between transcriptional epochs should not be confused with the action of cell cycle checkpoints that delineate cell cycle phases. The connection between the known molecular checkpoint mechanisms involving mainly protein-protein interactions and post-translational protein modifications and the transcriptional epochs might not be trivial or direct: partly, due to the delay between the gene and protein expression, and partly due to different parameters and constraints on the transcriptional and protein-protein interaction dynamics.

We can clearly observe the existence of the restriction point at the level of single cell transcriptome. In our notations, it belongs to the A-B segment of the cell cycle trajectory shown in **Figure 1A**,right. This transcriptional epoch separates post-mitotic (denoted as T1) and pre-replication parts of G1 phase, which corresponds to the classical definition of the R-point (e.g., from (Zetterberg et al., 1995)). Interestingly, in **Figure 1A**,right, one can observe that RNA velocity vectors reflect cells exiting from cell cycle and re-entering the cell cycle in the epoch between A and B turning points. Just after this transcriptional epoch, the expression of E2F transcription factors and Cyclin E start to increase as expected (**Figure 1C**).

We can also observe how, during each particular epoch, the components of a specific checkpoint mechanism are transcriptionally produced "just in time". For example, components of the G1 DNA damage checkpoint (e.g., CDC25A, CDKN1A) are produced during the T1s epoch of the cell cycle trajectory where the S phase starts, the components of G2 DNA damage checkpoints (e.g., CDC25B, CDC25C, CHEK2) are produced in the late part of the C-D epoch (T2s), and spindle checkpoint components (e.g., CDC20) are transcriptionally abundant during the mitosis-related epoch D-E (Tm) and after the cell division in T1 (**Figure 1C**). In this sense, the transcriptional dynamics prepare the correct ground for a proper succession of post-transcriptional events but the exact borders of the transcriptional epochs do not have to match the precise checkpoint timing.

Remarkably, within each of the identified transcriptional cell cycle epochs, the global dynamics of the transcriptome remain close to linear in the logarithmic scale. This allows us to suggest a simple model which can, for example, represent the collective

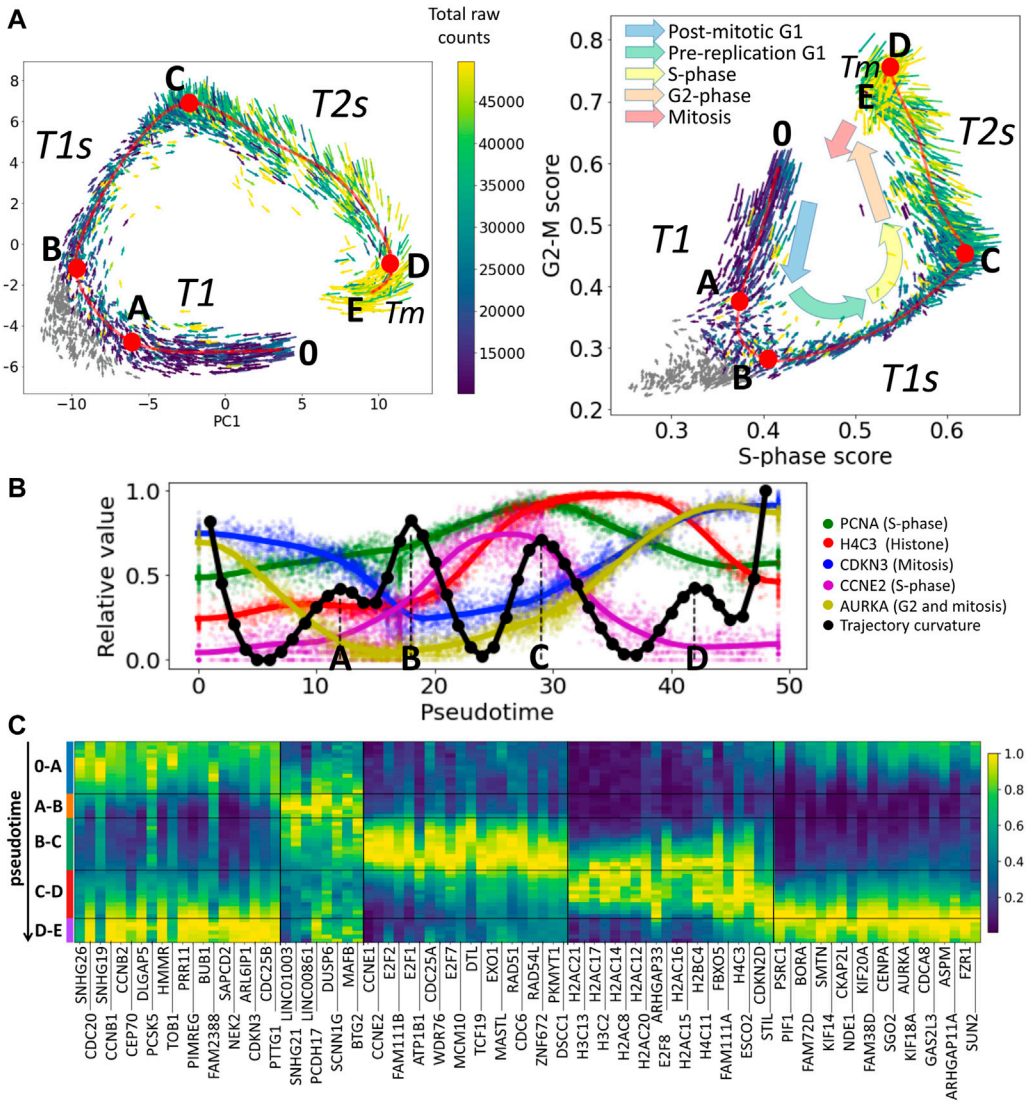


FIGURE 1 | Cell cycle trajectory (CCT) of CHLA9 Ewing sarcoma cell line in the single cell transcriptomic space. **(A)** Each cell is represented by an arrow reflecting the momentary direction and the speed of transcriptomic changes, estimated with RNA velocity. Two projections are shown, in the first two principal components and in the plane of S-phase and G2-M scores. The color of the arrows signifies either the total amount of RNA counts in the single cell profile (blue to yellow scale) or the cells in non-proliferative state (shown in grey). Red line shows an approximation of the cell cycle trajectory with a principal curve computed with ElPiGraph, directly in the 30-dimensional space of the first principal components of the dataset. Several particular positions along the trajectory (A,B,C,D) mark either the peaks of the Riemannian curvature of the principal curve (also shown in B) panel) or the beginning (0) and the end (E) of the trajectory. **(B)** Pseudotemporal transcriptomic dynamics of several cell cycle-related genes along CCT, shown relatively to the maximum value units. The pseudotime range is from 0 to 49, corresponding to the number of nodes in the approximation of the principal curve (50 nodes). In black, an estimation of the Riemannian curvature of the principal curve is shown, with peaks indicated by letters (A,B,C,D). **(C)** Pseudotemporal dynamics of genes whose expression is relatively high in one of the transcriptional epochs (trajectory segment) compared to other epochs. For each epoch the genes are ranked accordingly to the fold change of the mean expression of the gene in the epoch and outside the epoch. Only the genes having relatively large total variance across all cells are shown, and only top 20 genes maximum are shown per epoch for readability.

dynamics of the genes related to the S-phase and G2/M phases (see below).

3.2 Model of Cell Cycle as a Trajectory of Allometric Growth With Switches and Divisions

Based on the observations of the properties of the cell cycle trajectory in several scRNASeq datasets, we hypothesized that it can be recapitulated by a formal model of linear growth in logarithmic coordinates with switches and a cell division event. The suggested model is hybrid in nature, similar to some previously published models (Singhania et al., 2011; Noël et al., 2013). Namely, we distinguish the extrinsic observable cell state, characterized by continuous variables, and the intrinsic hidden cell state, characterized by discrete variables. The intrinsic state of a cell determines the parameters of the extrinsic dynamic process as in (Singhania et al., 2011).

Let the extrinsic state of a proliferating cell be determined by n substances quantified by their amounts, not their concentrations. Instead of their natural units (such as RNA counts), let us use the logarithms of these amounts. The cell is represented as an n-dimensional vector, and all possible combinations of these vector components define the cell configuration space. For our model, it is important that the considered n quantities are extensive measures, not intensive ones. Extensiveness here means that the total amount of a substance is a sum of the amounts found in different parts of a cell. A division (for two almost equal) daughter cells is formalized as a shift by the vector with all components equal −log 2 in this space. A relevant example of extensive quantity is the total amount of RNA molecules present in a cell, or the amount of any specific subset of RNA molecules, i.e., representing mRNAs of the genes involved in a particular process (such as mitosis or S-phase).

We assume that there exists a finite discrete set of intrinsic cell states. In each of these states, the cell follows a linear trajectory in the extrinsic and continuous cell state space. This trajectory extends until the cell meets a condition, where a switch into another intrinsic state of the cell happens, which changes the direction of the trajectory. For simplicity, we assume that the conditions of a switch can be described by a linear function. The cell movement continues until a particular condition is met in which the cell division event is triggered leading to the aforementioned translation of the vector representing the extrinsic cell state.

Let us introduce some mathematical notations and consider a deterministic automaton A whose complete state is represented by a pair (x, s), where $x \in R^n$ is a vector in n-dimensional continuous space (extrinsic state), and $s \in S$ is an integer number from a finite set $S = \{S_1, .., S_m\}$ (intrinsic state). In the rest of the study, we will call x a position of A and s an intrinsic state of A. We will denote the automaton A in position x and in the intrinsic state s as $A(x|s)$.

Each intrinsic state S_k is parameterized by a vector $a_k \in R^n$, $k = 1..m$ and by a linear manifold D_k of dimensionality $n - 1$ embedded in R^n (hyperplane), which we will call "the cell division hyperplane". D_k can be undefined, in this case, we denote $D_k = null$.

Let us also introduce a set of p functions $G = \{g_1, \ldots, g_p\}$, g_i: $S \to S$, which we will call switches. Each switch g_i is a map which converts an intrinsic state $s_j \in S$ into another intrinsic state $s_r \in S$. Each switch g_i is parametrized by a hyperplane L_i existing in R^n and inducing the switch function g_i each time the trajectory of the automaton intersects L_i (see **Figure 2A**).

Finally, we introduce the cell division event ϕ which is a map between two states of A, such that $\phi((x, s)) \to (x + d, s_d)$, where $d \in R^{n-}$ is a vector with negative components, and $s_d \in S$ is one of the possible intrinsic states of A.

We will characterize any hyperplane here by a linear functional $f(x|b, c) = b + < c, x >$, $b \in R$, $c \in R^n$, where $< , >$ denotes the standard scalar product between two vectors. Using such a functional, for any pair of vectors x_i, $x_j \in R^n$ we can determine if the linear segment connecting x_i and x_j intersects the hyperplane or not. If the segment intersects the hyperplane then $f(x_i)f(x_j) < 0$, and if it does not intersect then $f(x_i)f(x_j) > 0$. $f(x_i)f(x_j) = 0$ is satisfied only in a non-general position when either x_i or x_j is located exactly on the hyperplane.

The update rules for the automaton A are described as follows. The automaton is in some initial position x_0 and the intrinsic state s_0. It starts to move along the linear trajectory described by the equation $x = x_0 + a_0 t$, where a_0 is the vector of movement associated with the state s_0. This movement continues unless one of the two events happens. In the first case, A reaches the corresponding cell division plane D_0 (in case D_0 is not null). Then, the cell division event is triggered, $A(x|s) \to A(x + d|s_d)$. In the second case, x reaches a switch hyperplane L_j and then a switch of the intrinsic state of A happens without changing its position, $A(x|s_0) \to A(x|g_j(s_0))$. The movement continues along a new trajectory, corresponding to the new cell state, following the same rules: either the trajectory hits the cell division hyperplane or any of the switch planes.

To summarize, the automaton A is characterized by its position and the intrinsic state, see **Figure 2A**. The asymptotic (in the infinite time limit) temporal dynamics of A is parameterized by a set of cell division planes $D = D_i$, $i = 1,...,k$, a set of switch functions $G = \{g_i\}$, $i = 1, \ldots, p$, the corresponding switch hyperplanes $L = \{L_i\}$, $i = 1, \ldots, p$, and the parameters of the cell division event (namely, the translation vector d and the state after cell division s_d).

It is convenient to encode the state s as a binary sequence of length r representing the on-off states of r triggers. In this case, a switch can be thought of as changing only one particular trigger from on to off or vice versa. In many situations, this makes the description of switch functions g: $S \to S$ quite natural as explained below. Also, the state of the trigger might not be strictly binary but characterized by several discrete positions, for example {0, 1, 2}, just as it is the case in modeling multi-level discrete dynamics, where each discrete variable can take a value from a pre-defined finite set of levels.

The exact asymptotic trajectory of the automaton A can, in principle, depend on the initial position x_0 and the initial intrinsic state s_0 of A.

3.3 Simple Example of Dynamics With Switches and Cell Division Events

In the above-described switch-like dynamics, one can find examples of relatively complex behaviors even for simple model settings (**Figure 2**,B-E). As an illustration, we modeled a simple dividing automaton characterized by a position vector x with only two coordinates x_1, x_2. The automaton intrinsic state s encoded by only one binary trigger, so the automaton can be in two states $s = 0$ and $s = 1$, characterized by two vectors of movement a_0 and a_1, respectively. In order to be able to modify the trigger in both directions, we have to introduce two switch hyperplanes $L^{(+)}$ and $L^{(-)}$ with corresponding switch functions $g^{(+)} = 1$ (switch trigger on) and $g^{(-)} = 0$ (switch trigger off). Note that in this case the switch functions are constant, i.e., they map any state (which can be either 0 or 1) to a particular state. Let us also assume that the division event changes the automaton position but does not change its intrinsic state.

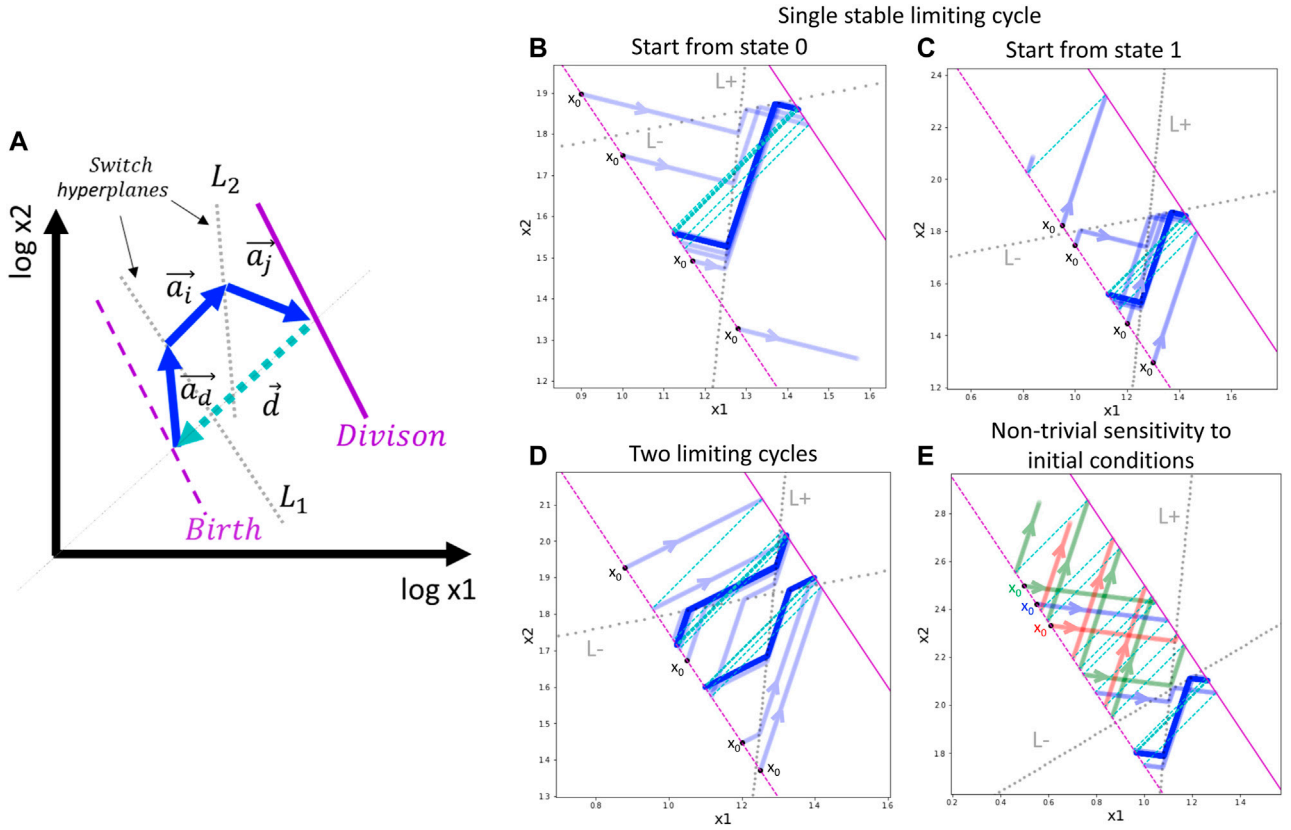


FIGURE 2 | General schema of switch-like dynamics and application to a toy model with a single trigger. **(A)** Schematic two-dimensional example of a limiting trajectory with division. The division hyperplane *D* is shown in purple, solid line. The birth hyperplane *B* is obtained from *D* by translation at vector *d*, shown in cyan (the most natural is to assume all the components of *d* to be –log 2). Two switch hyperplanes L_1 and L_2 are shown by dotted grey lines. The limiting cycling trajectory is represented by blue arrows. **(B,C)** Example of single limiting cycle in the switching dynamics. Depending on the initial state of the automaton and the initial position, the trajectory enters into the limit cycle or degenerates (goes to infinity). For the same parameters, four initial conditions are shown. The trajectory is plotted with semi-transparent blue color such that the intense blue line designates the trajectory cycling multiple times on top of itself. **(D)** Example of existence of two limit cycles. Depending on the initial state and position, the automaton ends up in one of the two possible limit cycles. **(E)** Example of non-trivial dependence of the switching dynamics on the initial position of the automaton. The trajectories drawn by different colors from three closely located initial positions are shown, with two leading to degenerated dynamics and one located in between the first two, leading to the limit cycle. In (B–E) panels, the initial position of the automaton is always shown at the birth hyperplane *B* (shown by dashed purple line), therefore, it is characterized by a single number.

In this simple toy example, by slightly varying parameters of the switching hyperplanes and the movement vectors, one can observe several interesting scenarios. Firstly, we observe that the automaton can approach and stay on a limit cycle trajectory, or it can diverge, meaning that one of the coordinates of the vector x goes to infinity or zero (**Figure 2**,B-C). Convergence or divergence to a limit cycle depends on the initial intrinsic state and the initial position of the automaton on the birth hyperplane.

In a more complex scenario, the switching dynamics trajectory can be characterized by two limit cycles that can be achieved from different initial intrinsic states and positions (**Figure 2D**).

By varying the positions of the switching hyperplanes in this toy example, one can observe the effect of non-trivial sensitivity to the initial conditions (**Figure 2E**). In this case, the birth hyperplane can be split into a sequence of alternating intervals of equal length such that starting from one interval, the dynamics finally converges to the limit cycle, and starting from another interval, the dynamics diverges to infinity.

3.4 Two-Dimensional Model of Cell Cycle Progression, Fitted to the Single Cell Transcriptomic Data

Let us denote the aggregate signal related to the activation of genes associated with the S-phase of the cell cycle program as *S*, and the signal related to the activity of genes in G2 and M phases as *M*. Therefore, we will characterize the position of the automaton by a vector (x_S, x_M), just as it is presented in **Figure 1A**, right panel. Let us denote the position of the turning points in the trajectory as $(x_S^{(i)}, x_M^{(i)})$, where $i \in \{0, A, B, C, D, E\}$.

We will encode the state of the system by the levels of two triggers, one associated with the *S* signal and another associated with the *M* signal. The three levels are denoted as a set

{2 = *synthesis*, 1 = *decay*, 0 = *degradation*}. Intuitively, these levels correspond to the state of active transcription of the corresponding set of transcripts ("synthesis"), absence of active transcription in which the transcripts are passively degraded according to some base rate ("decay"), and the process of active degradation when the transcripts are degraded more rapidly than the base rate ("degradation"). The state of the system is thus encoded by a pair of 3-level variables $i, j \in \{0, 1, 2\}$. The 2D vectors of linear movement a_{ij} are encoded by six rates $k_j^v, i \in \{0, 1, 2\}, v \in \{s, m\}$, such that $a_{ij} = (k_i^s, k_j^m)$. Following the intuition behind the introduced trigger levels, we assume constraints $k_2^v > 0, k_1^v < 0, k_0^v < 0, k_0^v < k_1^v < k_2^v$.

Let us introduce three switches. The first switch g_1 turns on the synthesis of both variables, i.e., g_1 (•, •) → (2, 2), where • designates any level of the trigger. The second switch turns off the synthesis of genes in S-phase: g_2 (2, •) → (1, •). The third switch turns off all the transcription, g_2 (•, •) → (1, 1). We assume that the division is possible only in the state (1, 1) with transcription switched off, and that after the division event, the cell enters into the state of active degradation of the cell cycle genes (0, 0).

The three introduced switches will be characterized by the corresponding switching hyperplanes. The first switch is triggered when the sum of the collective aggregated levels of expression of the genes involved in S and G2/M phases reaches some minimum $c_{\min}$, therefore, the linear functional associated with the first switch hyperplane is $f_1(x_s, x_m) = x_m + x_s - c_{\min}$. The second switch is triggered whenever the collective aggregated level of expression of S phase-associated genes reaches some maximum value $S_{\max}$, therefore, the linear functional associated with the second switch hyperplane is $f_2(x_S, x_M) = x_S - S_{\max}$. Finally, the third switch is triggered when the collective aggregated level of expression of G2/M phase-associated genes reaches some maximum value $M_{\max}$, therefore, the linear functional associated with the third switch hyperplane is $f_3(x_S, x_M) = x_M - M_{\max}$.

In the end, the cell division event is triggered when the collective aggregated level of expression of G2/M phase-associated genes crosses some threshold M_e, therefore, the linear functional associated with the division event is $f_d(x_S, x_M) = M_e - x_M$.

Let us define the number of parameters in this simple switching model. Three introduced switches are characterized by 4 parameters $c_{\min}$, $S_{\max}$, $M_{\max}$, M_e. There exist six rates k_i^v characterizing the movement vectors in the $9 = 3^2$ possible states, corresponding to all possible combinations of trigger levels. However, qualitatively, the dynamics in each automaton state is determined only by the direction of the corresponding vector and not its amplitude: therefore, one parameter per state visited is needed during the progression through the cell cycle. Under certain constraints on the rates formulated above, and also on the switch parameters (namely, $c_{\min} < S_{\max}, M_{\max}, M_e < M_{\max}$), the suggested model is constructed such that along the cell cycle trajectory only four states will be visited in a predefined order (0, 0) → (2, 2) → (1, 2) → (1, 1). Therefore, the total number of parameters equals 8.

Knowing the position of four characteristic points along the cell cycle trajectory, namely $(x_S^{(B)}, x_M^{(B)})$, $(x_S^{(C)}, x_M^{(C)})$, $(x_S^{(D)}, x_M^{(D)})$, $(x_S^{(E)}, x_M^{(E)})$, it is possible to completely parameterize the automaton. The starting and the ending point of the cell cycle trajectory must be connected by the relation $(x_S^{(0)}, x_M^{(0)}) = (x_S^{(E)}, x_M^{(E)}) + d$, where d is the vector with components $(-\log 2_{10}2, -\log 2_{10}2)$.

Therefore, we put $S_{max} = x_S^{(C)}, M_{max} = x_M^{(D)}, M_e = x_M^{(E)}$. Instead of using directly the B point, we will use the position of the non-proliferating cell with the maximum sum of the coordinates in the S, M plane, and we designate it as $x_S^{B'}, x_M^{B'}$ (other choices are also possible). Then $c_{min} = x_S^{(B')} + x_M^{(B')}$. Then we define rates:

$$k_2^v = \frac{x_v^{(C)} - x_v^{(B')}}{\|x^C - x^{B'}\|}, k_1^S = \frac{x_S^{(C)} - x_S^{(D)}}{\|x^C - x^D\|}, k_1^M = \frac{x_M^{(E)} - x_M^{(D)}}{\|x^{(E)} - x^{(D)}\|}, k_0^v = \frac{x_v^{(B')} - x_v^{(0)}}{\|x^{(0)} - x^{B'}\|}$$

The resulting steady state cell cycle trajectory is shown in **Figure 3**.

We denote the linear segments of the trajectory shown in **Figure 3** as T1, Ts, T2, Tm, assuming that they have significant overlap with G1, S, G2 and M phases correspondingly.

The suggested model describes 2D dynamics of the signals S, M which are empirically shown to explain most of the variance of all cell cycle genes in scRNASeq data (see below). However, higher-dimensional generalization of the suggested model is always possible. Also, in the model, we simplified the observed dynamics in **Figure 1A**, left which seems to contain five segments, with an additional curvature peak in point A. The segment A-B seems to contain non-proliferating cells, and might correspond to the transcriptional epoch most similar to the quiescent cell state, when the active degradation of the mitotic transcripts is completely finalized. The existence of this epoch is less pronounced in the S, M projection (**Figure 1A**,right), therefore we merged segments 0-A and A-B' as the first order approximation.

3.5 Connection Between the Effective Embedding Dimensionality of Cell Cycle Trajectory and the Number of Intrinsic States

The introduced cell cycle modeling framework is a simple and empirical model, lacking mechanistic details. Its main advantage is the possibility of analytical treatment of the most general geometrical cell cycle trajectory properties. In this section, we use this framework to prove a theorem connecting the number of the intrinsic states of the cell cycle trajectory and its intrinsic dimensionality.

This geometry is embedded into a space of omics measurements, whose dimensionality might be very high (e.g., expression of thousands of genes). However, we can assume that the intrinsic dimensionality (ID) of CCT is much smaller and that the extrinsic state of the cell progressing through the cell cycle can be characterized by n extensive variables, where n is relatively

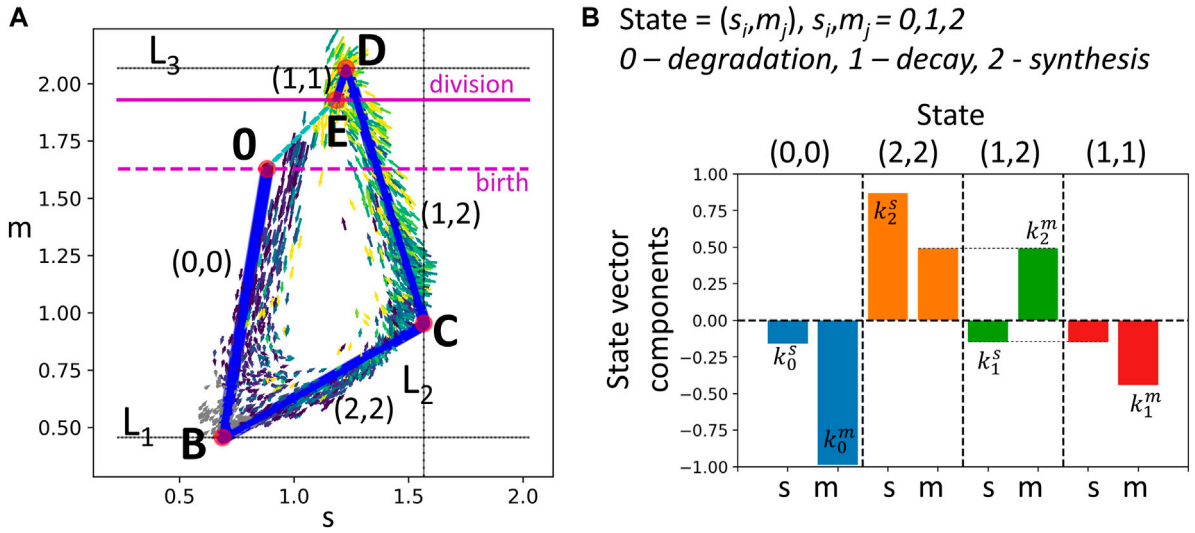


FIGURE 3 | Modeling transcriptomic cell cycle trajectory by an allometric growth with switches. **(A)** Piecewise linear cell cycle trajectory fit to the single cell RNASeq data (cell cycle trajectory, shown in **Figure 1A**,right). The model contains three switching planes L_1, L_2, L_3, and is characterized by four states. The states are encoded with two triggers, each possessing three possible levels 0,1,2, the biological meaning of which is specified in B). **(B)** The growth vectors associated with each state are encoded by rates k_i^S,k_j^M, such that the components of the growth vectors equal (k_i^S,k_j^S), where i and j are the levels of the corresponding triggers.

small. We will refer to n as CCT embedding dimensionality. Empirically, it can be estimated by studying the snapshot of dividing single cells profiled with a particular technology, and computing its global intrinsic dimensionality (ID), provided that other non cell cycle-related sources of heterogeneity could be dismissed in measurements. Estimating ID can be done using one of the many existing methods for ID estimation (Albergante et al., 2019; Bac and Zinovyev, 2020; Bac et al., 2021).

Let us establish the expected relation between n and the number of intrinsic states m of the automaton approximating CCT. We intend to claim that theoretically n should match m under some natural assumptions.

We first state that m cannot be smaller than n. In the theory of allometric growth with switches this statement has a character of strict theorem (see below), $m \geq n$. Secondly, we state that n is expected to be at least equal to m. Both statements are based on argumentation using "general position" statements. However, the former one is strictly necessary, while the latter one represents a feasible hypothesis.

Theorem on the number of intrinsic cell cycle states. *The number of segments m in the cell cycle trajectory modeled by the automaton with switches and linear growth in logarithmic coordinates is not less than the cell cycle trajectory intrinsic dimensionality n, or* $m \geq n$.

Proof. Let us consider the CCT dynamics in its n coordinates each of which represents an extensive variable. The variable extensiveness means, in particular, that its value, after the cell division moment, is divided by two. In logarithmic scale the cell division corresponds to the shift by vector $d \in R^n$ with n coordinates each of which equals −log 2. Each intrinsic state is associated with a growth vector $a_i \in R^n$, $i = 1..m$. All non-negative linear combinations of a_i form a convex cone $Q = \{\sum_{i=1}^{m} \lambda_i a_i\}, \lambda_i \geq 0$. If $m < n$ then the set of vectors $\{d, \{a_i, i = 1..m\}\}$ is almost always linear independent and $-d \notin Q$. Hence, −d is linearly separable from Q, according to the standard separability theorems. Linear separability of a point from a convex cone can be expressed as that for any non-zero $x \in Q$ we can find a linear function l () such that $l(d) = 0$ and $l(x) > 0$. This makes the periodic cell cycle model impossible, because the function $l(x)$ increases along any growth direction, since for any i and $\lambda > 0$ we have $l(x + \lambda a_i) = l(x) + \lambda l(a_i) > l(x)$, and after cell division $l()$ does not change since $l(x + d) = l(x) + l(d) = l(x)$. Therefore, the necessary condition of existence of stable cell cycle trajectory is $m \geq n$, when the set of vectors $\{d, \{a_i, i = 1..m\}\}$ is linearly dependent, and also such choice of a_i that $-d \in Q$. Only in this case one can satisfy the cyclic condition $\sum_i^m \lambda_i a_i + d = 0$ in general position of vectors $\{d, \{a_i, i = 1..m\}\}$.□

In simple words, this means that if $m < n$ then in a general position, each cell division (shift by d) moves a cell state out of the subspace defined by the growth vectors. The only way to make the trajectory stay in this subspace is to make the cell division vector d belong to this subspace that can be guaranteed only if $m \geq n$ (see **Figure 4**). The condition $m \geq n$ is necessary but not sufficient for a model to converge to a limit cycle. For example, in **Figure 7**, $m = n = 2$ (the theorem condition is satisfied) but the limit cycle in the model can be achieved only from some initial conditions and for some choice of vectors a_0, a_1.

Note that the proven Theorem is more general than the model of allometric growth with switches itself since it does not assume any particular shape of the switching surfaces L_k: they can be linear or nonlinear. Another generality consists in that the vector d can have any non-zero coordinates, not necessarily equal to −log 2.

Examples in **Figures 2**, **3** shows the case $n = 2$, $m > n$. The cell cycle trajectory modeled in **Figure 3** contains $m = 4$ segments in 2D, which makes the vectors $a_i \in R^2$, $i = 1, \ldots 4$ linearly

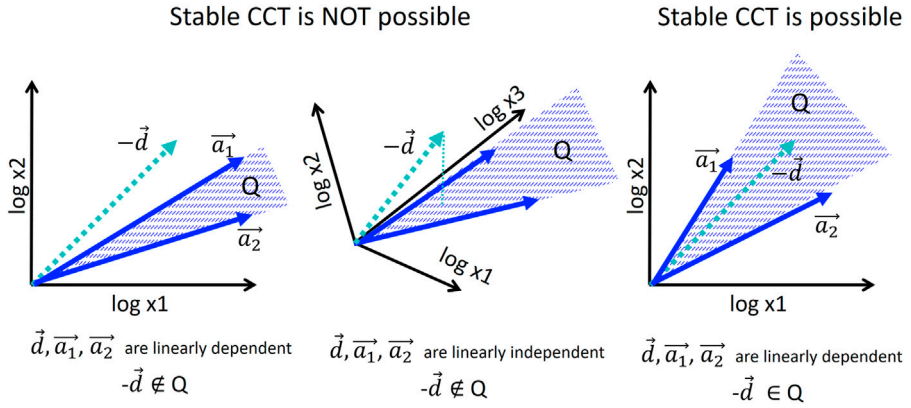


FIGURE 4 | Condition of existence of stable cell cycle trajectory in the model of allometric growth with switches. For illustration, only two growth vectors a_1, a_2 are considered, and 2D or 3D embedding space. Stable piecewise linear trajectory is possible only if the negative of the cell division vector $-d$ belongs to the convex cone $Q = \sum_i^m \lambda_i a_i, \lambda_i \geq 0$. Only in this case, the cyclic equality $\sum_i^m \lambda_i a_i + d = 0$ is possible. In general position, the condition can be met only when $m \geq n$, where n is the dimensionality of the trajectory space (see text for the formal proof).

dependent, and, of course, $d \in R^2$. The cell cycle model based on allometric growth is not contradictory in this case.

Now let us formulate our second statement. We can recall that vectors a_i are confined to the n-dimensional intrinsic subspace of CCT by projection from the multi-dimensional ambient space of all elementary measurements. The choice of n depends on our estimate of the CCT intrinsic dimensionality. However, movement along vectors a_i can be also seen in the complete space with thousands of coordinates. In this space, for sufficiently small m, any m vectors will almost always be linearly independent. Only projection into smaller than m-dimensional space will guarantee that these vectors are linearly dependent. This makes us hypothesize: if m segments are observed in CCT piecewise linear approximation in any linear projection then the most natural choice for n is at least m, i.e., $n \geq m$. Combining the two statements ($m \geq n$ and $n \geq m$) allows us to state that the correspondence $m = n$ is the most natural expectation for a cell cycle trajectory.

We explicitly verified this correspondence for the trajectory shown in **Figure 1**. The curvature analysis suggests the existence of five segments for the cell cycle trajectory reconstructed in the subspace of 30 first principal components of the complete dataset. However, some of these components might correspond to the variance not related to the progression through the cell cycle. In order to diminish the possible role of this variance, we considered a reduced version of the dataset confined to cell cycle-related genes only. We estimated the global intrinsic dimensionality, using six different linear ID estimators from scikit-dimension Python package (Bac et al., 2021), and it varied from 2 to 7, with average value 4.0. The scree plot shows existence of two dominant eigenvalues explaining 83% of total variance, indicating that the trajectory is relatively flat and located close to a 2D linear manifold. However, the residual variance demonstrated visible patterns related to transcriptional epochs in at least the first four principal components (**Figure 5**). The distribution of projections on the first four principal components well separated some transcriptional epochs (**Figure 5**,diagonal). Also, projections in higher dimensions highlighted the existence of sharp turning points between the segments which were less clear in the 2D projection on the first two principal components.

In addition, we split the data points into five classes according to projection on five segments of the principal curve (0-A, A-B, B-C, C-D, D-E), each of which is approximately linear. For each of this class, we computed the unity vector corresponding to the direction of the first principal component in the space of cell cycle genes with 198 dimensions. Afterwards, we estimated the effective rank of the matrix composed of five vectors representing the directions of the transcriptional epochs in the multi-dimensional space (see Methods), and it appeared to be 4, which indicates to that at least four out of five vectors determining the trajectory segments can be considered linearly independent.

As a result, we concluded that the embedding dimensionality for the transcriptomic cell cycle trajectory can be estimated as close to four. Therefore, restricting the trajectory to the plane of aggregate collective expressions of genes associated with S phase and G2/M phase (which roughly corresponds to the first two principal components) is a useful but incomplete approximation of CCT dynamics. Our reasoning suggests searching for additional biologically meaningful and statistically independent scores describing the progression through the cell cycle. The concrete gene expression dynamics shown in **Figure 1B** provides a hint in this direction, but a careful and complete investigation of this question should be a subject of a separate study. As an additional argument, we can mention that some mathematical cell cycle models based on a fit to real data are four-dimensional (Singhania et al., 2011).

3.6 Extending the Modeling Formalism to Piecewise Smooth Trajectories: Simple Kinetic Model of Cell Cycle at Transcriptomic Level

The piecewise-linear model of automaton with switches described in the previous sections is phenomenological

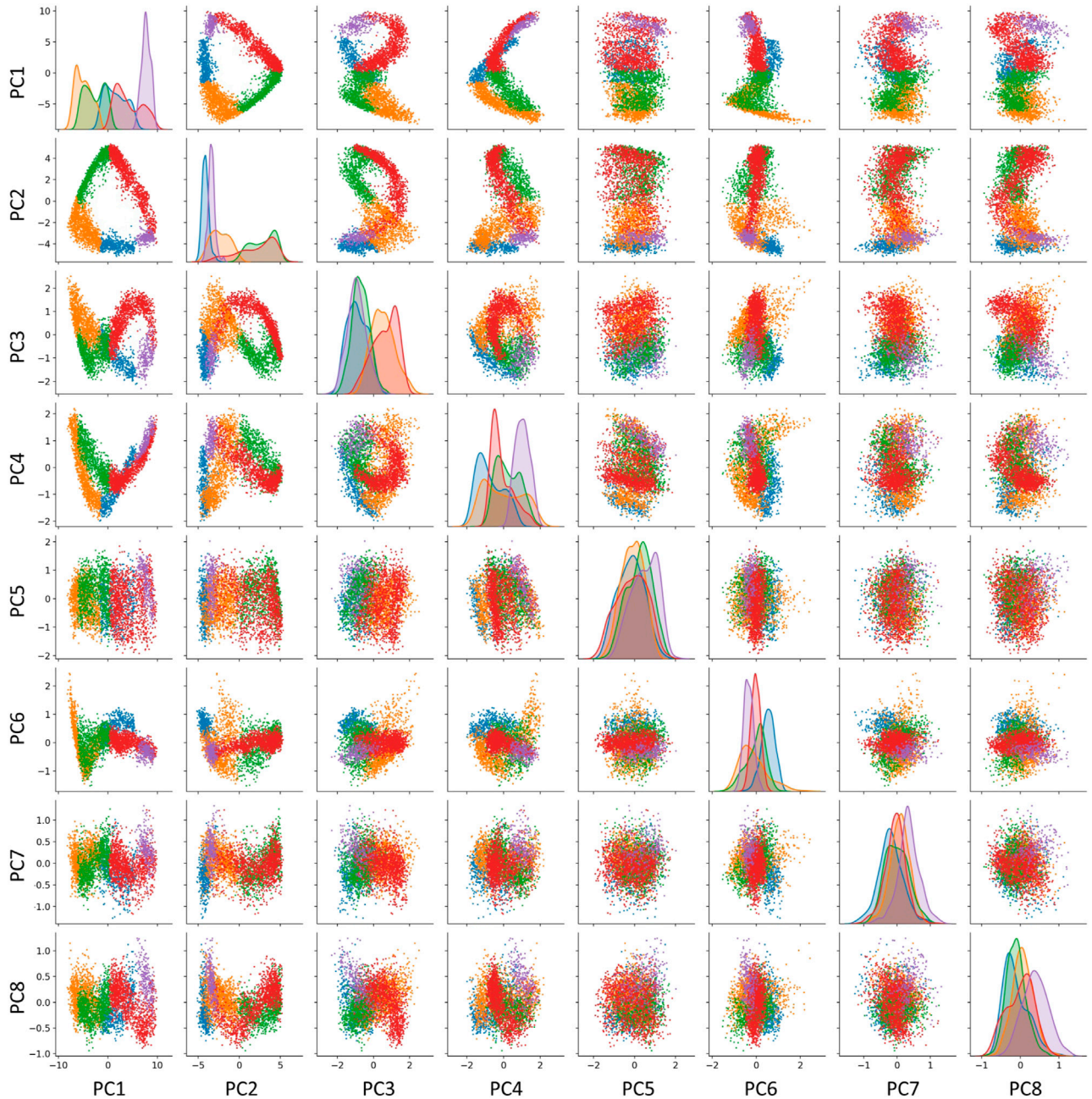


FIGURE 5 | Visualizing the transcriptomic cell cycle trajectory of CHLA9 cell line in projections on the first eight principal components, computed in the subspace of known cell cycle genes. The data points are partitioned according to the segmentation of the CCT into five transcriptomic epochs, also shown in **Figure 1**, 0-A (blue), A-B (orange), B-C (green), C-D (red), D-E (purple).

and lacks any notion of physical time and connection to the underlying kinetics of the lumped expression of genes involved in S phase and G2/M phases. A simple way to make it more concrete but still analytically tractable consists in introducing explicit processes of synthesis and degradation of the corresponding quantities, with kinetic rates changing in time. The simplest form of such dependence is piecewise-constant, with changes in the value of kinetic rates corresponding to the observed switches between transcriptional epochs of cell cycle progression.

Assuming the same epochs of cell cycle progression as above, and the same notations for variables (S, M, lumped expression of genes involved in S and G2/M phases correspondingly), their dynamics can be expressed as:

$$\begin{cases} \dfrac{dS}{dt} = k_t^S(t) - k_d^S(t)S \\ \dfrac{dM}{dt} = k_t^M(t) - k_d^M(t)M \end{cases}. \quad (1)$$

These equations must be accompanied by circular boundary conditions

$$\begin{cases} S(T) = S_f S(0) \\ M(T) = M_f M(0) \end{cases}, \quad (2)$$

where S_f, $M_f > 1$ are some numbers describing the drop of the lumped cell cycle variables after the moment of cell division. The most natural choice for them is S_f, $M_f = 2$, as before: however, here we prefer not to fix these parameters and rather fit them from the actually observed trajectory.

There exist several reasons for which S_f and M_f might appear in the range $1 \le S_f, M_f \le 2$ and not be equal. The most important of them is the technical biases introduced by sampling a limited amount of RNA, in the process of single cell transcriptome sequencing. It can lead to the situation when after cell division, the amount of RNA decreases non-uniformly between molecular processes. In particular, in all our experiments, we do observe the total amount of RNA reads does not decrease exactly by 2.0 and is rather close to 1.7-1.8. The decrease of the individual gene expression after cell division in terms of the number of reads, forms a bell-shaped distribution around this value with standard deviation close to 0.2.

Equation 1 with piecewise-constant in time kinetic rates and the boundary conditions 2) can be solved analytically for arbitrary number of levels in the piecewise-constant functions $k_t(t)$, $k_d(t)$. The resulting dynamics in the plane $\log S(t), \log M(t)$ represents a cell cycle trajectory parameterized by physical time, which consists of piecewise-smooth segments of three types. If a segment is characterized by $k_t^S(t) = k_t^M(t) = 0$ then the corresponding segment is linear in the logarithmic coordinates (since the underlying dynamics is exponentially decaying). If a segment is characterized by $k_d^S(t) = k_d^M(t) = 0$ then the corresponding segment is also linear in both logarithmic and initial coordinates. For a segment where at least one degradation k_d^* and one production kinetic rate k_t^* are positive, the dynamics follows a nonlinear curve in the logarithmic space, which remains monotonous (each of the coordinates does not change the derivative sign). The nonlinearity of the segment becomes important when one of the variables is in a stage exponentially increasing or decreasing, while the other is in a linear or close to saturation stage. Otherwise, the segment remains close to a line in logarithmic coordinates.

In order to choose the number of constant levels of the kinetic rates, we studied the averaged RNA velocity values along the cell cycle as a function of pseudotime (see **Figures 6A,B**). For the S variable, we decided to keep only one non-zero level of $k_t^S(t)$ during the transcriptional epoch Ts, and two levels of $k_d^S(t)$, one for the exit from mitosis epoch and one for the rest of the dynamics. The choice was similar for M variable, but we took into account that a boost of expression of the lumped $G2/M$ genes is visible in the beginning of the transcriptional epoch T2s, just after switching off the S phase genes. During mitosis we assumed that all production rates are zero, corresponding to the lack of transcription in the M phase. The resulting choice of levels for the kinetic rates is shown in **Figure 6C**.

The advantage of the proposed simple model of cell cycle trajectory is that it is fully analytically tractable and its parameters can be uniquely fit to the cell cycle trajectory observed in single cell data, given some biologically meaningful constraints. Thus, assuming that the duration of mitosis is by order of magnitude faster than the T1s epoch, for CHLA9 cell line one estimates the ratio between transcriptional epochs T2s and T1s close to 1.0 and the value of transcriptional boost of G2/M genes in T2s epoch close to 2.5-fold (**Figure 6C**). The determined values of all other parameters can be found in the Jupyter notebook at https://github.com/auranic/CellCycleTrajectory_SegmentModel.

3.7 Fitting Parameters of the Simple Kinetic Cell Cycle Model

Using the choice of levels for piecewise constant kinetic rates shown in **Figure 6C**, we could derive the dependence of the initial state of the cell cycle from the kinetic rates and the durations of four transcriptional epochs T_1, T_{1s}, T_{2s}, T_m:

$$\begin{cases} S(0) = \dfrac{k_t^S}{k_d^{S,2}} \dfrac{e^{k_d^{S,2}T_{1s}} - 1}{S_f e^{k_d^{S,2}(T_{1s}+T_{2s}+T_m)} - e^{-k_d^S T_1}} \\ M(0) = \dfrac{k_t^M}{k_d^{M,2}} \dfrac{p \cdot e^{k_d^{M,2}(T_{1s}+T_{2s})} - (p-1)e^{k_d^{M,2}T_{1s}} - 1}{M_f e^{k_d^{M,2}(T_{1s}+T_{2s}+T_m)} - e^{-k_d^M T_1}}. \end{cases} \quad (3)$$

Starting from the initial point of the trajectory $S(0)$, $F(0)$ it is possible to analytically write down the coordinates of all other borders of the transcriptional epochs:

$$\begin{cases} S(T_1) = S(0)e^{-k_d^S T_1} \\ M(T_1) = M(0)e^{-k_d^M T_1} \\ S(T_1+T_{1s}) = \dfrac{k_t^S}{k_d^{S,2}}\left(1 - \left(1 - \dfrac{k_d^{S,2}}{k_t^S}S(T_1)\right)e^{-k_d^{S,2}T_{1s}}\right) \\ M(T_1+T_{1s}) = \dfrac{k_t^M}{k_d^{M,2}}\left(1 - \left(1 - \dfrac{k_d^{M,2}}{k_t^M}{}^{*}M(T_1)\right)e^{-k_d^{M,2}T_{1s}}\right) \\ S(T_1+T_{1s}+T_{2s}) = S(T_1+T_{1s})e^{-k_d^{S,2}T_{2s}} \\ M(T_1+T_{1s}+T_{2s}) = \dfrac{p \cdot k_t^M}{k_d^{M,2}}\left(1 - \left(1 - \dfrac{k_d^{M,2}}{p \cdot k_t^M}{}^{*}M(T_1+T_{1s})\right)e^{-k_d^{M,2}T_{2s}}\right) \\ S(T) = S(T_1+T_{1s}+T_{2s})e^{-k_d^{S,2}T_m} \\ M(T) = M(T_1+T_{1s}+T_{2s})e^{-k_d^{M,2}T_m}, \end{cases} \quad (4)$$

where $T = T_1 + T_{1s} + T_{2s} + T_m$ is the full duration of the cell cycle. One can estimate the position of these points from the analysis of observed cell cycle trajectory curvature ((s_0, m_0), (s_1, m_1), ($s_{\max}$, m_s), (s_m, $m_{\max}$), (s_t, m_t), shown by red points in **Figure 6D**)) by requiring that the model trajectory should pass as close as possible to them. This defines an optimization problem which can be easily solved numerically by iterations, using the simplest

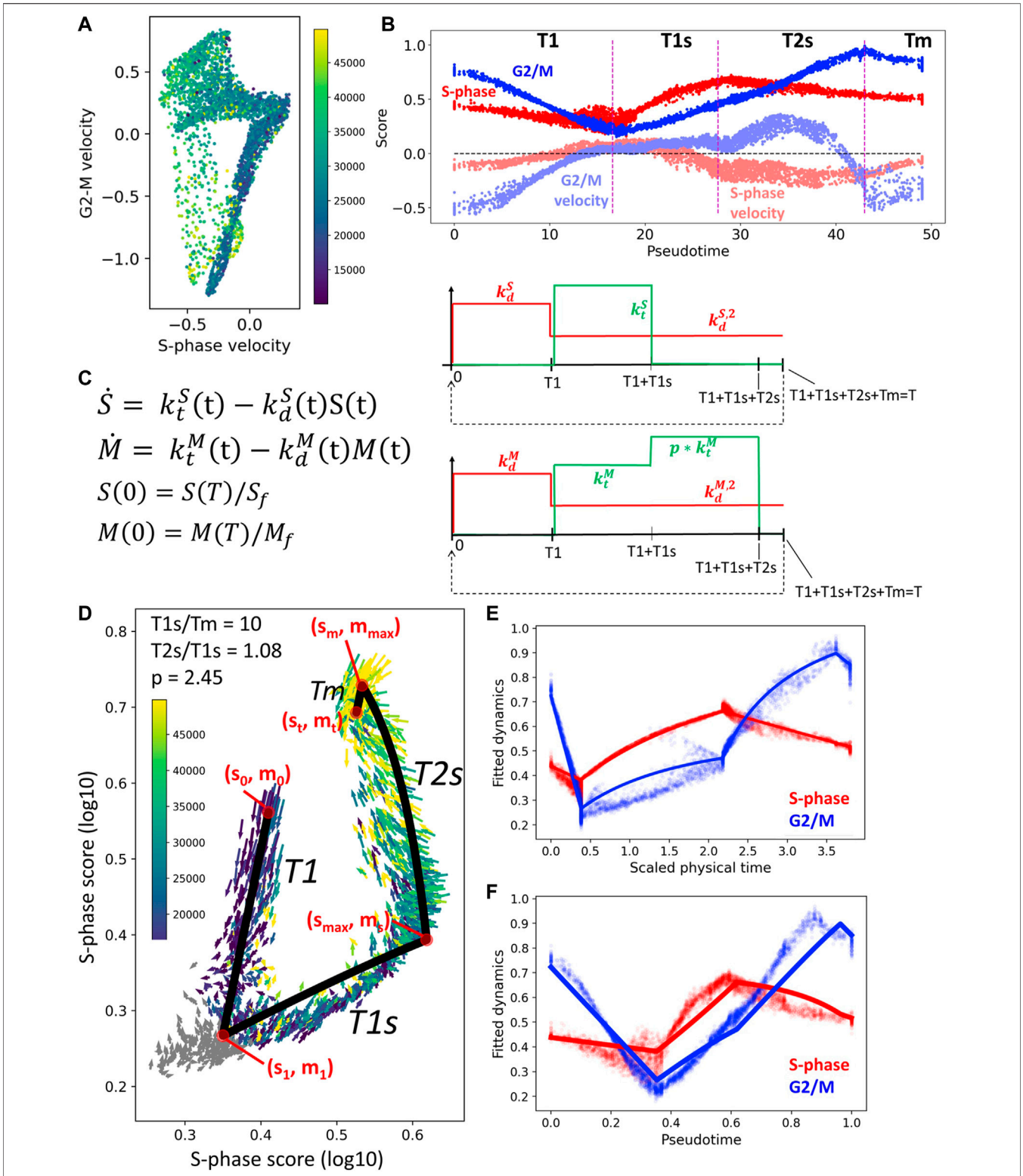


FIGURE 6 | Simple kinetic model of cell cycle transcriptome dynamics. **(A)** Mean RNA velocity values for S-phase and G2/M genes. **(B)** Pseudotemporal dynamics of S-phase and G2/M scores (shown with more intense color) and mean RNA velocity values (shown with semi-transparent color). **(C)** Description of the simple kinetic model of cell cycle transcriptome. Model equations are shown on the left and the changes in the values of kinetic rates (degradation, in red, and synthesis, in green). **(D)** Result of fitting the model dynamics to cell cycle transcriptome dynamics observed in CHLA9 cell line. **(E,F)** Inferred physical time and pseudotemporal dynamics of cell cycle transcriptome in CHLA9 cell line.

fixed-point algorithm. The details of parameter fitting are provided in the Jupyter notebook at https://github.com/auranic/CellCycleTrajectory_SegmentModel.

We note that this optimization does not allow us to determine all the model parameters uniquely, since they enter in the aforementioned optimization functional as certain combinations (as simple rational functions), namely, $\frac{k_t^S}{k_d^{S,2}}$, $\frac{k_t^M}{k_d^{M,2}}$, $k_d^S T_1$, $k_d^M T_1$, $k_d^{S,2} T_{1s}$, $k_d^{M,2} T_{1s}$, $k_d^{S,2} T_{2s}$, $k_d^{M,2} T_{2s}$, $k_d^{S,2} T_m$, $k_d^{M,2} T_m$. Two other parameters M, S define the observed cell division vector in **(2)**. One extra parameter p denotes transcriptional production acceleration of $G2/M$ genes during the transcriptional epoch T2s compared to the transcriptional epoch T1s (**Figure 6C**). Not all these quantities are independent, some of them are connected through nonlinear relations:

$$\frac{k_d^{S,2} \cdot T_{2s}}{k_d^{S,2} \cdot T_{1s}} = \frac{k_d^{M,2} \cdot T_{2s}}{k_d^{M,2} \cdot T_{1s}}, \frac{k_d^{S,2} \cdot T_m}{k_d^{S,2} \cdot T_{1s}} = \frac{k_d^{M,2} \cdot T_m}{k_d^{M,2} \cdot T_{1s}}, \tag{5}$$

which overall gives 11 independent combinations of parameters provided 10 measurable coordinates of cell trajectory turning points in **Figure 6D**.

Altogether, this means that 1) one needs to introduce at least one additional constraint in order to make the trajectory reconstruction unique and 2) physical time of the epochs T_1, T_{1s}, T_{2s}, T_m can not be uniquely computed from the cell cycle trajectory observed in the plane of S-, G2/M-phase scores. From the analysis of **Eq. 5** it follows that the model can be uniquely parameterized if one will constrain one of the three quantities p, $\frac{T_{2s}}{T_{1s}}$, $\frac{T_m}{T_{1s}}$. Finally, it is convenient to fix the durations T_1, T_{1s} to some arbitrary values which allows to determine parameters k_d^S, k_d^M and the ratios $\frac{T_{2s}}{T_{1s}}$, $\frac{T_m}{T_{1s}}$.

In our numerical experiments, we fixed the values of T_1 and T_{1s} to their corresponding pseudotemporal durations (as the corresponding fractions of the total length of the cell cycle trajectory). We also fixed the ratio $\frac{T_m}{T_{1s}} = 10$, assuming that the mitosis must be fast in physical time compared to the transcriptional epoch including activating the expression of the genes involved in the S-phase.

3.8 Simulating Cell Cycle Trajectories With Various Durations of Temporal Transcriptional Epochs

After fitting the kinetic parameters for an observable in the S-phase vs G2/M score plane cell cycle trajectory, one can perturb the parameters and investigate how the trajectory geometry depends on them.

In real life scRNASeq datasets, we observe that CCT geometry can appear very different in various biological systems. When projecting onto the plane of standard scores of S-phase and G2/M phase genes, scRNASeq datasets might not always reveal the circular nature of CCT. In some cases, the circular structure is not at all detectable via this projection, (**Figure 7**), and the two scores might be connected via a strong positive or negative correlation. Also, in some systems we observed co-existence of several CCT shapes, like it is the case in the U2OS cell line dataset (GSE146773). The univariate histograms of two score distributions might be characterized by bi- or uni-modal character.

Quite strikingly, we were able to reproduce these patterns qualitatively by fitting the kinetic parameters to the CHLA9 scRNASeq dataset, and then by manipulating the durations of $T1$, $T1s$ and $T2s$ transcriptional epochs and producing computer-simulated trajectory examples. Thus, significant reduction in the duration of both $T1$ and $T1s$ epochs led to the negative correlation pattern between S-phase and G2/M scores. This could be interpreted as drastic reduction of the G1 cell cycle phase. In real life datasets, such pattern has been observed in human embryonic stem cells (dataset GSE85917).

If both $T1$ and $T2s$ were shortened then this led to the increase of the positive correlation between two scores, (**Figure 7**). This pattern was indeed observed in human bone marrow and human neural epithelial stem cell-related single cell datasets (GSE99095 and GSE81475).

3.9 Predicting Cell Line Doubling Time From the Geometrical Properties of Cell Cycle Trajectory

The developed simple kinetic model leads to a simple prediction which can be validated: the total length of the transcriptomic cell cycle trajectory must diminish in rapidly dividing cells. This can be interpreted as a consequence of the fact that in a rapid proliferation process, during the post-mitotic G1 phase (T1 transcriptional epoch), there is not enough time to degrade all mitotic transcripts produced before the cell division moment, so they are reused in the consequent cell cycle phases, shortening the subsequent G1 phase.

We verified this prediction in a relatively large collection of cell line scRNASeq datasets. Using the data from Cellosaurus database, we identified those few ones for which the cell line doubling time has been estimated, and for which the number of available good quality single cell profiles exceeded 300.

We used the total length of the principal circle fit in the 2D plane of the scaled to maximum equals 1 cell cycle phase scores, as a proxy to quantify the level of CCT contraction (see Methods). This measure was correlated with cell line doubling time in hours. Two cell lines CHLA10 and SCC25 appeared to be strong outliers from otherwise significant positive regression line (Pearson correlation 0.931, p-value = 10^{-5}) (**Figure 8**). When this regression line was used as a predictor, CHLA10 cell line was predicted to have doubling time around 64 h (instead of determined by database search of around 32 h) and for SCC25 around 78 instead of 50 h. It is known that cell line doubling time can vary depending on the growth conditions, so we hypothesized that this variability could explain the appearance of two outliers. If two of them were kept in the regression calculation, it remained significant but less strong (Pearson correlation 0.67, p-value = 0.01).

3.10 Code Availability

The Python notebooks allowing the reader to reproduce all the computations presented in this manuscript are freely available

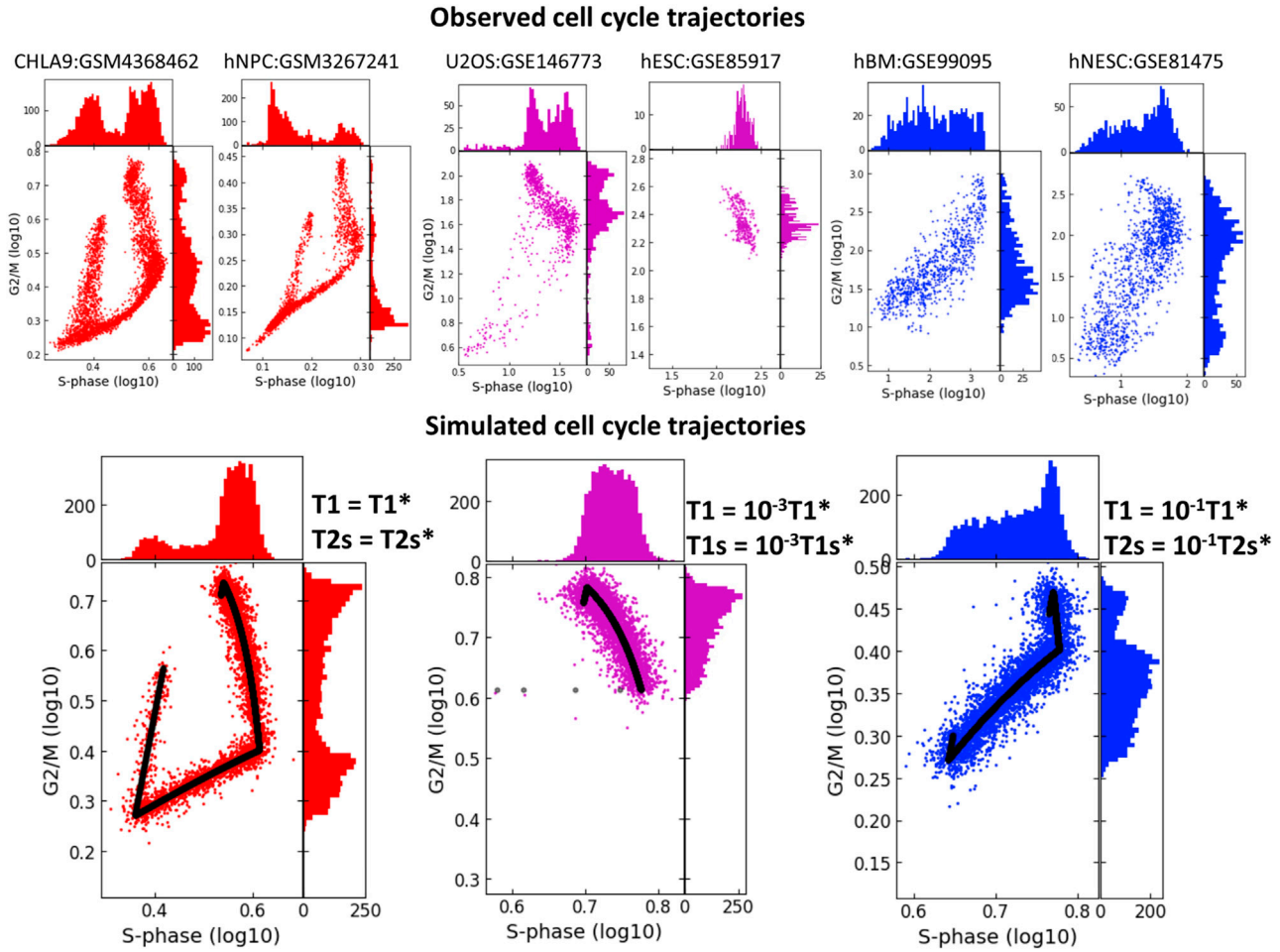


FIGURE 7 | Studying the effect of shortening the durations of transcriptional epochs T1 and T1s or T1 and T2s on the geometry of cell cycle trajectory projected onto the S-phase and G2/M-phase scores plane. The simulated trajectories (in the lower part of the figure) are produced by taking the parameters of the CHLA9 fit of model dynamics (red plot) and changing the durations of T1 and T1s epochs (violet plot) or the durations of T1 and T2s epochs (blue plot). Each simulation shows the trajectory (black line) sampled with Laplacian noise added, with score distribution histograms shown at the plot margins. The upper part of the plot shows six real-life cell cycle trajectories observed in different systems, with GEO identifiers indicated. In each plot title either cell line name is provided, or hNPC means human neural precursor cells, hESC - human embryonic stem cell, hBM - human bone marrow, hNESC - human neural epithelial stem cell.

from https://github.com/auranic/CellCycleTrajectory_SegmentModel.

4 DISCUSSION

This paper provides a framework for analyzing the cell cycle trajectories using single cell omics measurements such as scRNASeq data. Unlike the previously suggested model of the trajectory as a flat circle, we provide arguments that at least in some conditions the piecewise-linear in logarithmic coordinates approximation appears to fit the single cell transcriptomic data and to be biologically tractable. In particular, it allows us to delineate transcriptional epochs of cell cycle at which the corresponding segment of the trajectory remains close to linear in logarithmic coordinates which corresponds to locally allometric changes of the transcriptome.

We suggest two modeling formalisms to recapitulate the cell cycle transcriptomic dynamics as a sequence of switches. The first one is purely phenomenological and describes the dynamics as a change of states of a hidden automaton, leading to the switches of parameters of allometric growth, followed by a shift representing the cell division event. The advantage of this formalism is that it allows us to treat most general properties of cell cycle trajectory geometry.

In particular, we could prove a fundamental theorem on the number of intrinsic cell cycle states, which connects the number of linear segments in the trajectory and the embedding dimensionality of the cell cycle trajectory. The nature of this theorem, relying on “general position”-type arguments, is reminiscent of the well-known results imposing constraints on the number of the system’s internal states and the effective dimensionality of its environment, in several fields of science. For example, the Gause’s law of competitive exclusion and its

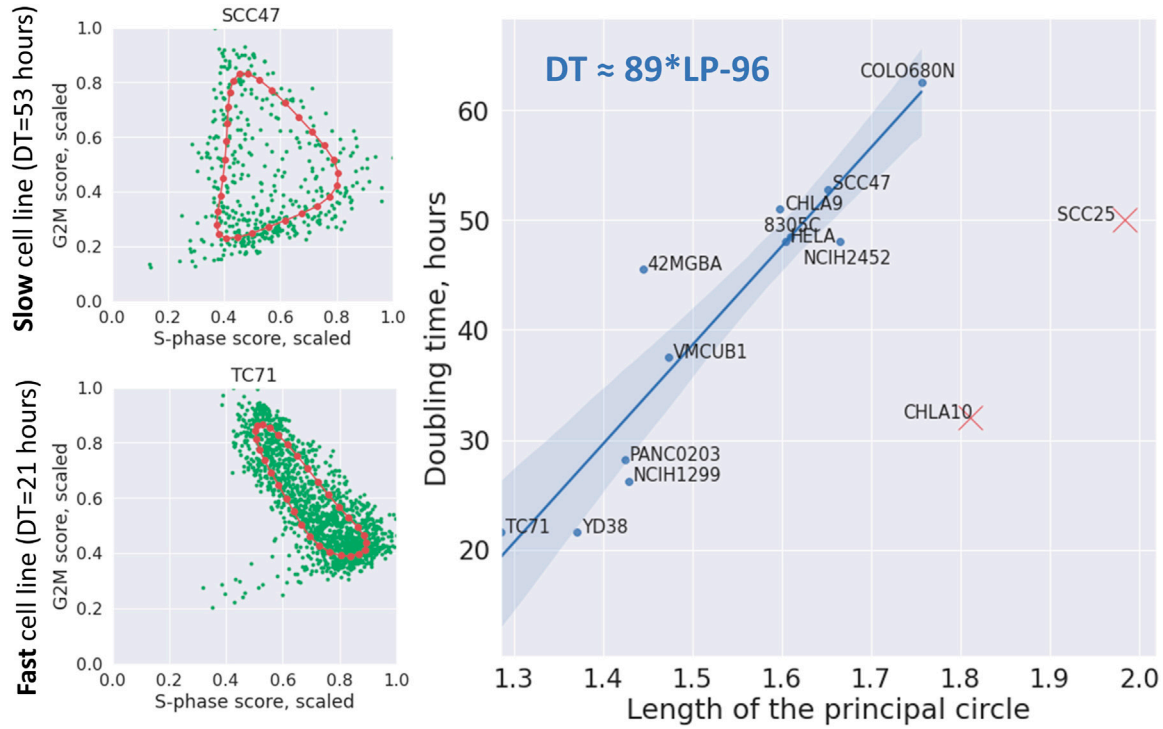


FIGURE 8 | Dependence of cell line doubling time (DT) on the length of the principal circle (LP) approximating the cell cycle trajectory in the 2D plane of scaled (divided by the maximum value) S-phase and G2M scores. On the left two examples of principal circles are shown in red, and cells in green. On the right the linear regression line with confidence intervals is shown connecting the length of the principal circle with cell line doubling time (Pearson correlation 0.931, p-value = 10^{-5}). The regression formula is shown on the plot in top left corner. Two cell lines indicated by red crosses were eliminated from the regression as evident outliers.

generalizations states that the number of competing species is limited by the effective number of resources, characterizing the environment (Gauze, 1934; Gorban, 2007). The famous Gibbs' phase rule in thermodynamics connects the effective number of the intensive variables with the number of components and phases in a system at thermodynamic equilibrium (Gibbs, 1961; Alper, 1999). All these results are also similar in terms of practical difficulties related to determining the effective system's dimensionality.

From the physico-chemical point of view, the effective dimensionality is the number of the substances "lumps" in the cell cycle kinetics. Lumping-analysis produces a partition of all chemical species into a few groups and then considers these groups ("lumps") as independent entities (Wei and Kuo, 1969). "Amounts" of these lumps are the combinations of the amounts of the chemical species (Li and Rabitz, 1989; Li and Rabitz, 1990). The theorem on the number of intrinsic cell cycle states that the number of lumps n does not exceed the number of the internal states of the cell cycle transcription machinery. This means that kinetics allows reduction of the huge-dimensional space of all components to $n \leq m$ number of aggregated lumps.

The second modeling formalism that we suggested connects the geometric properties of the cell cycle trajectory to the underlying transcriptional kinetics and physical time. It uses the simplest chemical kinetics equations with kinetic rates represented as piecewise-constant functions of time. We show that the suggested model is fully analytically tractable and, under some biologically transparent assumptions, allows unique determination of its independent parameter combinations. This type of modeling allowed us to explicitly study the relation between pseudotime and physical time.

The precise connection between physical time and pseudotime (geometric time) in the cell cycle is worth studying in more detail since this is the central question in the dynamic phenotyping approach in general (Golovenkin et al., 2020). Some of these relations can be potentially quantified from exploring the variations of point density along the inferred trajectories (Chen et al., 2019). Related to this, one can expect non-trivial phenomena in studying the cell cycle trajectory, such as effects of partial cell population synchronization under assumption of equal cell cycle durations in individual cells. This effect can lead to the appearance of density peaks in the reconstructed cell cycle trajectories that cannot be explained by nonlinear relation between physical time and pseudotime (Gorban, 2007).

As one of the applications of the suggested modeling formalism, we performed several numerical experiments on changing the durations of the transcriptional epochs overlapping with G1 or G2 cell cycle phases. We observed that these parameters might have a drastic effect on the shape of the CCT geometry and the form of the univariate variable distributions. This model prediction can be qualitatively confirmed by observing CCT properties of several *in vitro* and *in vivo* systems. The effect of CCT shrinkage might be

relevant in characterizing the cell cycle properties in various conditions: for example, when one can manipulate the activity of an oncogene (Aynaud et al., 2020). We show that the CCT geometry can be predictive to estimate the cell line doubling time which can be a proxy of cell cycle duration.

The relation between transcriptomic dynamics and the established definitions of cell cycle phases and cell cycle checkpoints has been discussed and even quantified using standard molecular biology techniques (Giotti et al., 2019; Hsiao et al., 2020). In this study, we deliberately leave open the question on defining the exact cell cycle phase borders from the transcriptomic CCT geometry. We found that this relation can not be the exact match: one of the reasons for this is delayed production of proteins, and dependence of the cell cycle progression from post-translational protein modifications. The transcriptomic dynamics is relatively slow, and activation of protein synthesis is switched on in advance, leaving time for producing enough proteins needed at a certain stage of the cell cycle molecular program. Same is true for the process of degradation of RNAs involved in cell cycle: a cell needs enough time after mitosis to degrade all cell cycle-related transcripts.

The suggested formalism is not limited to transcriptomic data. It looks promising to analyze the geometrical properties of cell cycle trajectory measured in unsynchronized cell populations profiled at various levels of molecular description, including epigenetics and protein expression, when the datasets of sufficient volume and quality will become available.

A more mechanistic description of the cell cycle has been already proposed in the context of yeast or mammalian cells (Tyson, 1991; Novák and Tyson, 2004). The mathematical models can be based on chemical kinetics or on discrete or hybrid frameworks but in all cases, the difficulty when constructing these models is to select the genes that can capture the main features of the cell cycle and the different events that allow the switch from one phase to another. We anticipate that the type of analyses presented here could orient the choice of these genes and inform on their dynamics.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article. The processed scRNA-Seq data used in this study are available from http://doi.org/10.5281/zenodo.5017357. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

AZ, MS, and AG designed the study and performed analytical calculations. AZ, AF, CG, and AC contributed to computer simulations and code development. AZ, AC, AF, and CG participated in data collection and analysis. LC, EB, AC, and CG interpreted the analysis results. AZ and MS wrote the manuscript draft. All authors read the manuscript and edited it.

FUNDING

This work was supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute), The work was supported by the Ministry of Science and Higher Education of the Russian Federation (Project No. 075-15-2021-634) and by European Union's Horizon 2020 program (grant No. 826121, iPC project). These funding sources had no role in the design, execution and interpretation of the results of this study.

REFERENCES

Albergante, L., Bac, J., and Zinovyev, A. (2019). "Estimating the Effective Dimension of Large Biological Datasets Using Fisher Separability Analysis," in *Proceedings of the International Joint Conference on Neural Networks*. doi:10.1109/ijcnn.2019.8852450

Albergante, L., Mirkes, E., Bac, J., Chen, H., Martin, A., Faure, L., et al. (2020). Robust and Scalable Learning of Complex Intrinsic Dataset Geometry via ElPiGraph. *Entropy* 22 (3), 296. doi:10.3390/e22030296

Alper, J. S. (1999). The Gibbs Phase Rule Revisited: Interrelationships between Components and Phases. *J. Chem. Educ.* 76 (11), 1567–1569. doi:10.1021/ed076p1567

Altinok, A., Lévi, F., and Goldbeter, A. (2007). A Cell Cycle Automaton Model for Probing Circadian Patterns of Anticancer Drug Delivery. *Adv. Drug Deliv. Rev.* 59 (9-10), 1036–1053. doi:10.1016/J.ADDR.2006.09.022

Aynaud, M.-M., Mirabeau, O., Gruel, N., Grossetête, S., Boeva, V., Durand, S., et al. (2020). Transcriptional Programs Define Intratumoral Heterogeneity of Ewing Sarcoma at Single-Cell Resolution. *Cel Rep.* 30 (6), 1767–1779. doi:10.1016/j.celrep.2020.01.049

Azizi, E., Carr, A. J., Plitas, G., Cornish, A. E., Konopacki, C., Prabhakaran, S., et al. (2018). Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* 174 (5), 1293–1308. doi:10.1016/j.cell.2018.05.060

Bac, J., Mirkes, E. M., Gorban, A. N., Tyukin, I., and Zinovyev, A. (2021). Scikit-dimension: a Python Package for Intrinsic Dimension Estimation. *Entropy Basel* 23:1368. doi:10.3390/e23101368

Bac, J., and Zinovyev, A. (2020). Lizard Brain: Tackling Locally Low-Dimensional yet Globally Complex Organization of Multi-Dimensional Datasets. *Front. Neurorobot.* 13. doi:10.3389/fnbot.2019.00110

Bairoch, A. (2018). The Cellosaurus, a Cell-Line Knowledge Resource. *J. Biomol. Tech.* 29 (2), 25–38. doi:10.1038/s41588-020-00726-610.7171/jbt.18-2902-002

Bauer, E. (1935). *Theoretical Biology*. Budapest: Akademiai Kiado.

Bernard, D., Mondesert, O., Gomes, A., Duthen, Y., Lobjois, V., Cussat-Blanc, S., et al. (2019). A Checkpoint-Oriented Cell Cycle Simulation Model. *Cell CycleZIP* 18 (8), 795–808. doi:10.1080/15384101.2019.1591125/SUPPL_FILE/KCCY_A_1591125_SM1721

Bray, N. L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal Probabilistic RNA-Seq Quantification. *Nat. Biotechnol.* 34 (5), 525–527. doi:10.1038/nbt.3519

Chen, H., Albergante, L., Hsu, J. Y., Lareau, C. A., Lo Bosco, G., Guan, J., et al. (2019). Single-cell Trajectories Reconstruction, Exploration and Mapping of Omics Data with STREAM. *Nat. Commun.* 10 (1). doi:10.1038/s41467-019-09670-4

Chen, K. C., Calzone, L., Csikasz-Nagy, A., Cross, F. R., Novak, B., and Tyson, J. J. (2004). Integrative Analysis of Cell Cycle Control in Budding Yeast. *MBoC* 15 (8), 3841–3862. doi:10.1091/mbc.E03-11-0794

Chen, K. C., Calzone, L., Csikasz-Nagy, A., Cross, F. R., Novak, B., and Tyson, J. J. (2004). Integrative Analysis of Cell Cycle Control in Budding Yeast. *MBoC* 15 (8), 3841–3862. doi:10.1091/mbc.e03-11-0794

Deritei, D., Rozum, J., Ravasz Regan, E., and Albert, R. (20199201). A Feedback Loop of Conditionally Stable Circuits Drives the Cell Cycle from Checkpoint to Checkpoint. *Sci. Rep.* 9 (11), 1–19. doi:10.1038/s41598-019-52725-1

Dominguez, D., Tsai, Y.-H., Gomez, N., Jha, D. K., Davis, I., and Wang, Z. (2016). A High-Resolution Transcriptome Map of Cell Cycle Reveals Novel Connections between Periodic Genes and Cancer. *Cell Res* 26 (8), 946–962. doi:10.1038/cr.2016.84

Fauré, A., Naldi, A., Chaouiya, C., and Thieffry, D. (2006). Dynamical Analysis of a Generic Boolean Model for the Control of the Mammalian Cell Cycle. *BioinformaticsOxford Acad.* 22 (14), e124–31. doi:10.1093/bioinformatics/btl210

Gauze, G. F. (1934). *The Struggle for Existence*. Baltimore: The Williams I& Wilkins company. Available from: https://www.biodiversitylibrary.org/item/23409.

Gibbs, J. W. (1961). *The Scientific Papers*, 1. New York: Thermodynamics. Dover.

Giotti, B., Chen, S.-H., Barnett, M. W., Regan, T., Ly, T., Wiemann, S., et al. (2019). Assembly of a Parts List of the Human Mitotic Cell Cycle Machinery. *J. Mol. Cel Biol.* 11 (8), 703–718. doi:10.1093/jmcb/mjy063

Golovenkin, S. E., Bac, J., Chervov, A., Mirkes, E. M., Orlova, Y. V., Barillot, E., et al. (2020). Trajectories, Bifurcations, and Pseudo-time in Large Clinical Datasets: Applications to Myocardial Infarction and Diabetes Data. *GigaScience* 9 (11), 1–20. doi:10.1093/gigascience/giaa128

Gorban, A. N. (2018). Model Reduction in Chemical Dynamics: Slow Invariant Manifolds, Singular Perturbations, Thermodynamic Estimates, and Analysis of Reaction Graph .

Gorban, A. N. (2007). Selection Theorem for Systems with Inheritance. *Math. Model. Nat. Phenom.* 2 (4), 1–45. doi:10.1051/mmnp:2008024

Holford, N. H. G., and Anderson, B. J. (2017). Allometric Size: the Scientific Theory and Extension to normal Fat Mass. *Eur. J. Pharm. Sci.* 109, S59–S64. doi:10.1016/j.ejps.2017.05.056

Hsiao, C. J., Tung, P., Blischak, J. D., Burnett, J. E., Barr, K. A., Dey, K. K., et al. (2020). Characterizing and Inferring Quantitative Cell Cycle Phase in Single-Cell RNA-Seq Data Analysis. *Genome Res.* 30 (4), 611–621. doi:10.1101/gr.247759.118

Hunt, T. (1991). Cell Cycle Gets More Cyclins. *Nature* 350 (6318), 462–463. doi:10.1038/350462a0

Hunt, T., Nasmyth, K., and Novák, B. (2011). The Cell Cycle. *Phil. Trans. R. Soc. B* 366 (1584), 3494–3497. doi:10.1098/rstb.2011.0274

Ingolia, N. T., and Murray, A. W. (2004). The Ups and downs of Modeling the Cell Cycle, *Curr Biol*, 14 R771–R777. doi:10.1016/j.cub.2004.09.018

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2020). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 48 (D1), D498–D503. doi:10.1093/nar/gkz1031

Kinker, G. S., Greenwald, A. C., Tal, R., Orlova, Z., Cuoco, M. S., McFarland, J. M., et al. (2020). Pan-cancer Single-Cell RNA-Seq Identifies Recurring Programs of Cellular Heterogeneity. *Nat. Genet.* 52 (11), 1208–1218. doi:10.1038/s41588-020-00726-6

Kotani, S. (2002). A Computational Model of Mammalian Cell Cycle Using Petri Nets. *Genome Inform.* 460, 459–460. doi:10.11234/GI1990.13.459

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., et al. (2018). RNA Velocity of Single Cells. *Nature* 560 (7719), 494–498. doi:10.1038/s41586-018-0414-6

Leng, N., Chu, L.-F., Barry, C., Li, Y., Choi, J., Li, X., et al. (2015). Oscope Identifies Oscillatory Genes in Unsynchronized Single-Cell RNA-Seq Experiments. *Nat. Methods* 12 (10), 947–950. doi:10.1038/nmeth.3549

Li, G., and Rabitz, H. (1990). A General Analysis of Approximate Lumping in Chemical Kinetics. *Chem. Eng. Sci.* 45 (4), 977–1002. doi:10.1016/0009-2509(90)85020-E

Li, G., and Rabitz, H. (1989). A General Analysis of Exact Lumping in Chemical Kinetics. *Chem. Eng. Sci.* 44 (6), 1413–1430. doi:10.1016/0009-2509(89)85014-6

Liang, S., Wang, F., Han, J., and Chen, K. (2020). Latent Periodic Process Inference from Single-Cell RNA-Seq Data. *Nat. Commun.* 11 (1441), 11. doi:10.1038/s41467-020-15295-9

Liu, Z., Lou, H., Xie, K., Wang, H., Chen, N., Aparicio, O. M., et al. (2017). Reconstructing Cell Cycle Pseudo Time-Series via Single-Cell Transcriptome Data. *Nat. Commun.* 8 (1), 1–9. doi:10.1038/s41467-017-00039-z

Mahdessian, D., Cesnik, A. J., Gnann, C., Danielsson, F., Stenström, L., Arif, M., et al. (2021). Spatiotemporal Dissection of the Cell Cycle with Single-Cell Proteogenomics. *Nature* 590 (7847), 649–654. doi:10.1038/s41586-021-03232-9

Miller, H. E., Gorthi, A., Bassani, N., Lawrence, L. A., Iskra, B. S., and Bishop, A. J. R. (2020). Reconstruction of ewing Sarcoma Developmental Context from Mass-Scale Transcriptomics Reveals Characteristics of Ewsr1-Fli1 Permissibility. *Cancers* 12 (4), 948. doi:10.3390/cancers12040948

Noel, V., Grigoriev, D., Vakulenko, S., and Radulescu, O. (2012). Tropical Geometries and Dynamics of Biochemical Networks Application to Hybrid Cell Cycle Models. *Electron. Notes Theor. Comp. Sci.* 284, 75–91. doi:10.1016/j.entcs.2012.05.016

Noël, V., Vakulenko, S., and Radulescu, O. (2013). A Hybrid Mammalian Cell Cycle Model. *Electron. Proc. Theor. Comput. Sci.* 125, 68–83. doi:10.4204/EPTCS.125.5

Novák, B., and Tyson, J. J. (2004). A Model for Restriction point Control of the Mammalian Cell Cycle. *J. Theor. Biol.* 230 (4), 563–579. doi:10.1016/j.jtbi.2004.04.039

Packard, G. C. (2017). The Essential Role for Graphs in Allometric Analysis. *Biol. J. Linn. Soc.* 120 (2), 468–473.

Pretzsch, H. (2020). The Course of Tree Growth. Theory and Reality. *For. Ecol. Manag.* 478, 118508. doi:10.1016/j.foreco.2020.118508

Radulescu, O., Gorban, A. N., Zinovyev, A., and Noel, V. (2012). Reduction of Dynamical Biochemical Reactions Networks in Computational Biology. *Front. Gene* 3, 131. doi:10.3389/fgene.2012.00131

Schwabe, D., Formichetti, S., Junker, J. P., Falcke, M., and Rajewsky, N. (2020). The Transcriptome Dynamics of Single Cells during the Cell Cycle. *Mol. Syst. Biol.* 16 (11), e9946. doi:10.15252/msb.20209946

Shkolnik, E. M. (1989). "Dynamic Theory of Cell Cycle," in *Dynamics of Chemical and Biological Systems* (Russian: Nauka plc. Siberian branch), 159–190.

Sible, J. C., and Tyson, J. J. (2007). Mathematical Modeling as a Tool for Investigating Cell Cycle Control Networks. *Methods* 41 (2), 238–247. doi:10.1016/j.ymeth.2006.08.003

Singhania, R., Sramkoski, R. M., Jacobberger, J. W., and Tyson, J. J. (2011). A Hybrid Model of Mammalian Cell Cycle Regulation. *Plos Comput. Biol.* 7 (2), e1001077. doi:10.1371/JOURNAL.PCBI.1001077

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., et al. (2016). Dissecting the Multicellular Ecosystem of Metastatic Melanoma by Single-Cell RNA-Seq. *Science* 352 (6282), 189–196. doi:10.1126/science.aad0501

Tyson, J. J. (1991). Modeling the Cell Division Cycle: Cdc2 and Cyclin Interactions. *Proc. Natl. Acad. Sci.* 88 (16), 7328–7332. doi:10.1073/pnas.88.16.7328

Tzamali, E., Tzedakis, G., and Sakkalis, V. (2020). Modeling How Heterogeneity in Cell Cycle Length Affects Cancer Cell Growth Dynamics in Response to Treatment. *Front. Oncol.* 10, 1552. doi:10.3389/FONC.2020.01552/BIBTEX

Wei, J., and Kuo, J. C. W. (1969). Lumping Analysis in Monomolecular Reaction Systems. Analysis of the Exactly Lumpable System. *Ind. Eng. Chem. Fund.* 8 (1), 114–123. doi:10.1021/i160029a019

White, C. R., Marshall, D. J., Alton, L. A., Arnold, P. A., Beaman, J. E., Bywater, C. L., et al. (2019). The Origin and Maintenance of Metabolic Allometry in Animals. *Nat. Ecol. Evol.* 3 (4), 598–603. doi:10.1038/s41559-019-0839-9

Zetterberg, A., Larsson, O., and Wiman, K. G. (1995). What Is the Restriction point? *Curr. Opin. Cel Biol* 7, 835–842. doi:10.1016/0955-0674(95)80067-0

Zhou, X., Yang, M., Liu, Z., Li, P., Xie, B., and Peng, C. (2021). Dynamic Allometric Scaling of Tree Biomass and Size. *Nat. Plants* 7 (1), 42–49. doi:10.1038/s41477-020-00815-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Zinovyev, Sadovsky, Calzone, Fouché, Groeneveld, Chervov, Barillot and Gorban. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 15 February 2022
doi: 10.3389/fmolb.2021.800721

Dynamic Control Balancing Cell Proliferation and Inflammation is Crucial for an Effective Immune Response to Malaria

Anuj Gupta[1], *Mary R. Galinski*[2†] *and Eberhard O. Voit*[1*†]

[1]*The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, United States,* [2]*Emory Vaccine Center, Yerkes National Primate Research Center, Department of Medicine, Division of Infectious Diseases, Emory University, Atlanta, GA, United States*

Malaria has a complex pathology with varying manifestations and symptoms, effects on host tissues, and different degrees of severity and ultimate outcome, depending on the causative *Plasmodium* pathogen and host species. Previously, we compared the peripheral blood transcriptomes of two macaque species (*Macaca mulatta* and *Macaca fascicularis*) in response to acute primary infection by *Plasmodium knowlesi*. Although these two species are very closely related, the infection in *M. mulatta* is fatal, unless aggressively treated, whereas *M. fascicularis* develops a chronic, but tolerable infection in the blood. As a reason for this stark difference, our analysis suggests delayed pathogen detection in *M. mulatta* followed by extended inflammation that eventually overwhelms this monkey's immune response. By contrast, the natural host *M. fascicularis* detects the pathogen earlier and controls the inflammation. Additionally, *M. fascicularis* limits cell proliferation pathways during the log phase of infection, presumably in an attempt to control inflammation. Subsequent cell proliferation suggests a cell-mediated adaptive immune response. Here, we focus on molecular mechanisms underlying the key differences in the host and parasite responses and their coordination. *SICAvar* Type 1 surface antigens are highly correlated with pattern recognition receptor signaling and important inflammatory genes for both hosts. Analysis of pathogen detection pathways reveals a similar signaling mechanism, but with important differences in the glutamate G-protein coupled receptor (GPCR) signaling pathway. Furthermore, differences in inflammasome assembly processes suggests an important role of S100 proteins in balancing inflammation and cell proliferation. Both differences point to the importance of Ca^{2+} homeostasis in inflammation. Additionally, the kynurenine-to-tryptophan ratio, a known inflammatory biomarker, emphasizes higher inflammation in *M. mulatta* during log phase. Transcriptomics-aided metabolic modeling provides a functional method for evaluating these changes and understanding downstream changes in NAD metabolism and aryl hydrocarbon receptor (AhR) signaling, with enhanced NAD metabolism in *M. fascicularis* and stronger AhR signaling in *M. mulatta*. AhR signaling controls important immune genes like IL6, IFNγ and IDO1. However, direct changes due to AhR signaling could not be

OPEN ACCESS

Edited by:
Daniel Garrido, Pontificia Universidad Católica de Chile, Chile

Reviewed by:
Janet Cox-Singh, University of St Andrews, United Kingdom
Xinjun Zhang, University of California, Los Angeles, United States

***Correspondence:**
Eberhard O. Voit
eberhard.voit@bme.gatech.edu

[†]These authors have contributed equally to this work and share senior authorship

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *23 October 2021*
Accepted: *20 December 2021*
Published: *15 February 2022*

Citation:
Gupta A, Galinski MR and Voit EO (2022) Dynamic Control Balancing Cell Proliferation and Inflammation is Crucial for an Effective Immune Response to Malaria. Front. Mol. Biosci. 8:800721. doi: 10.3389/fmolb.2021.800721

established due to complicated regulatory feedback mechanisms associated with the AhR repressor (AhRR). A complete understanding of the exact dynamics of the immune response is difficult to achieve. Nonetheless, our comparative analysis provides clear suggestions of processes that underlie an effective immune response. Thus, our study identifies multiple points of intervention that are apparently responsible for a balanced and effective immune response and thereby paves the way toward future immune strategies for treating malaria.

Keywords: *Plasmodium knowlesi*, *Macaca fascicularis*, *Macaca mulatta*, kynurenine metabolism, tryptophan metabolism, AhR signaling, antigenic variation, surface antigens

HIGHLIGHTS

- *Macaca mulatta* and *Macaca fascicularis* are closely related macaque species that respond very differently to infection with the malaria pathogen *Plasmodium knowlesi*.
- Early detection, sensing of the pathogen with associated signaling, and balance between inflammation and cell proliferation are the most important differences in the immune response of the two hosts.
- Pathogen surface antigens of *SICAvar* Type 1 are most highly correlated with host immune and pathogen sensing mechanisms.
- Pre-infection differences in neutrophils and naïve CD4+ T cells result in differences in Ca^{2+} homeostasis, which ultimately balances inflammation and cell proliferation during the expansion log phase of the parasitemia.
- Dysregulation of ribosomal protein assembly in *Macaca fascicularis* causes p53-dependent growth arrest, which is essential for balancing the immune response and inflammation.
- Tryptophan metabolism and its key control gene KMO balance downstream energy metabolism and inflammation pathways through NAD^+ metabolism and AhR signaling, hence playing an important role in the balance of cell proliferation, immune response and inflammation.

INTRODUCTION

Malaria is one of the world's deadliest infectious diseases, with an estimated 229 million cases and 409,000 deaths reported in 2019 (World malaria report 2020, 2020). It is caused by parasites of the genus *Plasmodium*. Ethical reasons render investigations of molecular host-responses in malaria difficult in humans, because treatment of patients is obligatory as soon as they are diagnosed. Rodent malaria models have been widely used to expand our understanding of these infections, but present drawbacks due to major differences in human and mouse or rat genetics and physiology. By contrast, nonhuman primates (NHPs) are much closer to humans, and the clinical presentation of malaria and consequent immune responses are quite similar in humans and macaques (Coatney et al., 1971; Aikawa et al., 1992; Coatney et al., 2003; Gardner and Luciw, 2008; Craig et al., 2012; Joyner et al., 2015; Pasini et al., 2018).

Here, we contrast the drastically different responses of two evolutionarily close macaque species (Morales and Melnick, 1998; Tosi et al., 2000), the kra monkey (*Macaca fascicularis*, Mf) and the rhesus monkey (*Macaca mulatta*, Mm), to infection with the same pathogen, *Plasmodium knowlesi*. These two hosts are the most-studied model NHPs and their infections with various pathogens is studied as it is often comparable to those in humans (Van Binnendijk et al., 1995; El Mubarak et al., 2007a; Baroncelli et al., 2008; Sasseville and Mansfield, 2010; Salguero et al., 2021). They are evolutionarily so close (3.7 MYA) (Hedges et al., 2015) that one might expect a similar immune response to a common pathogen. Long before *P. knowlesi* became a zoonotic concern, Knowles and Gupta (1932) identified Mf as a natural host for *P. knowlesi* infection. Since then, numerous infection experiments have demonstrated that the two macaques respond very differently to infection (Garcia et al., 2004; El Mubarak et al., 2007b; Maiello et al., 2018; Lu et al., 2020). Whereas Mf develops a chronic infection that it tolerates relatively well, the *P. knowlesi* infection in Mm is fatal, unless the monkey is subjected to aggressive treatment. This outcome is somewhat surprising, as it is widely accepted that Mm typically outcompetes other macaque species, including Mf. A likely explanation is that Mf co-evolved with *P. knowlesi* within a large geographical area of Southeast Asia, whereas the distribution of Mm overlaps with that of *P. knowlesi* only slightly (Street et al., 2007; Singh and Daneshvar, 2013; Moyes et al., 2014; Gupta et al., 2021). Studies analyzing these differences have begun to show that Mf generally launches a more effective immune response (Waag et al., 1999; El Mubarak et al., 2007a; Pinski et al., 2021). However, a better understanding of the control of the biological programs that differentiate the immune responses is of utmost importance, because it will not only offer insights into the details of these responses but may also point to molecular targets that might lead to improved malaria treatments for humans.

In a recent transcriptomics study (Gupta et al., 2021), we analyzed the gene programs with which Mm and Mf respond to a *P. knowlesi* infection, initiated with infectious sporozoites. This comparative analysis revealed numerous transcriptomic similarities, but also notable differences. In particular, Mf, but not Mm, apparently detects this pathogen as early as the liver phase of the infection, prior to the parasite infecting the blood, and this correspondingly activates beneficial signaling pathways early on. Later in the infection, significant differences arise in each monkey's immune responses, which in Mm lead to extended inflammatory activities and prolonged inflammation. By contrast,

Mf contains the infection and controls inflammation by undergoing a transcriptional makeover toward cell proliferation that accompanies its recovery.

The goal of the present study is to shed light on some of the molecular mechanisms governing the different gene programs and thus the ultimate fates of the two macaque species. In particular, the study identifies and quantifies: differences in the detection of the pathogen, associated differences in the immune response, differences in cell proliferation that directly affect the immune response and indirectly inflammation and, finally, differences in pathways that regulate inflammation.

The detection of malarial parasites by the host immune system is driven by parasite-encoded surface proteins, including among others the Schizont-Infected Cell Agglutination (SICA) variant proteins (Lapp et al., 2009) that are expressed from the *SICAvar* gene family (al-Khedery et al., 1999; Wahlgren et al., 1999; Pain et al., 2008; Lapp et al., 2018; Galinski et al., 2018). The antigenicity and variability provided by these various proteins stimulates the production of antibody repertoires and immunogenicity that have been widely studied in the context of vaccine development (Ferreira et al., 2004; Ouattara et al., 2015; Rénia and Goh, 2016; França et al., 2017). The *Plasmodium* pathogen multiplies within infected red blood cells (iRBCs) and once matured these cells burst releasing new merozoite progeny that infect other RBCs. This process generates pathogen- and damage-associated molecular patterns (PAMPs and DAMPs), which in turn stimulate various pattern recognition receptor (PRR) signaling pathways in macrophages, monocytes, neutrophils and dendritic cells, and execute various immune mechanisms via protein kinase cascades (Alberts, 2002; Schroder and Tschopp, 2010). Co-expression analysis has been shown to be instrumental in determining these host-pathogen interactions (Lee et al., 2018). The neutrophils and macrophages not only target foreign content for phagocytosis but also trigger the inflammatory and adaptive immune response. Our previous analysis showed much stronger inflammation in Mm compared to Mf, which launches extensive measures to control cell proliferation. The balance between these pro- and anti-inflammatory mechanisms appears to be the key to resilience, and a deeper understanding of the underlying mechanisms is therefore of utmost importance (Cicchese et al., 2018). The energy-intensive nature of these processes makes metabolic processes like glycolysis and tryptophan (Trp) metabolism close accomplices in regulating the overall physiological dynamics. Furthermore, the feedback loop of Trp metabolism and Aryl hydrocarbon Receptor (AhR) signaling in controlling inflammatory cytokines is essential for this balance.

The complete dynamics of the entire immune response is obviously difficult to comprehend in full detail, as this response is systemic and involves uncounted facets, some evident but others subtle. Thus, while our comparative analysis clearly cannot convey a complete picture of all chains of causes and effects governing the responses by the two macaque species, it offers a first glimpse into some of the same and some of the differentiating processes evoked by the two monkey species. The study thereby opens a new avenue toward potential future strategies of immune-based malaria treatments and provides multiple promising candidates for interventions targeting a balanced and effective immune response.

RESULTS

Our analysis is based on data that were obtained with an experimental design (**Supplementary Figure S18**) recently detailed in (Peterson et al., 2021) and (Gupta et al., 2021). In this longitudinal study of *P. knowlesi* infections in Mm and Mf, peripheral blood and bone marrow samples were collected at various time points (TPs), including baseline (before infection), pre-patent (TP3 or 3 days post inoculation; dpi), log-phase (TP4 or eight dpi) and peak-phase (TP5 or 10 dpi). The first signs of parasitemia were observed six dpi, and the infection increased exponentially thereafter. The Mm subjects were euthanized by 10 dpi, at the time parasitemias were escalating to dangerous levels, to carry out necropsies and characterize the infected tissues. We previously observed that Mf shows very early signs of parasite detection by three dpi (Gupta et al., 2021). Even though the immune response was found to be similar between the hosts during the log-phase of the blood infection, Mf was found to switch its response near peak infection towards cell proliferation, which we concluded is a sign of recovery. In the current study we address these and other findings to shed additional light on the molecular mechanisms governing these processes.

Correlated Nonhuman Primate Host and *P. Knowlesi* Transcripts Suggest Common Signaling Mechanisms and the Expression of Key Pathogenic Proteins, Including SICA Antigens

It is to be expected that a mammalian host senses the presence of a parasite based on the detection of pathogenic macromolecules or signals from infected erythrocytes, which trigger signaling pathways in the host that in turn control the gene programs governing a systemic immune response (**Figure 1A**). In this section, we analyze the sensing-signaling process by means of co-expression networks, functional annotation, and logistic regression analysis.

Co-Expression Networks of Host and Parasite Genes

Genes with similar functionality often have correlated expression profiles, which may be identified using co-expression network analysis (Fuller et al., 2007). We adapted this approach by combining both host and pathogen transcripts in a weighted correlation network analysis (WGCNA) (Langfelder and Horvath, 2008) in order to identify modules of host and pathogen genes that act in synchrony. We refer to these modules based on their "hub genes." Specifically, the analysis resulted in three types of modules: (A) Host modules consisting exclusively of host genes; (B) Host majority modules with both host and parasite genes, but with a majority of host genes; and (C) *P. knowlesi* majority modules with both host and parasite genes, but with a majority of *P. knowlesi* genes (**Supplementary Figure S2**).

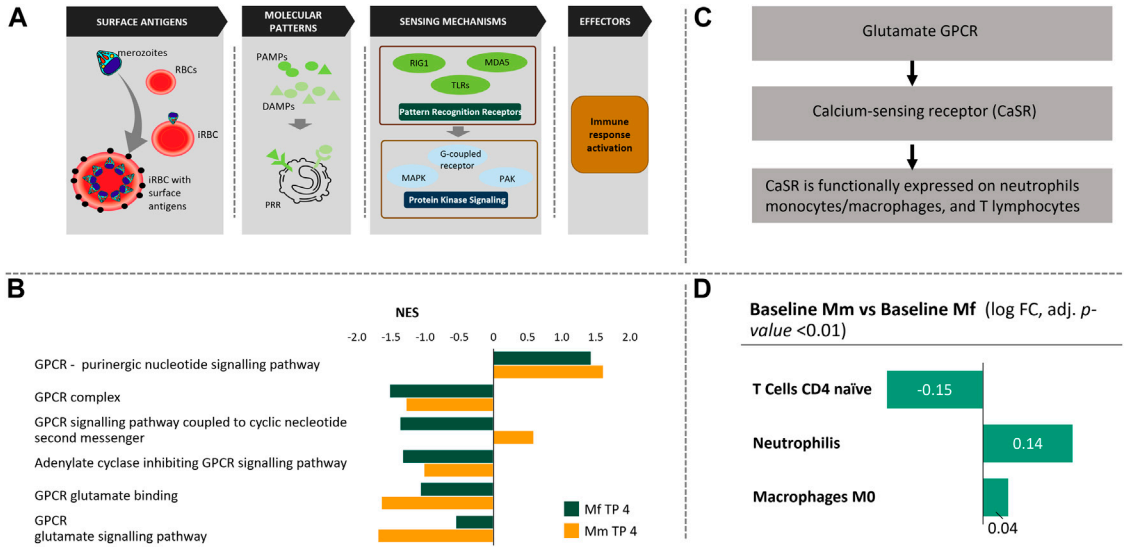


FIGURE 1 | Chain of events during the blood phase of *P. knowlesi* infection. **(A)** Pathogen and RBCs: Once released from the liver into the blood stream, merozoites invade uninfected RBCs leading to Infected RBCs (iRBC) with exposed surface antigens (schizont-infected cell agglutination antigens—SICA; black). The iRBCs are partially eliminated by macrophages, a process that triggers the production of pathogen/danger-associated molecular patterns (PAMPs/DAMPs). These PAMPs/DAMPs are sensed by other immune cells through Pathogen Sensing Mechanisms (PRR signaling), which activate various protein kinase signaling pathways. These signaling pathways are responsible for immune response activation that is mediated through various leukocytes. **(B)** Among several protein kinase signaling mechanisms, GPCR signaling pathways are enriched in both hosts. While purinergic nucleotide GPCRs are similarly enriched, glutamate GPCRs are noticeably different in the two hosts. **(C)** Glutamate GPCRs are responsible for calcium sensing and functionally expressed in neutrophils, monocytes, macrophages and T cells. **(D)** Comparative cell population deconvolution at baseline (i.e., before infection) shows these populations to be different, which might reflect an innate difference in Ca^{2+} signaling in the two host species.

Genes involved with essential functions form well-defined modules (**Supplementary Table S1A**). It is not surprising that most of the differentially expressed genes (DEGs) during the parasitemic log phase belong to the TOP1 module (immune response), followed by C1D and ATAD3A modules (with insignificant functional annotation) for both hosts. It is worth noting that the NF2 (tRNA metabolic process), SNRPD2 (ribosomal assembly) and RACGAP1 (mitotic cell cycle) modules are most highly correlated with the TOP1 module (Pearson correlation between eigengenes with $p < 0.01$, corrected for false discovery rate (FDR)), suggesting close orchestration between these essential functions. The interactions of host and pathogen genes are most evident through interactions between Type C modules and Type A or Type B modules. Significantly high correlations between host and *P. knowlesi* genes (**Supplementary Table S1B**) are found in modules ATAD3A (Type B) and PKNOH_S08507800 (Type C). Interestingly, 23 out of the 26 highly correlated *P. knowlesi* transcripts belong to Schizont-Infected Cell Agglutination variant antigen (*SICAvar*) Type 1 genes (al-Khedery et al., 1999; Pain et al., 2008; Lapp et al., 2018) (**Supplementary Table S1C**). The corresponding SICA variant antigens, which are expressed on the surface of infected erythrocytes (Howard et al., 1983) and associated with virulence (Galinski et al., 2018), show high correlations with several important host genes, including IL10, ELK4 and HSPA6. This suggests that *SICAvar* Type 1 transcripts play a role in regulating inflammation in the host, for example, directly through IL10 expression and indirectly by regulating stress signals through HSPA6 expression.

Parasite Gene Expression Affecting Host Co-Expression Modules

In order to create more functionally robust host modules, we used WGCNA with all host samples (including Baseline and TP3), while excluding *P. knowlesi* genes (**Supplementary Table S2**). Logistic regression followed by functional enrichment identified key modules changing during infection in both host species. The defense response module FBXO6, and modules GFRA2 and RASGEF1A (with insignificant functional annotation), were the most different during the log phase. Modules that were different included RPS19 (SRP-dependent co-translational protein targeting to membrane) and NR1H3 (cell activation involved in immune response). Integrating module membership data with host-pathogen transcripts correlation data (see Section on *co-expression networks*) highlighted *Plasmodium* proteins that affected each module. Most noticeable are *SICAvar* Type 1 (Pain et al., 2008; Lapp et al., 2018), Trp-rich antigen (Wang et al., 2015) and KIR-like proteins (Pain et al., 2008) affecting modules RPS19 and NR1H3, which differentiated the two hosts. Additionally, high correlation of hemoglobin complex module EPB42 with several pathogenic ribosomal proteins suggests a possible mechanism for digesting hemoglobin as an essential nutrition source for the pathogen (Counihan et al., 2021). Host specific WGCNA revealed

similar results with *SICAvar* Type 1 being most correlated to host immune modules (**Supplementary Notes**).

Pattern Recognition Receptor Signaling

The co-expression network analysis is able to detect important host-pathogen relationships that appear to be crucial for the two hosts. In order to focus on detection of the pathogen by a host, we modified the analysis to create a customized module of PRR signaling related genes, which allowed us to identify *Plasmodium* proteins that interact with their products. The most positively correlated pathogen proteins include KIR-like protein (Pain et al., 2008) and thioredoxin-like protein (Wang et al., 2018a; Yindom et al., 2012). Among the genes that negatively co-express with the PRR module are *SICAvar* Type II, AP-1 complex subunit sigma and histones H2A/H2B. The host genes most highly correlated with *P. knowlesi* genes include pathogen detection genes like IFIT3, PLA2G4C, MX1, OASL, DDX60, OAS2, RSAD2, MX2, DHX58, IFIH1, STAT1, FAS and TLR4 (for functional annotations (Huang et al., 2009), see **Supplementary Table S3A**). As reported before (Gupta et al., 2021), it is worth noting that most of these signaling genes are upregulated in Mf at TP3, which further supports the hypothesis of earlier pathogen detection in Mf than in Mm.

For many mammalian hosts, the PRR signaling pathway has been credited for detecting PAMPs or DAMPs, DNAs, and other large molecules (**Figure 1A**) (Mogensen, 2009; Roh and Sohn, 2018; Amarante-Mendes et al., 2018). In our case, the activity of this pathway is consistent with expression of PRR-related genes in both hosts during the appropriate infection time points (TPs; **Supplementary Table S3B**). Here we concentrate on the log phase of infection because this phase is associated with the most similar features between the two hosts, and any observed differences might highlight critical processes. It is interesting to note that many of the PRR signaling genes are in fact different between Mm and Mf. This difference implies that even though the more general PRR signaling pathway is activated in response to the detected pathogen, the specifics of the pathway operation are apparently different, which could be due either to the detected pathogenic content or the interpretation of the signaling event by the host's immune responses. The major genes differentiating the specifics of PRR signaling include TLR5, NLRP6, TNIP3, SLC15A4, SLC15A3, CD36 and CD300A. We had reported enriched pathways before, but the specific differences are not as easy to deduce (**Supplementary Table S4**). Such differences are evident in subsets of the TLR signaling cascade, especially in TICAM1/RIP1 mediated IKK complex recruitment. Higher expression of corresponding ubiquitination genes (UBE2D1, UBA52, RPS27A and BIRC2) in Mm suggests activation of NFκB (Festjens et al., 2007), which might be responsible for stronger inflammation.

These differences between hosts are carried forward toward responses by networks of protein kinases. We examined enrichment of several protein kinase cascades including mitogen-activated protein kinase (MAPK), G protein-coupled receptor (GPCR) systems and p21-activated cascades. Both host species exhibit higher activation of atypical cytokine activated MAPK4/6 signaling involving PAK (p21 activated kinases) (De la Mota-Peynado et al., 2011; Déléris et al., 2011) in comparison to the typical stress activated p38/MAPK signaling pathway (**Supplementary Figure S3A**). Major differences in protein kinase activities are associated with higher inhibition activity in Mm, which is probably due to peptidyl tyrosine dephosphorylation. Although not fully understood, this pathway has been implicated in both pro- and anti-cell proliferation roles (Kostenko et al., 2012) and might be responsible for downstream differences in p53 and HSP27 related cell cycle activity. Differential regulation of protein kinase C activity (**Supplementary Figure S3B**) might explain these differences (Saha et al., 2014).

Probably the most notable difference between the two species is observed in their GPCR signaling (**Figure 1B**). Although both species show similar positive enrichment of purinergic nucleotide GPCR signaling, differences in glutamate GPCR signaling highlight their differences in inflammation (**Figure 1B**). Purinergic nucleotide GPCR activity explains the upregulation of purine metabolism in malaria and points to a potential role of macrophages (Barberá-Cremades et al., 2016). Macrophage production is significantly upregulated in both hosts during log phase (**Supplementary Tables S5A,B**), with the same direction of fold change as the purinergic nucleotide GPCR signaling. In contrast to these similarities, glutamate GPCRs, which are Ca^{2+} sensing receptors, show clear differences between the two hosts in both binding and subsequent signaling pathways, thus suggesting downstream implications of calcium homeostasis (**Figure 1C**). As discussed later in the section *discussing the effects of inflammation*, Ca^{2+} homeostasis plays a crucial role in inflammation. Ca^{2+} sensing glutamate GPCRs are functionally expressed on neutrophils, monocytes, macrophages and T lymphocytes. An innate difference between the two hosts is their difference in these cell populations at baseline (**Figure 1D**, **Supplementary Table S5C**). Some of these differences have been corroborated in the literature (Koo et al., 2019).

Ribosomal Proteins Control p53 Pathway

Our previous work (Gupta et al., 2021) had suggested control over the p53 pathway during the log phase of the infection as a crucial difference between the immune responses of the two host species. Binding of p53 to its target response element leads to the expression of a multitude of genes with a spectrum of functions, including cell cycle growth arrest, DNA repair, cellular senescence and apoptosis (Haupt et al., 2002) (**Figure 2**). An active p53 pathway also protects cells against reactive oxygen species (ROS) through antioxidant genes like TP53INP1 (Sablina et al., 2005) (**Supplementary Figure S4**). Indeed, the early response of this pathway in Mf (at TP4) might be crucial in saving cells from apoptosis via PIG3 (TP53I3 gene) (Lee et al., 2010). Co-expression analysis, discussed in the earlier section on *parasite gene expression*, revealed several closely regulated modules controlling DNA binding, the mitochondrial envelope, and the mitotic cell cycle, all of which are more strongly enriched in Mf (**Supplementary Table S2**).

As observed in the previous section, certain MAPK signaling mechanisms potentially regulate the p53 pathway. Generically,

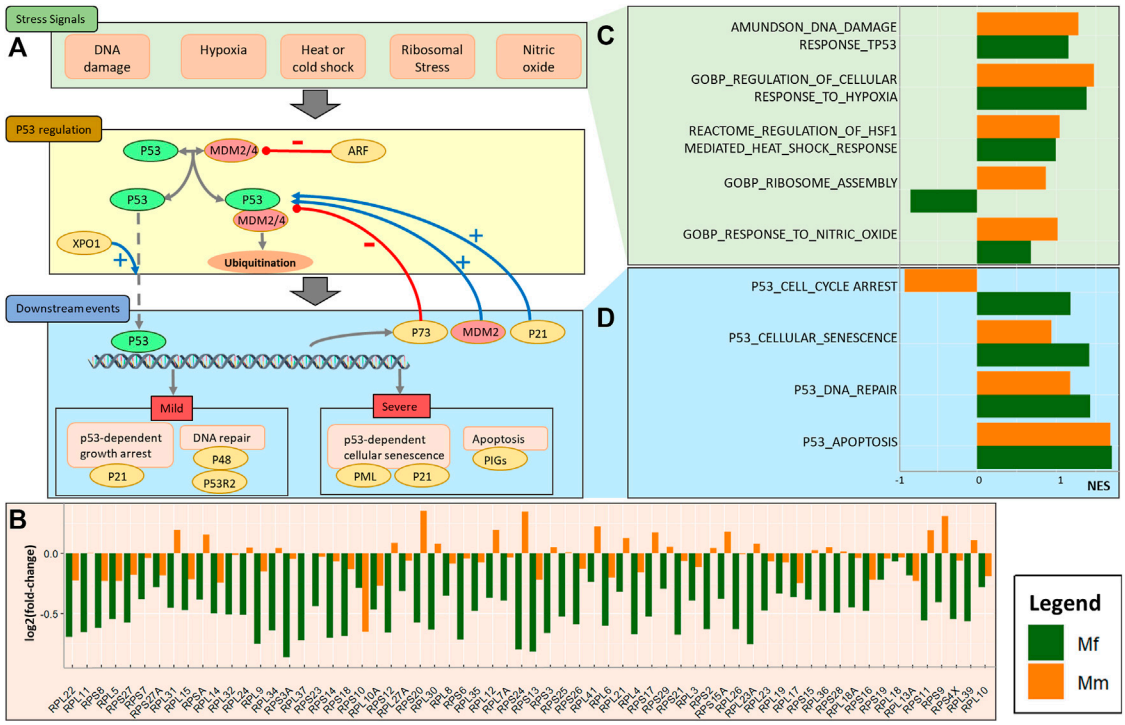


FIGURE 2 | Involvement of the p53 pathway. **(A)** Schematic of cause, regulation, and effect of p53 pathway activation during the log phase of a malarial infection. The overall process includes key stress signals that engage the p53 regulation pathway and result in regulation of downstream events. **(B)** Barplot for log2 fold-changes in the expression of ribosomal proteins between TP4 and baseline, comparing the two hosts. **(C)** Barplot for normalized enrichment scores (NESs) of stress signals involved in p53 activation. **(D)** Barplot for normalized enrichment scores of p53 related downstream events.

the p53 pathway is operational in the presence of molecular stresses and depends on their severity as well as other factors. Cellular stress signals that activate a p53 response include hypoxia, DNA damage, ribosomal and oxidative stresses, among others (Joerger and Fersht, 2016). These stresses are, of course, not independent of each other and manifest in an interrelated manner. In the case of a *P. knowlesi* infection, this interdependence can be seen in the enrichment of associated genes. In particular, the enrichment analysis demonstrates that ribosomal stress is a differentiating factor between the two host species, with stress caused by substantial downregulation of the ribosomal assembly complex in Mf (**Figure 2C**). This downregulation is achieved through the activity of various RNA polymerases (**Supplementary Figure S5**), and the significant downregulation of PLOR1C, POLR2E, POLR2A, POLR2J and POLR2L in Mf at TP4 suggests that these genes might be crucial for the control of p53.

The downregulation of associated ribosomal proteins (RPs) in Mf at TP4 (**Figure 2B**, **Supplementary Figure S6**) is indicative of alterations in ribosomal biosynthesis that results in unassembled RPs and 5S rRNA, which binds to the p53 inhibitors MDM2 and MDM4 and thereby prevents p53 degradation (Golomb et al., 2014; Haupt et al., 2019). As a consequence, p53 facilitates translation from its mRNA internal ribosome entry site (IRES). Indeed, the co-expression network analysis reveals high correlation of MDM4 with *SICAvar* Type 1 transcripts, which suggests direct control that might be crucial in this regulation.

As a consequence of RP downregulation, the p53 pathway in Mf is upregulated, which is reflected in higher levels of enrichment. However, the less pronounced changes in cell cycle arrest and DNA repair appear to be the strongest differentiating factors between the two species. The important genes involved in these processes include CDKN1A (p21), E2F7, PML and MDM2 (upregulated) and TP73 (downregulated).

Another facet of p53 pathway activation and control is provided by the transcription factor HSF1 (heat shock factor 1). Dysregulation of ribosomal biosynthesis processes leads to proteotoxic stress and a balance between these processes must be maintained (Albert et al., 2019). In comparison to Mf, Mm has higher ribosomal biosynthesis and senses higher proteotoxic stress during the log phase of infection, and these processes are further increased near the peak (**Supplementary Figure S7**). Of note in this context is the differential expression of chaperone-mediated protein folding genes—HSPA1A, HSPA8, DNAJB1 and FKBP4. This expression results in upregulation of HSF1 target genes in Mm. p53 has been shown to form multi-chaperone complexes with HSPA1, DNAJB1 or HSPA8, while FKBP4 is essential for its transport to the nucleus (Toma-Jonik et al., 2019). Among the apoptotic targets, ATF3 enrichment at both TP4 and TP5 in Mm highlights differences with Mf.

Finally, the control over cell proliferation in Mf is eased near the peak, which leads to upregulation of adaptive immune cells and, in particular CD4 memory activated and follicular helper cells (**Supplementary Table S7F**), both of which enhance the adaptive immune response by supporting B cells and CD8 T cells (MacLeod et al., 2009; Crotty, 2014).

Effects of Inflammation on Immune Response and Cell Proliferation

Control of cell proliferation in Mf during the log phase of infection constitutes a stark contrast to the elevated inflammatory response in Mm. This difference can be seen clearly in the enrichment of several inflammatory pathways, elevated expression of inflammatory genes, inflammasomes and inflammatory biomarkers like the kynurenine (Kyn)/Trp ratio (see **Supplementary Figures S8–S10** and next section). Co-expression analysis revealed that most of the inflammatory genes are part of the innate immune module (FBX06 module). Not surprisingly, this module is most significantly changed (according to logistic regression) in both hosts during log phase. Although the fold-change for this module is similar between the hosts, the lower adjusted p value (q) and higher log-odds (B) suggest a stronger role of this module in the innate immune response in Mf ($q < 3e\text{-}17/B > 32$) in comparison to Mm ($q < 1e\text{-}10/B > 17$). This module further highlights the differences between the hosts near peak infection as Mm ($q < 7e\text{-}7/B > 8$) maintains its immune response while Mf ($q < 4e\text{-}5/B > 1$) does not. The neutrophil activation and intracellular vesicle transport (FYB) module is most highly correlated to the innate immune module (FBX06). Also, worth noting is that both these modules are negatively correlated to ribosomal biosynthesis and localization modules RPS19, ZNF395, CHD6 and RASGEF1A. This finding implies an important, sustained balance between immune related inflammation and control over the cell cycle.

Several similarities in inflammation gene sets are found between the hosts, especially with respect to an LPS-like inflammatory response, which probably is a symptom of the NLRP3 inflammasome (**Supplementary Figure S8A**). This phenomenon might be attributed to significant upregulation of monocytes and monocyte-derived pro-inflammatory M1 macrophages (**Supplementary Tables S7A,B**), which are the first-line cells expressing inflammasome genes (Awad et al., 2017) (**Supplementary Figure S10**). Important differences are detected in the inflammatory response cytokine production and an antigenic stimulus (**Supplementary Figure S8B**). Even though the positive regulation of these functions is similarly enriched in the two hosts, as seen in the important genes NOD2, GPX1 and IL12B, the negative regulation shows a distinct and opposing enrichment. The main distinguishing genes include IL10, NLRP6 and ABCD1.

The two hosts show similar enrichment of the chronic inflammatory response; however, Mm has a higher acute inflammatory response (**Supplementary Figure S8B**), reaffirming the stronger inflammation in Mm during log phase. The chronic inflammation changes near the peak and is mostly driven by crucial genes like IL10, IDO1, TNF, TNFAIP3 and CXCL13 (**Supplementary Figure S9**).

Exploration of innate immune components of inflammation reveals a crucial difference in S100 proteins (**Figure 3**). Ca^{2+} sensing S100 proteins have a wide range of functionality that includes cell apoptosis, proliferation and inflammation (**Figure 3B**) (Fox and Man, 2019). Differential upregulation of S100A8, S100A9, S100A16 and S100P in Mm suggests a potential role of Ca^{2+} in inflammation (Wang et al., 2018b), while upregulation of S100A4, S100A2 and S100A3 in Mf suggests possible regulation of p53 (**Figure 3D**) (Pan et al., 2018; Boye and Mælandsmo, 2010). Since neutrophils release S100A8/A9 during inflammation, their differential expression mediates Ca^{2+} signaling, which positively regulates NLRP3 inflammasome assembly and pro-inflammatory activity of NFκB (**Figures 3B,C**) (Wang et al., 2018c; Xia et al., 2018). This inflammatory activity is further exacerbated by master regulator DDX3X (Fox and Man, 2019). Further enrichment of processes specifically associated with innate and adaptive immune processes reveals an interesting pattern that succinctly differentiates the responses of the two hosts. Namely, TLR4 signaling is stimulated by Ca^{2+} via S100 proteins (S100A8 and S100A9), which enhances the inflammatory activity of the NLRP3 inflammasome. These inflammatory pathways are responsible for IFNβ regulation and IL6 production. This finding directly complements earlier findings of differential Ca^{2+} glutamate GPCR activity (Section *PRR signaling*), which directly affects the inflammasome assembly.

Further exploration of differences in immunological signatures reveals several important similarities and differences between the two hosts (**Supplementary Figure S11**). Both hosts show significant enrichment towards FOXP3+ $CD4^+$ naïve T-reg cells (GSE37533 (Cipolletta et al., 2012), GSE42021 (Toker et al., 2013)), with gene sets pointing to the strongest enrichment of a thymic T-reg subset of intermediate maturation, $CD24^{int}$. A related important difference appears between the two hosts: Mf has a higher mature ($CD24^{low}$) subset while Mm has a higher immature ($CD24^{hi}$) subset. Both hosts have enriched naïve B cells (GSE42724 (Covens et al., 2013)), even though this change could not be confirmed by deconvolution analysis. Differences can be seen in gene sets derived from IL6 and IL10 stimulation as well. Taken together, these signatures suggest downregulation of key genes in Mf, which is not observed in Mm. Important genes that seem to be regulating this process in both Mm and Mf include IL6, IL6R, TGFB3, IL23A, IL10 and SOCS3.

Although not conclusive, pre-infection state differences in cell populations point to eventual differences in the immune response. For instance, at baseline, Mm has significantly more naïve $CD4^+$ T cells while Mf has higher levels of neutrophils (Koo et al., 2019) (**Figure 1D**, **Supplementary Table S5C**). Although there is no significant relative difference in cell populations during the log phase of infection between the hosts (**Supplementary Table S5E**), these initial pre-infection differences persist (**Supplementary Table S5D**) and may be a key differentiating factor in the immune response.

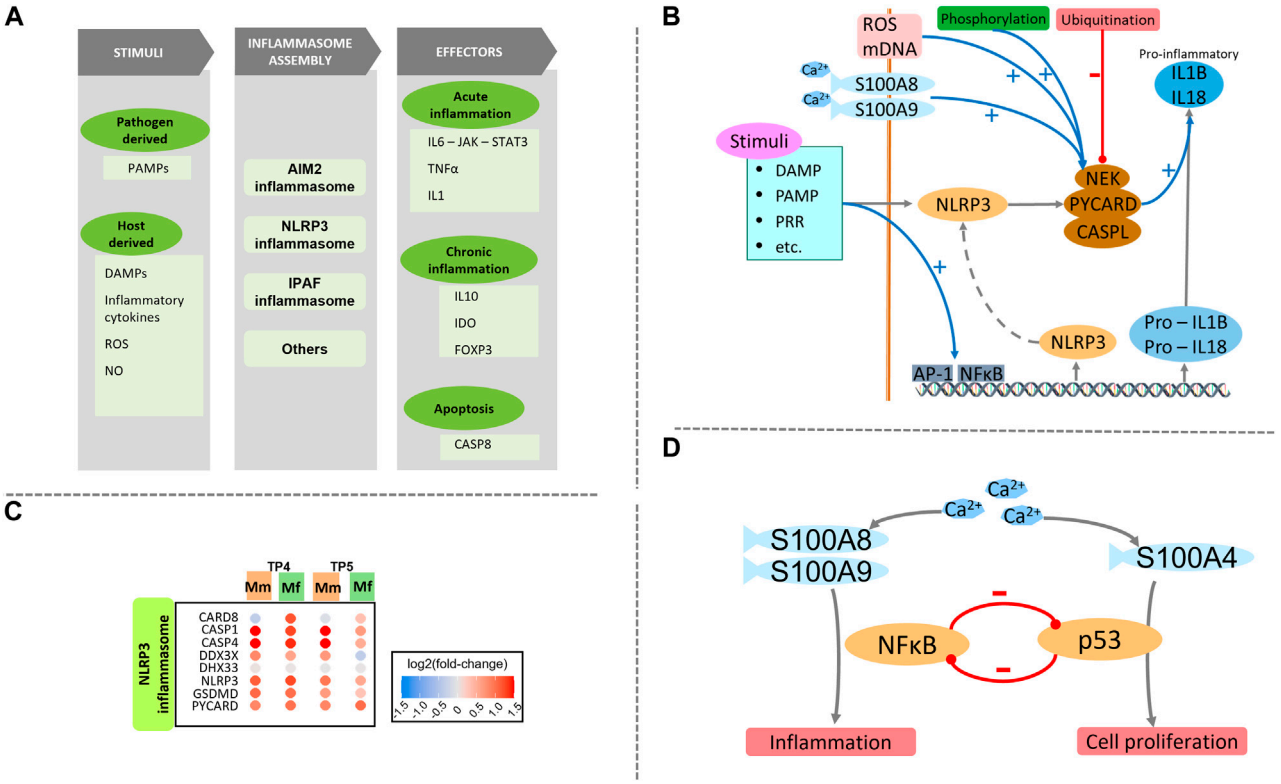


FIGURE 3 | (A) Schematic of the inflammasome assembly process. Various host and pathogen derived stimuli are responsible for initiation of the inflammasome assembly process. This process is very closely regulated by various signals and processes including ROS and Ca^{2+}. Several effectors execute important processes like inflammation and apoptosis. Different effectors are activated during acute and chronic phases. **(B)** Detailed schematic of NLRP3 inflammasome signaling including the initiation and regulation of the assembly process, followed by pro-inflammatory effectors. **(C)** Heatmap of genes involved with NLRP3 inflammasome assembly process are similarly enriched between the two hosts. **(D)** Balance and cross-regulation of p53 and NFκB showing importance of Ca^{2+} homeostasis.

Changes in Tryptophan Metabolism Suggests Higher NAD$^+$ Metabolism

Metabolomic and transcriptomic analysis of the Mf and Mm hosts revealed prominent differences in the expression of genes associated with Trp metabolism at TP4 and TP5 (**Figure 4**). Trp metabolism can coarsely be divided into pathways responsible for serotonin and melatonin, Nicotinamide Adenine Dinucleotide (NAD^+) and Kyn synthesis. Serotonin and related compounds are not of interest in the present context, and their concentrations in blood are very low. NAD^+ metabolism plays a crucial role in cellular energy regulation as well as the handling of ROS. The Kyn pathway is responsible for the biosynthesis of several metabolites that play key roles in immunomodulation and inflammation.

Both hosts had lower Trp and higher Kyn levels in the peripheral blood during the log and peak phases of infection in comparison with baseline levels (TP1 and TP2), as observed previously with *P. coatneyi* infection of rhesus macaques (Cordy et al., 2019; Colvin and Joice Cordy, 2020). Of special note here, however, is that the Kyn/Trp ratio, a known inflammatory biomarker, is reduced in Mf near peak infection, whereas it remains at the same level as during the log phase in Mm (**Figure 4**).

IFNγ signaling is responsible for upregulating the expression of IDO (Taylor and Feng, 1991; Sarkar et al., 2007; Banzola et al., 2018), which converts Trp into Kyn (**Figure 4A**). Even though IFNγ signaling is upregulated in both hosts, a higher degree of signaling in Mf near the log and peak phases results in higher IDO expression, which thereby leads to a higher conversion of Trp to Kyn in Mf than in Mm (**Figure 4C**). One might expect that this increased activity should lead to a higher level of Kyn. Yet, we observe lower levels of Kyn and a lower Kyn/Trp ratio in Mf (**Figure 4A**), which however is easily explained by the increased activity of the subsequent enzymes KYNU and KMO in the Kyn utilization pathway, which ultimately lead to higher NAD^+ biosynthesis and immunomodulatory activities. In addition, earlier downregulation of AhR and AADAT in Mf suggests potential differences in AhR signaling (see next section).

In order to elucidate the role of Trp and Kyn levels in the blood and then understand changes in Trp-Kyn metabolism in white blood cells (WBCs) during the infection, we adopted existing metabolic models of Trp metabolism in brain and liver (Stavrum et al., 2013) and adapted them to reflect Trp metabolism in WBCs. With this adapted model we can clearly differentiate Trp-Kyn metabolism in blood from brain and liver (**Supplementary Figure S12**). The model confirms that over 90% of Trp in whole blood is channeled toward Kyn through the activity of

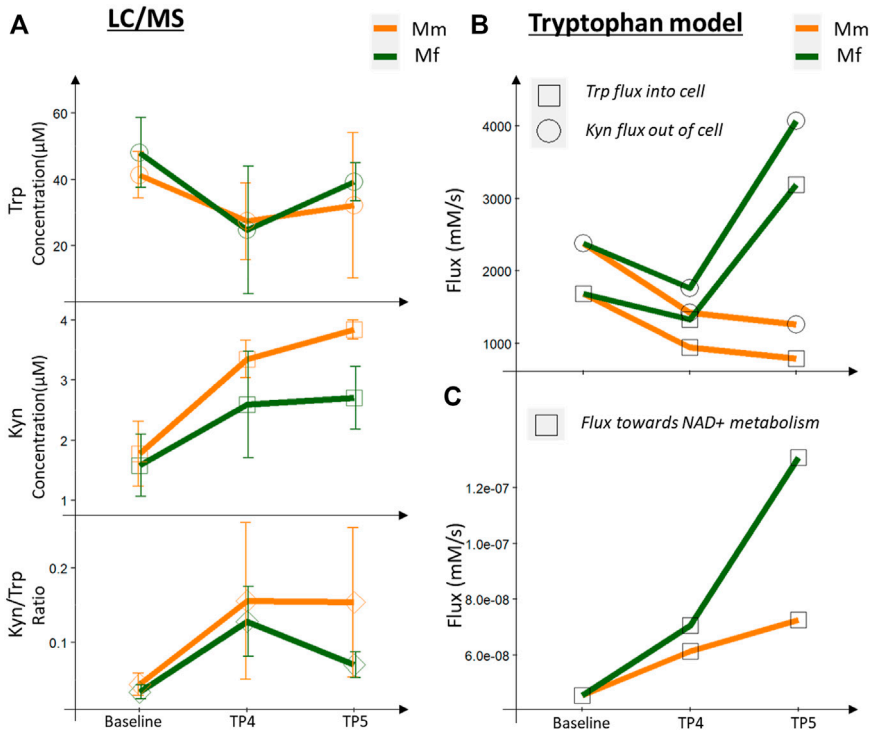


FIGURE 4 | Metabolomics (LC-MS measurements) and model predictions for tryptophan metabolism. **(A)** LC/MS measurements for Trp, Kyn and the Kyn/Trp ratio across infection timepoints, comparing Mm and Mf. **(B)** Fluxes predicted by Trp model for Trp consumed and Kyn produced by blood cells. **(C)** Flux towards NAD metabolism predicted by the Trp model.

IDO, compared to tryptophan-2,3-dioxygenase (TDO) in the liver model. Kyn itself is the substrate for different reactions, and flux control analysis (Fell, 1992; Fell, 2007) reveals that KMO is the most important control point for Kyn utilization. Extension of the blood model during the log phase of the infection shows for both host species that the transport of both Trp and Kyn through the cell membrane is lowered in comparison to the baseline (**Figure 4**). This finding is interesting as it potentially leads to serum levels of Trp and Kyn in both hosts that change significantly during this phase and lead to a higher Kyn/Trp ratio (**Figure 4**). At the same time, these increases are accompanied by a major reshuffling of fluxes, which affects the metabolite concentrations inside the cells.

Specifically, Trp can be metabolized through six reactions, among which the pathway toward Kyn is most important, based on relative fluxes (**Figure 4A**). Indeed, if the Trp concentration inside the cells is decreased, the effluxes out of the Trp pool are also decreased, with the notable exception of the Kyn pathway, which receives essentially a normal influx from Trp. This flux is important, because the pathway later leads to the formation of quinolinic acid, which is a precursor of NAD^+ and thus affects energy metabolism and redox handling. At this juncture, the differential expression of KMO becomes even more important as a control point for Trp metabolism: here, it causes a higher flux in Mf toward NAD^+ synthesis (**Figure 4C**). This enhanced flux from Trp to Kyn is compensated in both species by decreased fluxes from Trp toward protein synthesis and serotonin production (**Supplementary Figure S13**).

Near the peak of infection, the differences in the two hosts are particularly pronounced, with Trp and Kyn transported through the cell membrane at higher rates in Mf than in Mm. As a result, the fluxes through the Kyn pathway are higher in Mf (**Supplementary Figure S14**). At the same time, the Kyn/Trp ratio is lower in Mf during this phase of the infection, presumably due to the enhanced activity of KMO (**Figure 4A**). It is also worth noting that the higher flux toward NAD^+ metabolism persists in Mf (**Figure 4C**). Furthermore, the concentrations of other Kyn compounds remain high, and these are potential ligands of the AhR (see next section), which ultimately serves as a transcription factor for numerous genes (**Figure 4B**).

Among the other effluxes out of Trp, the indole-pyruvate and tryptamine pathways are also responsible for AhR activation (Hubbard et al., 2015; Roager and Licht, 2018) (next section). Trp is incorporated into proteins via tryptophanyl-tRNA synthestases (WARS proteins), a process that directly links Trp sensing to p53 activation (Yu et al., 2021). Changes in these fluxes during infection further show the central role of Trp metabolism (**Supplementary Figures S16A,B**).

Aryl Hydrocarbon Receptor Signaling and the Role of the Aryl Hydrocarbon Receptor Repressor in Controlling Aryl Hydrocarbon Receptor and HIF1A Signaling

AhR belongs to the basic helix–loop–helix-PER-ARNT-SIM (bHLH-PAS) superfamily of transcription factors where

multiple other members interact with each other and therefore affect each other's functionality. Prominent members include AhRR, ARNT and HIF1α (**Supplementary Figure S15**).

Most of the biologically active intermediates of the Kyn pathway, as well as several other compounds, can act as ligands for AhR (Murray et al., 2014; Gutiérrez-Vázquez and Quintana, 2018) (**Figure 4B**, **Supplementary Table S6**), which makes this receptor a central control point for multiple physiological changes, *e.g.*, in heme degradation, hypoxia and Trp metabolism. Once a ligand binds, AhR can form a complex with the nuclear transporter ARNT, which is translocated to the nucleus. Once in the nucleus, the AhR-ARNT complex binds to the ARE promoter region of numerous genes.

Different Kyn derivatives may act as ligands for AhR, and their dynamics differentiates the Mf and Mm hosts during infection. Additionally, IL4I1 activity leading to indole-pyruvate derivatives from Trp synthesis activates AhR (Zhang et al., 2020). Similarly, multiple other ligands have been associated with AhR activity, and some of these may constitute further differences between the two host responses (**Supplementary Figures S16C–F**). For instance, *Plasmodium*'s consumption of hemoglobin releases heme, which is metabolized (**Supplementary Figure S17**). Certain AhR ligands that are derived from heme metabolism (Kapitulnik and Gonzalez, 1993; Phelan et al., 1998) may point to additional differences between the two hosts.

Another level of control of AhR signaling occurs through the competition between AhR, AhRR and HIF1α for ARNT and thus for transport into the nucleus and binding to their corresponding response elements. Given the molecular similarity of the competitors, it is not surprising that most of the downstream genes are simultaneous targets of both the AhR-ARNT and the HIF1α-ARNT complexes (**Supplementary Table S7A**).

Exploring these targets through an enrichment analysis shows that both complexes act quite similarly during log phase (**Supplementary Figure S18B**, TP4). Yet, several differences emerge near the peak of infection (**Supplementary Figure S18B, TP5**). The most pronounced differences emerge with respect to higher upregulation of AhR targets and HIF1α targets in Mf, while targets of the AhR repressor AhRR are downregulated in Mf. It appears that the relative hypoxia stress is quite different between the two hosts (**Supplementary Figure S19**), but it is unclear how the balance is achieved between these complexes and their corresponding genes.

To shed light on the interference among these complexes, many of which share numerous common targets, we calculated enrichment of each subset of these targets (**Supplementary Figure S18**). Specifically, we divided the targets into three major groups (AhR targets, HIF1α targets, and AhR and HIF1α targets) and compared them with and without the AhRR binding site to account for repressor activity (**Supplementary Figure S18A**).

The effect of AhRR on AhR targets is quite clear in Mf, with lower enrichment of targets at both TP4 and TP5, as opposed to almost no effect in Mm. Corresponding effects of AhRR on HIF1α targets are not easily identified. At TP4, HIF1α targets with the AhRR binding site are more enriched than without AhRR. At TP5, AhRR containing HIF1α targets are enriched more in Mm and less in Mf.

As there are multiple levels of regulation, it is difficult to predict the activity of these targets without further experimentation. However, one may try to elucidate the specific functionality of these targets by identifying the key genes along with their functional annotation. The transcription factor complexes in question are associated with a wide range of genes with diverse functionality (**Figure 4**). Functional annotation of AhR and HIF1α targets shows their involvement in key process like the p53 pathway, heme metabolism, cell cycle related pathways, and immune related IFNγ and NFκB pathways (**Supplementary Table S7B**). The complex nature of this response makes it difficult to elucidate the specifics and differences during a *P. knowlesi* infection, but the activity of individual genes suggests potential outcomes. Their roles in immune and inflammatory processes are evident in the activity of genes like OASL, STAT3, IRF5, IL6, DDIT4, NRF2, REL, and LAG3. These IFNγ signaling genes create a positive feedback loop, because IFNγ directly regulates IDO expression, which leads to enhanced levels of the AhR ligand Kyn. The control over cell proliferation is evident in the operation of p53 and other cell cycle related genes like MXD1, FOS, BCL6, GADD45A, and CREBRF. Another possible contributor with respect to malarial infection is heme metabolism with target genes including CCND3, BLVRB, and KLF1.

DISCUSSION AND CONCLUSION

Malaria has haunted mankind throughout its history. Even after several decades of active research, malaria continues to be a severe global health concern with over 400,000 fatalities and about 3.2 billion people at risk annually. Among the six species of *Plasmodium* known to cause malaria in humans, *P. knowlesi* has become recognized as a major zoonosis in Southeast Asia (Cox-Singh, 2012; Barber et al., 2017; Zaw and Lin, 2019; Raja et al., 2020). A *P. knowlesi* infection in humans may range from mild to severe, with 6–10% of the cases considered severe (Singh et al., 2004; Daneshvar et al., 2009). A deeper knowledge of the details of *P. knowlesi* infections can be expected to provide a crucial basis for understanding the immune responses in general and for comparing resilient and severe malarial responses in particular. As a zoonotic species, *P. knowlesi* has the advantage that it can be studied in different NHP species (Pasini et al., 2018; Peterson et al., 2021). Among these NHP models, *Macaca mulatta* (Mm) and *M. fascicularis* (Mf) provide unique advantages specifically for comparing *P. knowlesi* infections with different disease progression. Namely, even though Mm and Mf are evolutionarily very close, Mm, once infected, suffers from increasing parasitemia, which is in almost all cases fatal if not treated, whereas Mf controls parasitemia and escapes death without treatment (Knowles and Gupta, 1932; Napier and Campbell, 1932; Peterson et al., 2021). These dramatic differences provide unparalleled opportunities to study the details of host physiology and immune responses in the context of host-parasite interactions and explore mechanisms of resilience in human malaria, and to potentially relate the findings to other diseases that may also show drastically different possible outcomes.

In previous work, we had established crucial differences in the transcriptomics of the two hosts that ultimately determines the outcome in terms of susceptibility and resilience (Gupta et al., 2021). As also noted in the clinical assessment by Peterson *et al.* (Peterson et al., 2021), transcriptomics analysis showed that Mf detects the pathogen earlier than Mm, and even though both host species mount a similar immune response, Mf starts controlling inflammation as early as the log phase of infection (Gupta et al., 2021). Subsequently, Mf switches the immune response towards cell proliferation pathways, which presumably aids recovery (Gupta et al., 2021). The current analysis explores the key findings further and explains the molecular functions that determine the mild or fateful outcome. Interestingly as well, early detection of the parasites by the Mf animals is also consistent with a rise in temperature in this species immediately upon patency, by seven dpi (Peterson et al., 2021).

The current results show consequential differences in signaling mechanisms beginning with the early detection of the presence of *P. knowlesi* pathogens by Mf. Once the merozoites invade the RBCs, they transform the iRBC and express different antigenic forms of surface molecules in an attempt to escape the immune response (Brown and Brown, 1965; Howard et al., 1983; Biggs et al., 1991). Specifically, antigenic variation of *P. knowlesi* SICA proteins is a main factor responsible for chronicity in Mm (Brown and Brown, 1965) (reviewed in Galinski et al. (2018)). Moreover, expression of *SICAvar* genes in *P. coatneyi* have been shown to change as chronic rhesus monkey infections are established, also suggesting a role for metabolites in regulating these changes (Cordy et al., 2019). Our correlation analysis of host and pathogen transcripts sheds light on possibly involved *SICAvar* Type 1 genes (al-Khedery et al., 1999; Pain et al., 2008; Lapp et al., 2018) along with correlated host genes. The specific correlations of individual transcripts from this large pathogen gene family—with 136 *SICAvar* members (Lapp et al., 2018)—could shed light on its transcripts and their variable gene expression, which may trigger different antibody responses. Additionally, correlations with host genes, especially the differentially responding IL10 and HSPA6 genes, can help associate parasite markers with the host immune response.

On the host side, differences in the mechanisms for pathogen detection and PRR signaling pathways are surprisingly subtle. However, these differences are magnified downstream with MAPK signaling. There is a close relation of these signaling cascades, especially the GPCR activity with the p53 pathway and cell cycle (Zhang and Liu, 2002; Goldsmith and Dhanasekaran, 2007; New and Wong, 2007). Ca^{2+} drives intracellular communication and interacts with GPCR to regulate various aspects of the cell cycle, and by extension, regulates inflammation and apoptosis during infection. This regulation is even further augmented by inflammasome activity (**Figure 5**). Specifically, some of the Ca^{2+} binding S100 proteins (S100A8, A9 and A4) might be differentiating factors between the two hosts. While S100A8 and S100A9 aid the inflammasome assembly, S100A4 assists with the regulation of the p53 pathway. Additionally, the inflammasome assembly process is regulated by multiple other factors including ROS, IL10 and transcription factor AP-1. These factors do not only relay the stress response but also seem to be important in regulating the p53 pathway.

The most strongly differentiating factor between the two species appears to be the control of cell proliferation by Mf during log phase via the p53 pathway, along with subsequent inhibition that leads to recovery. Similar stresses can trigger both inflammation and cell proliferation, but it appears that it is the stress related to fundamental ribosomal assembly that causes the inhibition of cell proliferation in Mf through the p53 pathway. Several upstream kinases have been shown to cause this stress. Since ribosomal assembly is one of the most energy intensive functions, inhibition of this fundamental function to conserve energy seems likely (Albert et al., 2019). Of course, that is not the sole purpose. In particular, we observe that ribosomal assembly leads to differences in p53 based cell cycle arrest and DNA repair in Mf. The interrelatedness of this pathway with p21, AP-1 and HSF1 activity provides additional regulators that might be responsible for balancing cell proliferation with inflammation.

Another known inflammation biomarker, the Kyn/Trp ratio, shows surprisingly deep integration with these processes. Even though the induction of IDO in malarial infection is quite often discussed (Sanni et al., 1998; Hansen et al., 2000; Tetsutani et al., 2007; Colvin and Joice Cordy, 2020; Santos et al., 2020), its biological significance for the immune response is in general poorly understood. Nonetheless, a mathematical model of the direct upregulation of IDO through IFNγ signaling quite clearly shows how the Kyn/Trp ratio changes during the infection (**Figure 5**). This metabolic model is able to shed light on several important, although indirect implications, such as the importance of KMO and KYNU in regulating fluxes, redirection of fluxes towards NAD^+ metabolism, and metabolite pools of kynurenine compounds as ligands for AhR. In summary, Trp metabolism diverts the fluxes towards the essential functions, and especially NAD^+ metabolism and protein synthesis. The higher activity in Mf also indicates that this host maintains essential functions in spite of the inflammation. Further analysis into the kynurenines shows an impactful control of AhRR in regulating both AhR and HIF1α related signaling. This process includes a competitive effect of multiple stresses, hypoxia and infection induced damage and cytokine response in determining the overall outcome.

Although this analysis dives deep into multiple molecular mechanisms that play crucial roles in permitting resilience of the host, it only paints a crude image of the immune response over time. For example, a more detailed longitudinal and immunologically based analysis of *SICAvar* gene expression and switching of SICA proteins in each host (and with different parasite species (Cordy et al., 2019)) is likely to advance our understanding of the different antibody responses and immune evasion mechanisms (reviewed in Galinski et al. (2018)). The combined analysis of immune response, inflammation and cell proliferation also seems to reveal Ca^{2+} as a crucial factor, which is known to play a role in iRBC egress (Glushakova et al., 2013). If this general finding can be validated and cross referenced with other bacterial and viral infections (Tran Van Nhieu et al., 2018; Chen et al., 2019; Crespi and Alcock, 2021), improved understanding of Ca^{2+} homeostasis might lead to novel targets that could naturally aid the immune response against *Plasmodium* infection. Similarly, the metabolic model we employed, adjusted for transcriptional changes during the infection, provides a deeper appreciation of the mechanisms of

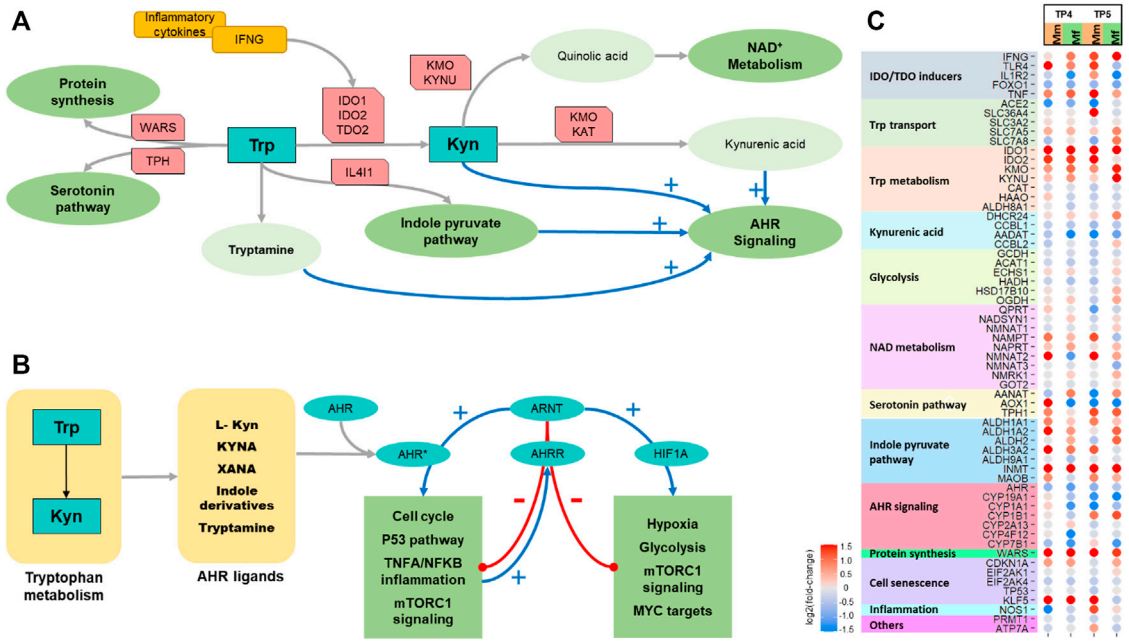


FIGURE 5 | Tryptophan metabolism. **(A)** Schematic showing key features of tryptophan metabolism. **(B)** Schematic showing AHR signaling. **(C)** Heatmap of differential expression of significant genes involved in tryptophan metabolism comparing the two hosts across TP4 and TP5.

Trp metabolism and could possibly be extended to identify targets that could predictably adjust metabolism to aid in resilience.

Overall, this work interprets transcriptional data and integrates them in a manner that provides deeper understanding of *Plasmodium* infections. It is hoped to suggest new avenues of studying malaria and identifying valid candidates for future drug development.

METHODS

Experimental Setup and Data Pre-Processing (Ribonucleic Acid Seq/ LC-MS)

The analysis described here expands on previously published studies (Gupta et al., 2021) with details about individual processes. Briefly, four male Mm and seven male Mf were infected with *P. knowlesi* sporozoites. Peripheral blood samples were extracted before (baseline) and after inoculation with sporozoites (pre-patent—TP3, log-phase—TP4 and peak-phase—TP5). These blood samples were used for transcriptomics and metabolomics analysis (**Supplementary Figure S1**).

To assess the transcriptome, samples were sequenced using Illumina Hi-seq 3,000, mapped using STAR and normalized using DESeq2. Details of the process were previously published in Gupta et al. (2021).

For metabolomics analyses, plasma samples were quantified using the AbsoluteIDQ p180 kit (Biocrates Life Sciences AG). Specifically, the metabolites were quantified using SCIEX Exion LC and a QTRAP 5500 mass spectrometer in only positive ionization mode with each sample injected using a separation column. Specific details of the process can be found with the corresponding submissions of MaHPIC data to PlasmoDB (https://plasmodb.org/plasmo/app/static-content/PlasmoDB/mahpic.html) with the Mm dataset available at MTBLS824 and the Mf dataset at MTBLS822 from the MetaboLights repository.

Enrichment Analyses

Differential expression (DE) of genes was calculated using DESeq2. Genes with low read counts were removed from analysis. The genes were modeled using the design—*Species* + *TimePoint* + *Species:TimePoint* and DE was calculated using Wald's test.

Gene set enrichment analysis was performed using the GSEA toolkit (version 4.0) of the Broad Institute. The gene sets used for the analysis were Hallmark (Liberzon et al., 2015), Reactome (Jassal et al., 2020), ImmuneSigDB (Godec et al., 2016) and Gene Ontology (Ashburner et al., 2000; Gene, 2021). The pre-ranked GSEA module of the toolkit (Subramanian et al., 2005) was used for the analysis, and all genes were ranked based on inverse of adjusted *p*-values and the sign of fold changes. Files of custom gene sets (gmt files) were created using R to contrast enrichment scores between comparable data sets. To compare gene sets across the two species and account for representation bias in individual gene sets, rank scores for all genes were used to calculate enrichment scores (ES), which were adjusted by normalization of gene set sizes. Gene sets with small (<15) and large (>500) overlaps were filtered out. This normalized enrichment score (NES) was used to contrast various gene sets.

Enrichment analysis for targets of AhR, AHRR and HIF1A was performed similarly to the method described above. The gene

sets for target genes for each were created using ChIP-Atlas (Oki et al., 2018) with ±5 Kb overlap with the transcription start site. NES values for each subset described in the *Results* were calculated with the method described above.

Weighted Gene Co-Expression Network Analysis

Weighted gene co-expression network analysis (WGCNA) was performed using the WGCNA package (version 1.70–3) (Zhang and Horvath, 2005; Langfelder and Horvath, 2008) in R to describe correlation patterns among genes. The analysis was performed in multiple ways to serve different purposes. The differences arose in the subsets of samples in datasets used for each analysis. First, for co-expression networks with both host and pathogen genes, only infection TPs (TP4 and TP5) were used for both hosts, as there are no pathogen transcripts at baseline and TP3. Next, to differentiate host-specific differences, subsets of each host for different infection TPs were used. Finally, all TPs for both hosts were used with host-only genes to form co-expression networks among host genes.

WGCNA analysis begins with creation of a Pearson correlation matrix of the expression of all gene pairs. These were used to filter highest correlated pairs where required. This step was followed by the creation of an approximately scale-free adjacency matrix, using a power function. The soft threshold parameter (B) for the power function in each case was determined based on the criterion of approximate scale-free topology, as described in the software manual (Langfelder and Horvath, 2008). The topological overlap matrix (TOM) was calculated to quantify the degree of overlap in shared neighbors. Finally, modules were created using a dynamic tree cut algorithm in WGCNA. To characterize each of the modules, module eigengenes and GO annotations were calculated. To calculate the similarities between various modules, Pearson correlation between eigengene vectors was used.

Deconvolution of Cell Populations

Cibersortx (Newman et al., 2019) was used to analyze gene expression data to obtain an estimation of abundances of individual cell types from mixed cell populations in the various blood samples. The LM22 signature matrix (Newman et al., 2015) was used as a cell type reference profile. Previously DESeq2 normalized expression data for all samples were used to estimate the abundances of the 22 cell types from whole blood.

To contrast various groups, the lmFit function (limma package) in R was used to model the cell populations as *Species* + *TimePoint* + *Species:TimePoint* and the eBayes function was used to compute log fold changes, *t* statistics, *p*-values and adjusted *p*-values, using the Benjamini–Hochberg method.

Dynamic Modeling of Tryptophan Metabolism

To understand the implications of transcriptomic changes during *P. knowlesi* infection, we used a well-established tryptophan metabolic model (Stavrum et al., 2013) and adjusted its parameters to represent changes in enzymatic activities in accordance to changes in the expression of corresponding genes (Tang et al., 2018).

The model was originally developed for liver tissue and had to be adapted for blood. Due to the lack of tissue specific enzyme concentration data, we used gene expression data for individual tissues (in this case blood *vs* liver data from the GTEx project (Consortium, 2013)) to form a crude estimate of enzymatic concentration. Each reaction rate v in the model is described with the Michaelis-Menten rate function

$$v = \frac{V_{\max} \cdot S}{K_m + S} \tag{1}$$

where $V_{\max}$ is the maximum reaction rate, S is the substrate concentration and K_m is the Michaelis constant. According to our assumption of proportionality between gene expression and enzyme activity (Tang et al., 2018), $V_{\max}$ is a function of enzyme concentration and enzymatic turnover K_{cat}. Since enzyme concentration is difficult to calculate, mRNA levels were used as approximate quantities:

$$V_{\max} = F.mRNA.K_{cat} \tag{2}$$

Here, F is a factor that converts expression values into enzyme concentrations and $mRNA$ is the measured expression.

Once the parameters were updated, the model was simulated to a steady state to obtain baseline metabolite concentrations and fluxes for the blood model.

Next, the kinetic parameters were updated by a factor corresponding to the fold change in gene expression in order to obtain the appropriate enzymatic activity, similar to **Eqs. 1** and **2**. For each case, the model was simulated to the steady state of all metabolite concentrations and fluxes were used for comparison of different scenarios.

For flux control analysis (Wildermuth, 2000) (**Eq. 3**), the control coefficients were calculated as

$$C^S_{v_i} = \frac{d \ln J}{d \ln v_i} \tag{3}$$

where $C^S_{v_i}$ is the flux control coefficient for the pathway flux J with small changes in enzyme activity v_i of step i.

DATA AVAILABILITY STATEMENT

Only publicly available datasets were analyzed in this study. They can be found at https://plasmodb.org/plasmo/app/static-content/PlasmoDB/mahpic.html (Experiments 06 and 07) or under BioProject accession PRJNA524357 (https://www.ncbi.nlm.nih.gov/bioproject/524357) and PRJNA526495 (https://www.ncbi.nlm.nih.gov/bioproject/526495).

ETHICS STATEMENT

All experiments involving NHPs were performed at the Yerkes National Primate Research Center (YNPRC), an AAALAC

International-accredited facility. All methods were carried out in accordance with relevant guidelines and regulations. Specifically, all procedures followed ARRIVE guidelines and were approved by Emory's IACUC and the Animal Care and Use Review Office (ACURO) of the US Department of Defense and followed accordingly. The Emory's IACUC approval number was PROTO201700484—YER-2003344-ENTRPR-A.

AUTHOR CONTRIBUTIONS

AG, MG and EV conceived the study; AG executed all analyses; all authors contributed to writing and editing the article.

FUNDING

This project was funded in part by the National Institute of Allergy and Infectious Diseases; National Institutes of Health, Department of Health and Human Services, which established the MaHPIC (Contract Nos. HHSN272201200031C; MG), the NIH Office of Research Infrastructure Programs/OD P51OD011132, the Defense Advanced Research Program Agency and the US Army Research Office via a cooperative agreement (Contract Nos. W911NF16C0008; MG0), which funded the Technologies for Host Resilience - Host Acute Models of Malaria to study Experimental Resilience (THoR's HAMMER) consortium, and the Georgia Research Alliance.

ACKNOWLEDGMENTS

The authors are very grateful to Regina Joice Cordy for generating the Biocrates targeted metabolomics data and performing initial analyses that provided insights into the role of kynurenine in the host response to malaria. They also acknowledge Luis L. Fonseca and other members of the Malaria Host-Pathogen Interaction Center (MaHPIC) Consortium (http://systemsbiology.emory.edu) for insightful discussions and for the NHP longitudinal infection and data management work on which this analysis is based (Peterson et al., 2021). The authors would also like to thank the Yerkes National Primate Research Center's Genomics Core for sequencing services.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2021.800721/full#supplementary-material

REFERENCES

Aikawa, M., Webster, K., Collins, W. E., Brown, A., Smith, C. D., Tegoshi, T., et al. (1992). A Primate Model for Human Cerebral Malaria: *Plasmodium Coatneyi*-Infected Rhesus Monkeys. *Am. J. Trop. Med. Hyg.* 46, 391–397. doi:10.4269/ajtmh.1992.46.391

al-Khedery, B., Barnwell, J. W., and Galinski, M. R. (1999). Antigenic Variation in Malaria. *Mol. Cel* 3, 131–141. doi:10.1016/s1097-2765(00)80304-4

Albert, B., Kos-Braun, I. C., Henras, A. K., Dez, C., Rueda, M. P., Zhang, X., et al. (2019). A Ribosome Assembly Stress Response Regulates Transcription to Maintain Proteome Homeostasis. *eLife* 8, e45002. doi:10.7554/eLife.45002

Alberts, B. (2002). *Molecular Biology of the Cell*. New York: Garland Science.

Amarante-Mendes, G. P., Adjemian, S., Branco, L. M., Zanetti, L. C., Weinlich, R., and Bortoluci, K. R. (2018). Pattern Recognition Receptors and the Host Cell Death Molecular Machinery. *Front. Immunol.* 9, 2379. doi:10.3389/fimmu.2018.02379

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene Ontology: Tool for the Unification of Biology. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Awad, F., Assrawi, E., Jumeau, C., Georgin-Lavialle, S., Cobret, L., Duquesnoy, P., et al. (2017). Impact of Human Monocyte and Macrophage Polarization on NLR Expression and NLRP3 Inflammasome Activation. *PLoS One* 12, e0175336. doi:10.1371/journal.pone.0175336

Banzola, I., Mengus, C., Wyler, S., Hudolin, T., Manzella, G., Chiarugi, A., et al. (2018). Expression of Indoleamine 2,3-Dioxygenase Induced by IFN-γ and TNF-α as Potential Biomarker of Prostate Cancer Progression. *Front. Immunol.* 9, 1051. doi:10.3389/fimmu.2018.01051

Barber, B. E., Rajahram, G. S., Grigg, M. J., William, T., and Anstey, N. M. (2017). World Malaria Report: Time to Acknowledge *Plasmodium Knowlesi* Malaria. *Malar. J.* 16, 135. doi:10.1186/s12936-017-1787-y

Barberá-Cremades, M., Baroja-Mazo, A., and Pelegrín, P. (2016). Purinergic Signaling during Macrophage Differentiation Results in M2 Alternative Activated Macrophages. *J. Leukoc. Biol.* 99, 289–299. doi:10.1189/jlb.1a0514-267rr

Baroncelli, S., Negri, D. R., Michelini, Z., and Cara, A. (2008). *Macaca mulatta*,fascicularisandnemestrinain AIDS Vaccine Development. *Expert Rev. Vaccin.* 7, 1419–1434. doi:10.1586/14760584.7.9.1419

Biggs, B. A., Gooze, L., Wycherley, K., Wollish, W., Southwell, B., Leech, J. H., et al. (1991). Antigenic Variation in *Plasmodium Falciparum*. *Proc. Natl. Acad. Sci.* 88, 9171–9174. doi:10.1073/pnas.88.20.9171

Boye, K., and Mælandsmo, G. M. (2010). S100A4 and Metastasis. *Am. J. Pathol.* 176, 528–535. doi:10.2353/ajpath.2010.090526

Brown, K. N., and Brown, I. N. (1965). Immunity to Malaria: Antigenic Variation in Chronic Infections of *Plasmodium Knowlesi*. *Nature* 208, 1286–1288. doi:10.1038/2081286a0

Chen, X., Cao, R., and Zhong, W. (2019). Host Calcium Channels and Pumps in Viral Infections. *Cells* 9. doi:10.3390/cells9010094

Cicchese, J. M., Evans, S., Hult, C., Joslyn, L. R., Wessler, T., Millar, J. A., et al. (2018). Dynamic Balance of Pro- and Anti-inflammatory Signals Controls Disease and Limits Pathology. *Immunol. Rev.* 285, 147–167. doi:10.1111/imr.12671

Cipolletta, D., Feuerer, M., Li, A., Kamei, N., Lee, J., Shoelson, S. E., et al. (2012). PPAR-γ Is a Major Driver of the Accumulation and Phenotype of Adipose Tissue Treg Cells. *Nature* 486, 549–553. doi:10.1038/nature11132

Coatney, C. W. G. R., Warren, M., and Contacos, P. G. (2003). *The Primate Malarias [original Book Published 1971*. Atlanta, GA: CDC.

Coatney, G. R., Collins, W. E., Warren, M., and Contacos, P. G. (1971). *The Primate Malarias*.

Colvin, H. N., and Joice Cordy, R. (2020). Insights into Malaria Pathogenesis Gained from Host Metabolomics. *Plos Pathog.* 16, e1008930. doi:10.1371/journal.ppat.1008930

Consortium, G. T. (2013). The Genotype-Tissue Expression (GTEx) Project. *Nat. Genet.* 45, 580–585. doi:10.1038/ng.2653

Cordy, R. J., Patrapuvich, R., Lili, L. N., Cabrera-Mora, M., Chien, J. T., Tharp, G. K., et al. (2019). Distinct Amino Acid and Lipid Perturbations Characterize Acute versus Chronic Malaria. *JCI Insight* 4, e125156. doi:10.1172/jci.insight.125156

Counihan, N. A., Modak, J. K., and de Koning-Ward, T. F. (2021). How Malaria Parasites Acquire Nutrients from Their Host. *Front. Cel Develop. Biol.* 9. doi:10.3389/fcell.2021.649184

Covens, K., Verbinnen, B., Geukens, N., Meyts, I., Schuit, F., Van Lommel, L., et al. (2013). Characterization of Proposed Human B-1 Cells Reveals Pre-plasmablast Phenotype. *Blood* 121, 5176–5183. doi:10.1182/blood-2012-12-471953

Cox-Singh, J. (2012). Zoonotic Malaria. *Curr. Opin. Infect. Dis.* 25, 530–536. doi:10.1097/qco.0b013e3283558780

Craig, A. G., Grau, G. E., Janse, C., Kazura, J. W., Milner, D., Barnwell, J. W., et al. (2012). The Role of Animal Models for Research on Severe Malaria. *Plos Pathog.* 8, e1002401. doi:10.1371/journal.ppat.1002401

Crespi, B., and Alcock, J. (2021). Conflicts over Calcium and the Treatment of COVID-19. *Evol. Med. Public Health* 9, 149–156. doi:10.1093/emph/eoaa046

Crotty, S. (2014). T Follicular Helper Cell Differentiation, Function, and Roles in Disease. *Immunity* 41, 529–542. doi:10.1016/j.immuni.2014.10.004

Daneshvar, C., Davis, T. M. E., Cox-Singh, J., Rafa'ee, M. Z., Zakaria, S. K., Divis, P. C. S., et al. (2009). Clinical and Laboratory Features of HumanPlasmodium knowlesiInfection. *Clin. Infect. Dis.* 49, 852–860. doi:10.1086/605439

De la Mota-Peynado, A., Chernoff, J., and Beeser, A. (2011). Identification of the Atypical MAPK Erk3 as a Novel Substrate for P21-Activated Kinase (Pak) Activity. *J. Biol. Chem.* 286, 13603–13611. doi:10.1074/jbc.m110.181743

Déléris, P., Trost, M., Topisirovic, I., Tanguay, P.-L., Borden, K. L. B., Thibault, P., et al. (2011). Activation Loop Phosphorylation of ERK3/ERK4 by Group I P21-Activated Kinases (PAKs) Defines a Novel PAK-Erk3/4-MAPK-Activated Protein Kinase 5 Signaling Pathway. *J. Biol. Chem.* 286, 6470–6478. doi:10.1074/jbc.m110.181529

El Mubarak, H. S., Yüksel, S., van Amerongen, G., Mulder, P. G. H., Mukhtar, M. M., Osterhaus, A. D. M. E., et al. (2007). Infection of Cynomolgus Macaques (*Macaca fascicularis*) and Rhesus Macaques (*Macaca mulatta*) with Different Wild-type Measles Viruses. *J. Gen. Virol.* 88, 2028–2034. doi:10.1099/vir.0.82804-0

Fell, D. A. (1992). Metabolic Control Analysis: a Survey of its Theoretical and Experimental Development. *Biochem. J.* 286 (Pt 2) (Pt 2), 313–330. doi:10.1042/bj2860313

Fell, D. (2007). *Understanding the Control of Metabolism*. London: Portland Press.

Ferreira, M. U., da Silva Nunes, M., and Wunderlich, G. (2004). Antigenic Diversity and Immune Evasion by Malaria Parasites. *Clin. Vaccin. Immunol* 11, 987–995. doi:10.1128/cdli.11.6.987-995.2004

Festjens, N., Vanden Berghe, T., Cornelis, S., and Vandenabeele, P. (2007). RIP1, a kinase on the crossroads of a cell's decision to live or die. *Cell Death Differ* 14, 400–410. doi:10.1038/sj.cdd.4402085

El Mubarak, H. S., Yüksel, S., van Amerongen, G., Mulder, P. G. H., Mukhtar, M. M., Osterhaus, A. D. M. E., et al. Infection of Cynomolgus Macaques (*Macaca fascicularis*) and Rhesus Macaques (*Macaca mulatta*) with Different Wild-type Measles Viruses. 88 (2007) 2028–2034.doi:10.1099/vir.0.82804-0

Fox, D., and Man, S. M. (2019). DDX3X: Stressing the NLRP3 Inflammasome. *Cell Res* 29, 969–970. doi:10.1038/s41422-019-0250-8

França, C. T., White, M. T., He, W. Q., Hostetler, J. B., Brewster, J., Frato, G., et al. (2017). Identification of Highly-Protective Combinations of *Plasmodium Vivax* Recombinant Proteins for Vaccine Development. *eLife* 6, e28673. doi:10.7554/eLife.28673

Fuller, T. F., Ghazalpour, A., Aten, J. E., Drake, T. A., Lusis, A. J., and Horvath, S. (2007). Weighted Gene Coexpression Network Analysis Strategies Applied to Mouse Weight. *Mamm. Genome* 18, 463–472. doi:10.1007/s00335-007-9043-3

Galinski, M. R., Lapp, S. A., Peterson, M. S., Ay, F., Joyner, C. J., Le Roch, K. G., et al. (2018). Plasmodium Knowlesi:a Superbin Vivononhuman Primate Model of Antigenic Variation in Malaria. *Parasitology* 145, 85–100. doi:10.1017/s0031182017001135

Garcia, M. A., Bouley, D. M., Larson, M. J., Lifland, B., Moorhead, R., Simkins, M. D., et al. (2004). Outbreak of Mycobacterium Bovis in a Conditioned Colony of Rhesus (*Macaca mulatta*) and Cynomolgus (*Macaca fascicularis*) Macaques. *Comp. Med.* 54, 578–584.

Gardner, M. B., and Luciw, P. A. (2008). Macaque Models of Human Infectious Disease. *ILAR J.* 49, 220–255. doi:10.1093/ilar.49.2.220

Gene, C. (2021). The Gene Ontology Resource: Enriching a GOld Mine. *Nucleic Acids Res.* 49, D325–D334. doi:10.1093/nar/gkaa1113

Glushakova, S., Lizunov, V., Blank, P. S., Melikov, K., Humphrey, G., and Zimmerberg, J. (2013). Cytoplasmic Free Ca2+ Is Essential for Multiple Steps in Malaria Parasite Egress from Infected Erythrocytes. *Malar. J.* 12, 41. doi:10.1186/1475-2875-12-41

Godec, J., Tan, Y., Liberzon, A., Tamayo, P., Bhattacharya, S., Butte, A. J., et al. (2016). Compendium of Immune Signatures Identifies Conserved and Species-specific Biology in Response to Inflammation. *Immunity* 44, 194–206. doi:10.1016/j.immuni.2015.12.006

Goldsmith, Z. G., and Dhanasekaran, D. N. (2007). G Protein Regulation of MAPK Networks. *Oncogene* 26, 3122–3142. doi:10.1038/sj.onc.1210407

Golomb, L., Volarevic, S., and Oren, M. (2014). p53 and Ribosome Biogenesis Stress: The Essentials. *FEBS Lett.* 588, 2571–2579. doi:10.1016/j.febslet.2014.04.014

Gupta, A., Styczynski, M. P., Galinski, M. R., Voit, E. O., and Fonseca, L. L. (2021). Dramatic Transcriptomic Differences in *Macaca mulatta* and *Macaca fascicularis* with *Plasmodium Knowlesi* Infections. *Sci. Rep.* 11, 19519. doi:10.1038/s41598-021-98024-6

Gutiérrez-Vázquez, C., and Quintana, F. J. (2018). Regulation of the Immune Response by the Aryl Hydrocarbon Receptor. *Immunity* 48, 19–33. doi:10.1016/j.immuni.2017.12.012

Hansen, A. M., Driussi, C., Turner, V., Takikawa, O., and Hunt, N. H. (2000). Tissue Distribution of Indoleamine 2,3-dioxygenase in normal and Malaria-Infected Tissue. *Redox Rep.* 5, 112–115. doi:10.1179/135100000101535384

Haupt, S., Mejía-Hernández, J. O., Vijayakumaran, R., Keam, S. P., and Haupt, Y. (2019). The Long and the Short of it: the MDM4 Tail So Far. *J. Mol. Cel Biol.* 11, 231–244. doi:10.1093/jmcb/mjz007

Haupt, Y., Robles, A. I., Prives, C., and Rotter, V. (2002). Deconstruction of P53 Functions and Regulation. *Oncogene* 21, 8223–8231. doi:10.1038/sj.onc.1206137

Hedges, S. B., Marin, J., Suleski, M., Paymer, M., and Kumar, S. (2015). Tree of Life Reveals Clock-like Speciation and Diversification. *Mol. Biol. Evol.* 32, 835–845. doi:10.1093/molbev/msv037

Howard, R. J., Barnwell, J. W., and Kao, V. (1983). Antigenic Variation of *Plasmodium Knowlesi* Malaria: Identification of the Variant Antigen on Infected Erythrocytes. *Proc. Natl. Acad. Sci.* 80, 4129–4133. doi:10.1073/pnas.80.13.4129

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211

Hubbard, T. D., Murray, I. A., and Perdew, G. H. (2015). Indole and Tryptophan Metabolism: Endogenous and Dietary Routes to Ah Receptor Activation. *Drug Metab. Dispos* 43, 1522–1535. doi:10.1124/dmd.115.064246

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2020). The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* 48, D498–d503. doi:10.1093/nar/gkz1031

Joerger, A. C., and Fersht, A. R. (2016). The P53 Pathway: Origins, Inactivation in Cancer, and Emerging Therapeutic Approaches. *Annu. Rev. Biochem.* 85, 375–404. doi:10.1146/annurev-biochem-060815-014710

Joyner, C., Barnwell, J. W., and Galinski, M. R. (2015). No More Monkeying Around: Primate Malaria Model Systems Are Key to Understanding *Plasmodium Vivax* Liver-Stage Biology, Hypnozoites, and Relapses. *Front. Microbiol.* 6, 145. doi:10.3389/fmicb.2015.00145

Kapitulnik, J., and Gonzalez, F. J. (1993). Marked Endogenous Activation of the CYP1A1 and CYP1A2 Genes in the Congenitally Jaundiced Gunn Rat. *Mol. Pharmacol.* 43, 722–725.

Knowles, R., and Gupta, B. M. D. (1932). A Study of Monkey-Malaria, and its Experimental Transmission to Man. *Ind. Med. Gaz.* 67, 301–320.

Koo, B.-S., Lee, D.-H., Kang, P., Jeong, K.-J., Lee, S., Kim, K., et al. (2019). Reference Values of Hematological and Biochemical Parameters in Young-Adult Cynomolgus Monkey (*Macaca fascicularis*) and Rhesus Monkey (*Macaca mulatta*) Anesthetized with Ketamine Hydrochloride. *Lab. Anim. Res.* 35, 7. doi:10.1186/s42826-019-0006-0

Kostenko, S., Dumitriu, G., and Moens, U. (2012). Tumour Promoting and Suppressing Roles of the Atypical MAP Kinase Signalling Pathway ERK3/4-MK5. *Jms* 7, 9. doi:10.1186/1750-2187-7-9

Langfelder, P., and Horvath, S. (2008). WGCNA: an R Package for Weighted Correlation Network Analysis. *BMC Bioinformatics* 9, 559. doi:10.1186/1471-2105-9-559

Lapp, S. A., Geraldo, J. A., Chien, J.-T., Ay, F., Pakala, S. B., Batugedara, G., et al. (2018). PacBio Assembly of aPlasmodium Knowlesigenome Sequence with Hi-C Correction and Manual Annotation of theSICAvargene Family. *Parasitology* 145, 71–84. doi:10.1017/s0031182017001329

Lapp, S. A., Korir, C. C., and Galinski, M. R. (2009). Redefining the Expressed Prototype SICAvar Gene Involved in *Plasmodium Knowlesi* Antigenic Variation. *Malar. J.* 8, 181. doi:10.1186/1475-2875-8-181

Lee, H. J., Georgiadou, A., Walther, M., Nwakanma, D., Stewart, L. B., Levin, M., et al. (2018). Integrated Pathogen Load and Dual Transcriptome Analysis of Systemic Host-Pathogen Interactions in Severe Malaria. *Sci. Transl Med.* 10. doi:10.1126/scitranslmed.aar3619

Lee, J.-H., Kang, Y., Khare, V., Jin, Z.-Y., Kang, M.-Y., Yoon, Y., et al. (2010). The P53-Inducible Gene 3 (PIG3) Contributes to Early Cellular Response to DNA Damage. *Oncogene* 29, 1431–1450. doi:10.1038/onc.2009.438

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. *Cel Syst.* 1, 417–425. doi:10.1016/j.cels.2015.12.004

Lu, S., Zhao, Y., Yu, W., Yang, Y., Gao, J., Wang, J., et al. (2020). Comparison of Nonhuman Primates Identified the Suitable Model for COVID-19. *Sig Transduct Target. Ther.* 5, 157. doi:10.1038/s41392-020-00269-6

MacLeod, M. K. L., Clambey, E. T., Kappler, J. W., and Marrack, P. (2009). CD4 Memory T Cells: what Are They and what Can They Do. *Semin. Immunol.* 21, 53–61. doi:10.1016/j.smim.2009.02.006

Maiello, P., DiFazio, R. M., Cadena, A. M., Rodgers, M. A., Lin, P. L., Scanga, C. A., et al. (2018). Rhesus Macaques Are More Susceptible to Progressive Tuberculosis Than Cynomolgus Macaques: a Quantitative Comparison. *Infect. Immun.* 86, e00505–17. doi:10.1128/IAI.00505-17

Mogensen, T. H. (2009). Pathogen Recognition and Inflammatory Signaling in Innate Immune Defenses. *Clin. Microbiol. Rev.* 22, 240–273. doi:10.1128/cmr.00046-08

Morales, J. C., and Melnick, D. J. (1998). Phylogenetic Relationships of the Macaques (Cercopithecidae:), as Revealed by High Resolution Restriction Site Mapping of Mitochondrial Ribosomal Genes. *J. Hum. Evol.* 34, 1–23. doi:10.1006/jhev.1997.0171

Moyes, C. L., Henry, A. J., Golding, N., Huang, Z., Singh, B., Baird, J. K., et al. (2014). Defining the Geographical Range of the Plasmodium Knowlesi Reservoir. *Plos Negl. Trop. Dis.* 8, e2780. doi:10.1371/journal.pntd.0002780

Murray, I. A., Patterson, A. D., and Perdew, G. H. (2014). Aryl Hydrocarbon Receptor Ligands in Cancer: Friend and Foe. *Nat. Rev. Cancer* 14, 801–814. doi:10.1038/nrc3846

Napier, L. E., and Campbell, H. G. M. (1932). Observations on a Plasmodium Infection Which Causes Hæmoglobinuria in Certain Species of Monkey. *Ind. Med. Gaz.* 67, 246–249.

New, D. C., and Wong, Y. H. (2007). Molecular Mechanisms Mediating the G Protein-Coupled Receptor Regulation of Cell Cycle Progression. *Jms* 2, 2. doi:10.1186/1750-2187-2-2

Newman, A. M., Liu, C. L., Green, M. R., Gentles, A. J., Feng, W., Xu, Y., et al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nat. Methods* 12, 453–457. doi:10.1038/nmeth.3337

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., et al. (2019). Determining Cell Type Abundance and Expression from Bulk Tissues with Digital Cytometry. *Nat. Biotechnol.* 37, 773–782. doi:10.1038/s41587-019-0114-2

Oki, S., Ohta, T., Shioi, G., Hatanaka, H., Ogasawara, O., Okuda, Y., et al. (2018). ChIP-Atlas: a Data-Mining Suite Powered by Full Integration of Public ChIP-Seq Data. *EMBO Rep.* 19. doi:10.15252/embr.201846255

Ouattara, A., Barry, A. E., Dutta, S., Remarque, E. J., Beeson, J. G., and Plowe, C. V. (2015). Designing Malaria Vaccines to Circumvent Antigen Variability. *Vaccine* 33, 7506–7512. doi:10.1016/j.vaccine.2015.09.110

Pain, A., Böhme, U., Berry, A. E., Mungall, K., Finn, R. D., Jackson, A. P., et al. (2008). The Genome of the Simian and Human Malaria Parasite *Plasmodium Knowlesi*. *Nature* 455, 799–803. doi:10.1038/nature07306

Pan, S.-C., Li, C.-Y., Kuo, C.-Y., Kuo, Y.-Z., Fang, W.-Y., Huang, Y.-H., et al. (2018). The P53-S100a2 Positive Feedback Loop Negatively Regulates Epithelialization in Cutaneous Wound Healing. *Sci. Rep.* 8, 5458. doi:10.1038/s41598-018-23697-5

Pasini, E. M., Zeeman, A.-M., Voorberg-van Der Wel, A., and Kocken, C. H. M. (2018). *Plasmodium Knowlesi*: a Relevant, Versatile Experimental Malaria Model. *Parasitology* 145, 56–70. doi:10.1017/s0031182016002286

Peterson, M. S., Joyner, C. J., Brady, J. A., Wood, J. S., Cabrera-Mora, M., Saney, C. L., et al. (2021). Clinical Recovery of *Macaca fascicularis* Infected with *Plasmodium Knowlesi*. *Malar. J*, 486. doi:10.1186/s12936-021-03925-6

Phelan, D., Winter, G. M., Rogers, W. J., Lam, J. C., and Denison, M. S. (1998). Activation of the Ah Receptor Signal Transduction Pathway by Bilirubin and Biliverdin. *Arch. Biochem. Biophys.* 357, 155–163. doi:10.1006/abbi.1998.0814

Pinski, A. N., Maroney, K. J., Marzi, A., and Messaoudi, I. (2021). Distinct Transcriptional Responses to Fatal Ebola Virus Infection in Cynomolgus and Rhesus Macaques Suggest Species-specific Immune Responses. *Emerging Microbes & Infections* 10, 1320–1330. doi:10.1080/22221751.2021.1942229

Raja, T. N., Hu, T. H., Kadir, K. A., Mohamad, D. S. A., Rosli, N., Wong, L. L., et al. (2020). Naturally Acquired Human *Plasmodium Cynomolgi* and *P. Knowlesi* Infections, Malaysian Borneo. *Emerg. Infect. Dis.* 26, 1801–1809. doi:10.3201/eid2608.200343

Rénia, L., and Goh, Y. S. (2016). Malaria Parasites: The Great Escape. *Front. Immunol.* 7, 463. doi:10.3389/fimmu.2016.00463

Roager, H. M., and Licht, T. R. (2018). Microbial Tryptophan Catabolites in Health and Disease. *Nat. Commun.* 9, 3294. doi:10.1038/s41467-018-05470-4

Roh, J. S., and Sohn, D. H. (2018). Damage-Associated Molecular Patterns in Inflammatory Diseases. *Immune Netw.* 18, e27. doi:10.4110/in.2018.18.e27

Sablina, A. A., Budanov, A. V., Ilyinskaya, G. V., Agapova, L. S., Kravchenko, J. E., and Chumakov, P. M. (2005). The Antioxidant Function of the P53 Tumor Suppressor. *Nat. Med.* 11, 1306–1313. doi:10.1038/nm1320

Saha, K., Adhikary, G., Kanade, S. R., Rorke, E. A., and Eckert, R. L. (2014). p38δ Regulates P53 to Control p21Cip1 Expression in Human Epidermal Keratinocytes. *J. Biol. Chem.* 289, 11443–11453. doi:10.1074/jbc.m113.543165

Salguero, F. J., White, A. D., Slack, G. S., Fotheringham, S. A., Bewley, K. R., Gooch, K. E., et al. (2021). Comparison of Rhesus and Cynomolgus Macaques as an Infection Model for COVID-19. *Nat. Commun.* 12, 1260. doi:10.1038/s41467-021-21389-9

Sanni, L. A., Thomas, S. R., Tattam, B. N., Moore, D. E., Chaudhri, G., Stocker, R., et al. (1998). Dramatic Changes in Oxidative Tryptophan Metabolism along the Kynurenine Pathway in Experimental Cerebral and Noncerebral Malaria. *Am. J. Pathol.* 152, 611–619.

Santos, R. O. d., Cruz, M. G. S. d., Lopes, S. C. P., Oliveira, L. B., Nogueira, P. A., Lima, E. S., et al. (2020). A First *Plasmodium Vivax* Natural Infection Induces Increased Activity of the Interferon Gamma-Driven Tryptophan Catabolism Pathway. *Front. Microbiol.* 11, 400. doi:10.3389/fmicb.2020.00400

Sarkar, S. A., Wong, R., Hackl, S. I., Moua, O., Gill, R. G., Wiseman, A., et al. (2007). Induction of Indoleamine 2,3-Dioxygenase by Interferon-γ in Human Islets. *Diabetes* 56, 72–79. doi:10.2337/db06-0617

Sasseville, V. G., and Mansfield, K. G. (2010). Overview of Known Non-human Primate Pathogens with Potential to Affect Colonies Used for Toxicity Testing. *J. Immunotoxicology* 7, 79–92. doi:10.3109/15476910903213521

Schroder, K., and Tschopp, J. (2010). The Inflammasomes. *Cell* 140, 821–832. doi:10.1016/j.cell.2010.01.040

Shaulian, E., and Karin, M. (2001). AP-1 in Cell Proliferation and Survival. *Oncogene* 20, 2390–2400. doi:10.1038/sj.onc.1204383

Singh, B., and Daneshvar, C. (2013). Human Infections and Detection of Plasmodium Knowlesi. *Clin. Microbiol. Rev.* 26, 165–184. doi:10.1128/cmr.00079-12

Singh, B., Sung, L. K., Matusop, A., Radhakrishnan, A., Shamsul, S. S., Cox-Singh, J., et al. (2004). A Large Focus of Naturally Acquired *Plasmodium Knowlesi* Infections in Human Beings. *The Lancet* 363, 1017–1024. doi:10.1016/s0140-6736(04)15836-4

Stavrum, A.-K., Heiland, I., Schuster, S., Puntervoll, P., and Ziegler, M. (2013). Model of Tryptophan Metabolism, Readily Scalable Using Tissue-specific Gene Expression Data. *J. Biol. Chem.* 288, 34555–34566. doi:10.1074/jbc.m113.474908

Street, S. L., Kyes, R. C., Grant, R., and Ferguson, B. (2007). Single Nucleotide Polymorphisms (SNPs) Are Highly Conserved in Rhesus (*Macaca mulatta*) and Cynomolgus (*Macaca fascicularis*) Macaques. *BMC Genomics* 8, 480. doi:10.1186/1471-2164-8-480

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-wide Expression Profiles. *Proc. Natl. Acad. Sci.* 102, 15545–15550. doi:10.1073/pnas.0506580102

Tang, Y., Gupta, A., Garimalla, S., Galinski, M. R., Styczynski, M. P., Fonseca, L. L., et al. (2018). Metabolic Modeling Helps Interpret Transcriptomic Changes during Malaria. *Biochim. Biophys. Acta (Bba) - Mol. Basis Dis.* 1864, 2329–2340. doi:10.1016/j.bbadis.2017.10.023

Taylor, M. W., and Feng, G. (1991). Relationship between Interferon-γ, Indoleamine 2,3-dioxygenase, and Tryptophan Catabolism. *FASEB j.* 5, 2516–2522. doi:10.1096/fasebj.5.11.1907934

Tetsutani, K., To, H., Torii, M., Hisaeda, H., and Himeno, K. (2007). Malaria Parasite Induces Tryptophan-Related Immune Suppression in Mice. *Parasitology* 134, 923–930. doi:10.1017/s0031182007002326

Toker, A., Engelbert, D., Garg, G., Polansky, J. K., Floess, S., Miyao, T., et al. (2013). Active Demethylation of the Foxp3 Locus Leads to the Generation of Stable Regulatory T Cells within the Thymus. *J.I.* 190, 3180–3188. doi:10.4049/jimmunol.1203473

Toma-Jonik, A., Vydra, N., Janus, P., and Widłak, W. (2019). Interplay between HSF1 and P53 Signaling Pathways in Cancer Initiation and Progression: Non-oncogene and Oncogene Addiction. *Cell Oncol.* 42, 579–589. doi:10.1007/s13402-019-00452-0

Tosi, A. J., Morales, J. C., and Melnick, D. J. (2000). Comparison of Y Chromosome and mtDNA Phylogenies Leads to Unique Inferences of Macaque Evolutionary History. *Mol. Phylogenet. Evol.* 17, 133–144. doi:10.1006/mpev.2000.0834

Tran Van Nhieu, G., Dupont, G., and Combettes, L. (2018). Ca2+ Signals Triggered by Bacterial Pathogens and Microdomains. *Biochim. Biophys. Acta (Bba) - Mol. Cel Res.* 1865, 1838–1845. doi:10.1016/j.bbamcr.2018.08.007

Van Binnendijk, R. S., Van Der Heijden, R. W. J., and Osterhaus, A. D. M. E. (1995). "Monkeys in Measles Research," in *Measles Virus*. Editors V. ter Meulen and M. A. Billeter (Berlin, Heidelberg: Springer Berlin Heidelberg), 135–148. doi:10.1007/978-3-642-78621-1_9

Waag, D. M., Byrne, W. R., Estep, J., Gibbs, P., Pitt, M. L., and Banfield, C. M. (1999). Evaluation of Cynomolgus (*Macaca fascicularis*) and Rhesus (*Macaca mulatta*) Monkeys as Experimental Models of Acute Q Fever after Aerosol Exposure to Phase-I Coxiella Burnetii. *Lab. Anim. Sci.* 49, 634–638.

Wahlgren, M., Fernandez, V., Chen, Q., Svärd, S., and Hagblom, P. (1999). Waves of Malarial Var-Iations. *Cell* 96, 603–606. doi:10.1016/s0092-8674(00)80569-3

Wang, B., Lu, F., Cheng, Y., Chen, J.-H., Jeon, H.-Y., Ha, K.-S., et al. (2015). Immunoprofiling of the Tryptophan-Rich Antigen Family in *Plasmodium Vivax*. *Infect. Immun.* 83, 3083–3095. doi:10.1128/iai.03067-14

Wang, S., Song, R., Wang, Z., Jing, Z., Wang, S., and Ma, J. (2018). *Inflammation*, 9. doi:10.0A8/A9

Wang, S., Song, R., Wang, Z., Jing, Z., Wang, S., and Ma, J. (2018). S100A8/A9 in Inflammation. *Front. Immunol.* 9, 1298. doi:10.3389/fimmu.2018.01298

Wang, W., Huang, P., Jiang, N., Lu, H., Zhang, D., Wang, D., et al. (2018). A Thioredoxin Homologous Protein of *Plasmodium Falciparum* Participates in Erythrocyte Invasion. *Infect. Immun.* 86. doi:10.1128/IAI.00289-18

Wildermuth, M. C. (2000). Metabolic Control Analysis: Biological Applications and Insights. *Genome Biol.* 1, REVIEWS1031. doi:10.1186/gb-2000-1-6-reviews1031

World Malaria Report 2020. (2020)*20 Years of Global Progress and Challenges*., Geneva: World Health Organization, .

Xia, C., Braunstein, Z., Toomey, A. C., Zhong, J., and Rao, X. (2018). S100 Proteins as an Important Regulator of Macrophage Inflammation. *Front. Immunol.* 8, 1908. doi:10.3389/fimmu.2017.01908

Yindom, L. M., Forbes, R., Aka, P., Janha, O., Jeffries, D., Jallow, M., et al. (2012). Killer-cell Immunoglobulin-like Receptors and Malaria Caused by Plasmodium Falciparum in the Gambia. *Tissue Antigens* 79, 104–113. doi:10.1111/j.1399-0039.2011.01818.x

Yu, Y. C., Han, J. M., and Kim, S. (2021). Aminoacyl-tRNA Synthetases and Amino Acid Signaling. *Biochim. Biophys. Acta (Bba) - Mol. Cel Res.* 1868, 118889. doi:10.1016/j.bbamcr.2020.118889

Zaw, M. T., and Lin, Z. (2019). Human *Plasmodium Knowlesi* Infections in South-East Asian Countries. *J. Microbiol. Immunol. Infect.* 52, 679–684. doi:10.1016/j.jmii.2019.05.012

Zhang, B., and Horvath, S. (2005). A General Framework for Weighted Gene Co-expression Network Analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article17. doi:10.2202/1544-6115.1128

Zhang, W., and Liu, H. T. (2002). MAPK Signal Pathways in the Regulation of Cell Proliferation in Mammalian Cells. *Cel Res* 12, 9–18. doi:10.1038/sj.cr.7290105

Zhang, X., Gan, M., Li, J., Li, H., Su, M., Tan, D., et al. (2020). Endogenous Indole Pyruvate Pathway for Tryptophan Metabolism Mediated by IL4I1. *J. Agric. Food Chem.* 68, 10678–10684. doi:10.1021/acs.jafc.0c03735

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Gupta, Galinski and Voit. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 23 February 2022
doi: 10.3389/fmolb.2022.811145

Three-Dimensional Model of Sub-Plasmalemmal Ca^{2+} Microdomains Evoked by T Cell Receptor/CD3 Complex Stimulation

Diana Gil[1], *Björn-Philipp Diercks*[1], *Andreas H. Guse*[1] *and Geneviève Dupont*[2]*

[1]The Calcium Signalling Group, Department of Biochemistry and Molecular Cell Biology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, [2]Unit of Theoretical Chronobiology, Faculté des Sciences CP231, Université Libre de Bruxelles (ULB), Brussels, Belgium

Ca^{2+} signalling plays an essential role in T cell activation, which is a key step to start an adaptive immune response. During the transition from a quiescent to a fully activated state, Ca^{2+} microdomains of reduced spatial and temporal extents develop in the junctions between the plasma membrane and the endoplasmic reticulum (ER). These microdomains rely on Ca^{2+} entry from the extracellular medium, via the ORAI1/STIM1/STIM2 system that mediates store operated Ca^{2+} entry Store operated calcium entry. The mechanism leading to local store depletion and subsequent Ca^{2+} entry depends on the activation state of the cells. The initial, smaller microdomains are triggered by D-*myo*-inositol 1,4,5-trisphosphate (IP_3) signalling in response to T cell adhesion. T cell receptor (TCR)/CD3 stimulation then initiates nicotinic acid adenine dinucleotide phosphate signalling, which activates ryanodine receptors (RYR). We have recently developed a mathematical model to elucidate the spatiotemporal Ca^{2+} dynamics of the microdomains triggered by IP_3 signalling in response to T cell adhesion (Gil et al., 2021). This reaction-diffusion model describes the evolution of the cytosolic and endoplasmic reticulum Ca^{2+} concentrations in a three-dimensional ER-PM junction and was solved using COMSOL Multiphysics. Modelling predicted that adhesion-dependent microdomains result from the concerted activity of IP_3 receptors and pre-formed ORAI1-STIM2 complexes. In the present study, we extend this model to include the role of RYRs rapidly after TCR/CD3 stimulation. The involvement of STIM1, which has a lower K_D for Ca^{2+} than STIM2, is also considered. Detailed 3D spatio-temporal simulations show that these Ca^{2+} microdomains rely on the concerted opening of ~7 RYRs that are simultaneously active in response to the increase in NAADP induced by T cell stimulation. Opening of these RYRs provoke a local depletion of ER Ca^{2+} that triggers Ca^{2+} flux through the ORAI1 channels. Simulations predict that RYRs are most probably located around the junction and that the increase in junctional Ca^{2+} concentration results from the combination between diffusion of Ca^{2+} released through the RYRs and Ca^{2+} entry through ORAI1 in the junction. The computational model moreover provides a tool allowing to investigate how Ca^{2+} microdomains occur, extend and interact in various states of T cell activation.

Keywords: T cells, ER-PM junctions, ryanodine receptors, NAADP, COMSOL, computational model, store operated calcium entry, ca^{2+} signalling

OPEN ACCESS

Edited by:
Ernesto Perez-Rueda, Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Alexandre Lewalle, King's College London, United Kingdom
Thomas John Brett, Washington University in St. Louis, United States

***Correspondence:**
Geneviève Dupont
Genevieve.Dupont@ulb.be

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *08 November 2021*
Accepted: *24 January 2022*
Published: *23 February 2022*

Citation:
Gil D, Diercks B-P, Guse AH and Dupont G (2022) Three-Dimensional Model of Sub-Plasmalemmal Ca^{2+} Microdomains Evoked by T Cell Receptor/CD3 Complex Stimulation. Front. Mol. Biosci. 9:811145. doi: 10.3389/fmolb.2022.811145

INTRODUCTION

Calcium signaling plays a crucial role in the activation of T cells and the adaptative immune response. In particular, it controls transcriptional activation, proliferation, differentiation or secretion of cytokines (Feske, 2007; Trebak & Kinet, 2019). Increases of the free cytosolic Ca^{2+} concentration rely on Ca^{2+} release from the endoplasmic reticulum (ER) and on Ca^{2+} entry from the extracellular medium. Mobilization of internal Ca^{2+} follows the increase in D-*myo*-inositol 1,4,5-trisphosphate (IP_3) and in nicotinic acid adenine dinucleotide phosphate (NAADP), via IP_3 receptors (IP_3R) and type 1 ryanodine receptor (RYR1), respectively (Streb et al., 1983; Wolf et al., 2015). Ca^{2+} entry relies on the ORAI/STIM system that allows Ca^{2+} entry in the cytosol, at a rate that is regulated by the concentration of Ca^{2+} in lumen of the ER (Putney, 2009). When Ca^{2+} dissociates from the Ca^{2+} sensors stromal interaction molecules 1 (STIM1) and 2 (STIM2) located in the ER membrane, STIM molecules aggregate and move to so-called "junctional spaces". These regions correspond to the narrow cytosolic spaces between the ER and PM membranes, at locations where these membranes are separated by distances smaller than 20 nm. There, STIM molecules can recruit ORAI1 to form Ca^{2+} channels allowing Ca^{2+} to enter into the cytoplasm. This process is known as capacitative or store operated Ca^{2+} entry (SOCE). The relation between SOCE and ER Ca^{2+} concentration is nonlinear, with a K_D for half activation of the order of 200 μM when it depends on the dissociation of Ca^{2+} from STIM1 and of 400 μM when it depends on the dissociation of Ca^{2+} from STIM2 (Stathopulos et al., 2006; Brandman et al., 2007; Luik et al., 2008).

Upon TCR/CD3 stimulation, second messengers create a substantial release of Ca^{2+} from the ER which, together with the resulting activation of SOCE, leads to a rise in the free cytosolic Ca^{2+} concentration in the whole T cell. This global Ca^{2+} increase contrasts with the locally restricted, sub-plasmalemmal Ca^{2+} increases of short duration (~50 ms) that can be observed as a consequence of adhesive interactions (Weiss and Diercks, unpublished results), or in the first seconds following TCR/CD3 stimulation (Wolf et al., 2015; Diercks et al., 2018). The two types of events are known as Ca^{2+} microdomains and have similar spatio-temporal characteristics. Yet, they have different molecular origins. The adhesion dependent Ca^{2+} microdomains rely on a pathway involving focal adhesion kinase (FAK), phospholipase C (PLC) and IP_3Rs (Weiss and Diercks, unpublished results). Due to IP_3R-mediated Ca^{2+} release and subsequent SOCE, they also rely on ORAI1-mediated Ca^{2+} entry. Computational simulations of the interplay between ORAI1 and IP_3R in a 3D configuration simulating an ER-PM junction have confirmed that the local depletion of ER Ca^{2+} created by the opening of a few IP_3R can trigger the opening of ORAI1 in the junction, even in conditions of a full ER (Gil et al., 2021). Interestingly, these non-TCR/CD3 dependent Ca^{2+} microdomains require the existence of pre-formed complexes of ORAI1 and STIM2 that were demonstrated experimentally (Diercks et al., 2018). Ca^{2+} microdomains characterized by somewhat larger amplitude (340 ± 11 nM *vs* 290 ± 12 nM) are observed during the first ~15 s following TCR/CD3 stimulation (Wolf et al., 2015; Diercks et al., 2018; reviewed in Guse et al., 2021). As the non-TCR/CD3 dependent microdomains, these signals also involve ORAI1, but in addition they require NAADP signaling and RYR1 dependent Ca^{2+} release from ER (Wolf et al., 2015; Diercks et al., 2018). As another difference, at this stage, pre-formed ORAI1/STIM complexes involve both STIM1 and STIM2 isoforms (Ahmad et al., 2021).

Ca^{2+} microdomains represent a crucial step for the successful activation of T cells. Reported durations of the signals triggered by cell adhesion and by formation of NAADP in the first seconds upon TCR/CD3 stimulation are 44 ± 4 ms and 64 ± 3 ms, respectively (Diercks et al., 2018). They extend on 0.216 ± 0.004 μm^2. Because of these limited temporal and spatial extents, the investigation of Ca^{2+} microdomains is technically limited by the resolution of the microscopic imaging system used. The spatial and temporal resolution of the imaging system used to characterize T cell Ca^{2+} microdomains is approx. 368 nm and 20–25 ms, respectively (Wolf et al., 2015). Mathematical modelling thus represents a useful complementary tool to investigate their molecular origin, in the line of the numerous studies devoted to small scale Ca^{2+} events (Solovey et al., 2008; Swaminathan et al., 2009; Thul et al., 2009; Rückl & Rüdiger, 2016; Walker et al., 2017).

In a previous study (Gil et al., 2021), we adapted the realistic three-dimensional mathematical description of the ER-PM junction proposed by McIvor et al. (2018) to simulate adhesion-dependent Ca^{2+} microdomains arising in T cells. This model describes Ca^{2+} dynamics in a confined 3D configuration corresponding to a junctional cytosolic space and the adjacent sub-PM ER, taking into account Ca^{2+} influx through ORAI1 and IP_3R, Ca^{2+} pumping into the ER through SERCA, and diffusion within the cytosolic and ER compartment. The IP_3Rs are supposed to be located close to the junctional space (Thillaiappan et al., 2017). Simulations using COMSOL Multiphysics showed that the spontaneous activity of ~3 IP_3Rs create a local depletion of ER Ca^{2+} that suffices to trigger the opening of ORAI1 channels located in the junction and thus, the onset of a microdomain. Because of the presence of pre-formed complexes of ORAI1 and STIM2 in unstimulated cells, opening of ORAI1 indeed rapidly follows the dissociation of Ca^{2+} from STIM2. Predictions of this model are in agreement with recent observations in HEK293 cells reporting that constitutive STIM2 clusters in ER-PM junctions sense decreases in local ER Ca^{2+} mediated by IP_3Rs (Ahmad et al., 2021). Moreover, IP_3R channel activity near the junctions was shown to favour STIM2 clustering in the junction.

In this study, we modified our previous model of the T cell junctions to address the molecular mechanism that underlies the TCR/CD3-evoked and NAADP and RYR-dependent Ca^{2+} microdomains occurring in the first ~15 s that follow TCR/CD3 stimulation. We first used modeling to find out whether RYR1 are located inside or around the ER-PM junctions. The analysis was based on comparisons between simulated and experimental results both in WT and ORAI1$^{-/-}$ T cells. The next issue related to the number of RYR1 involved in the formation of the junctional Ca^{2+} microdomains, which cannot

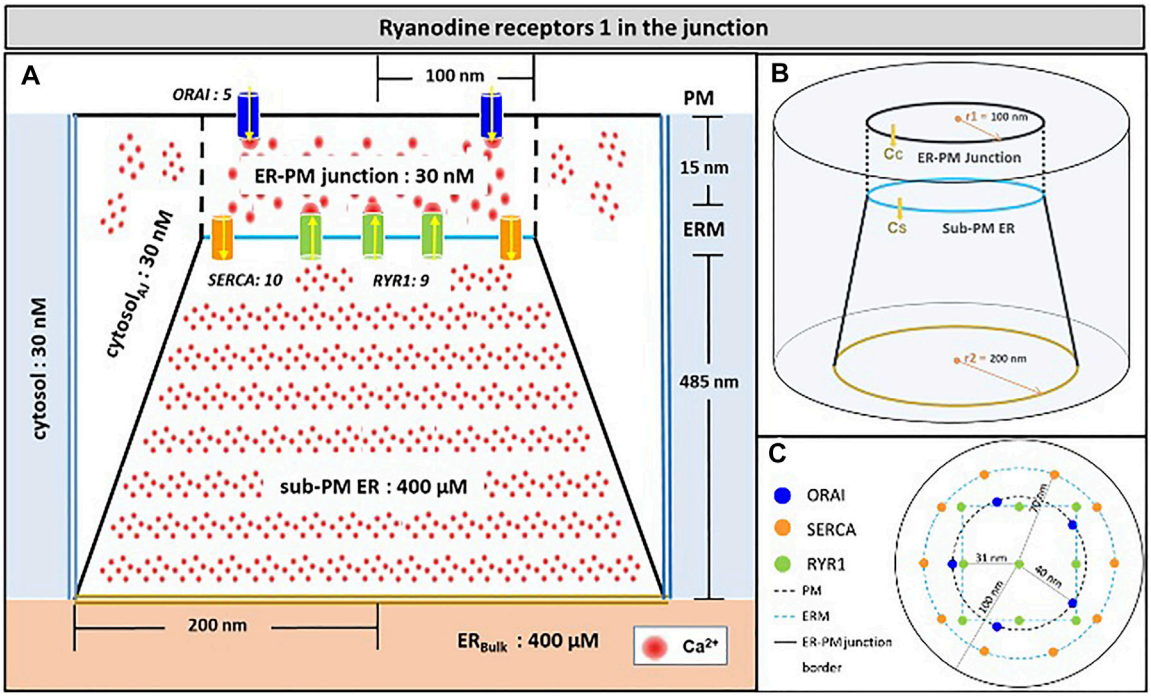


FIGURE 1 | Schematic representation of the model geometry of the ER-PM junction and sub-PM ER used to investigate the origin of the Ca^{2+} microdomains in T cells with nine RYR1 inside the junction **(A)** Frontal diagram showing the dimensions of the cone that represents the sub-PM ER, of the junction and of the portion of the cytosol considered in the simulations. ORAI1 channels are in blue, SERCA pumps in orange and RYR1 in green. Plain lines represent membrane boundaries; dashed lines, fictitious limits between the junction and the cytosol and double lines indicate the limits of the simulated system. The resting Ca^{2+} concentrations considered as initial conditions and boundary conditions in the two compartments are indicated. **(B)** 3D view of the model geometry **(C)** Upper view of the positions of the ORAI1 channels on the PM, in blue, and of the SERCA pumps and RYR1 (in a chessboard manner) on the ERM, in orange and green respectively. Not to scale. This geometry is based on McIvor et al. (2018). See text for details.

be directly inferred from experimental observations. By contrast, modeling can determine the number of RYR1 that must open simultaneously to create the local depletion of ER Ca^{2+} triggering the appropriate level of SOCE activation. Conclusions about this number were next validated by an independent estimation of the increase in the RYR1 open probability triggered by the NAADP formation in TCR/CD3 stimulated T cells. Finally, we took advantage of the great flexibility provided by computational modeling to investigate the respective roles played by the ER Ca^{2+} channels (IP_3R or RYR) and the Ca^{2+} sensors (STIM1 and STIM2) in shaping the characteristics of the Ca^{2+} microdomains created by the openings of the related ORAI1 channels. This analysis allowed us to propose a unifying description of the molecular mechanism underlying T cells Ca^{2+} microdomains from adhesion to early TCR/CD3 stimulation.

Description of the Mathematical Model

Because the model was fully described in Gil et al. (2020), we provide a concise description of its main features in this section. The spatial geometry is shown in **Figure 1**. The junction is a 15 nm-wide (Wu et al., 2006; Hogan, 2015) three-dimensional space between the PM and an ER portion located close to it. The junction communicates with the adjacent cytoplasm, a portion of which is modelled explicitly. The free cytosolic Ca^{2+} concentration, defined by C_C, is initially set at 30 nM (Diercks et al., 2018). In the rest of the cytoplasm, which is not modelled explicitly, C_C is fixed at this same value. Similarly, the evolution of ER Ca^{2+} concentration (C_S) is simulated in the sub-PM ER, which is in contact with the bulk of the ER where Ca^{2+} concentration is fixed at 400 µM (Lewis, 2011). The PM portion located in the junction contains five ORAI channels and the ER membrane, 10 SERCA and nine RYR1s as reported (Hogan, 2015; McIvor et al., 2018; Jayasinghe et al., 2018; Yin et al., 2008). Seen from above (**Figure 1C**), SERCA pumps form a ring surrounding ORAI1 channels and RYR1s are arranged on a square lattice. Given the large size of these channels (Lanner et al., 2010), the 31 nm distance between the pores of the channels considered in **Figure 1** corresponds to a close packing of RYR1.

Membranes, schematized as simple full lines in **Figure 1A**, correspond to no flux boundary conditions, except across channels and pumps where corresponding fluxes are simulated. The flux through ORAI channels is given by

$$J_{ORAI} = f\left(C_S^{loc}\right)\frac{I_{ORAI}}{F \cdot z \cdot A_o} \quad (1)$$

with I_{ORAI} the maximal single channel current, F the Faraday constant, z the charge of a Ca^{2+} ion and A_O the surface of the channel pore. $f\left(C_S^{loc}\right)$ is a function of the average local concentration of ER Ca^{2+} around the pore of the closest RYR1s. This step-wise function determines the level of ORAI1

activation that can take four values depending on the amount of bound Ca^{2+}-free STIM. In the first phase after TCR/CD3 stimulation of T lymphocytes, preformed complexes of ORAI1, STIM1 and STIM2 have been detected by FRET experiments and super-resolution microscopy (Weiss and Diercks, unpublished results). We thus consider the activation of ORAI1 by heterotetramers of STIM1 and STIM2 (STIM1/2) and modified $f(C_S^{loc})$ accordingly. See **Supplementary Information** for a detailed explanation.

In the ER membrane, Ca^{2+} flux from the ER to the cytosol through the RYR is given by

$$J_{RyR} = \frac{I_{RyR}}{F \cdot z \cdot A_{RyR}} \cdot \frac{(C_s - C_C)}{(C_{s,0} - C_{C,0})} \quad (2)$$

with I_{RYR} the current through the RYR, which takes the value of 0.35 pA (Guo et al., 2012). A_{RYR} is the surface of the channel pore. The second factor in **Eq. (2)** allows to scale the current to take the actual gradient across the channel pore into account, where $C_{S,0}$ and $C_{C,0}$ represent resting concentrations of Ca^{2+} in the ER and in the cytosol (Mazel et al., 2009).

Finally, SERCA pumps are considered as bidirectional as in McIvor et al. (2018) and described by

$$J_{SERCA} = \frac{Q}{A_S} * V_{max} * \left[\frac{\left(\frac{C_C}{K_F}\right)^{n2} - \left(\frac{C_S}{K_R}\right)^{n2}}{1 + \left(\frac{C_C}{K_F}\right)^{n2} + \left(\frac{C_S}{K_R}\right)^{n2}} \right] \quad (3)$$

with V_{max} its maximal velocity, *n2* the Hill coefficient and K_F and K_R the pump affinity for cytosolic (C_C) and ER (C_S) calcium, respectively. A_s is the surface of the pore and Q is a temperature coefficient initially introduced by McIvor et al. (2018). To approximate the partial differential equations (PDE), we used the finite element method (FEM) and simulation software COMSOL Multiphysics 5.5 (http://www.comsol.com), more specifically the Transport of Diluted Species interface that is used to compute the concentration field of a dilute solute in a solvent. We chose a backward differentiation formula (BDF) to compute the time steps with a relative tolerance of 0.005 that controls the relative error in each step. The system is solved using the iterative linear solver GMRES (Generalized Minimum Residual). For further details regarding the system discretization and the use of COMSOL Multiphysics, please refer to the authors.

RESULTS

Ca^{2+} Microdomains Simulated by the Opening of Type 1 Ryanodine Receptors Localized in the ER-PM Junction do Not Rely on ORAI1 Opening

Upon TCR/CD3 stimulation, NAADP-evoked Ca^{2+} release through RYR1 acts in concert with Ca^{2+} entry through ORAI1/STIM complexes to create Ca^{2+} microdomains. These microdomains last for 64 ± 3 ms and reach amplitudes of 340 ± 11 nM (Diercks et al., 2018). Although it is known that ORAI1 and STIM are arranged in pre-formed complexes in the ER-PM junctions of T cells, the exact location of the RYR1s responsible for the decrease in $[Ca^{2+}]_{ER}$ in the sub-PM ER remains to be determined. As described in the presentation of the model and schematized in **Figure 1**, in the model we first considered that RYR1s are located in the junction, facing the PM as in cardiac dyadic clefts (Jones et al., 2018). In this section, we evaluated if this arrangement allows to reproduce experimental observations.

We simulated the junction schematized in **Figure 1** considering an increasing number of open RYR1s during 64 ms. A few milliseconds after RYR1 opening, a stable profile of Ca^{2+} increase in the junction is observed (**Figures 2A–F**, **Supplementary Figure S2**). Although C_c can locally reach concentrations close to 20 μM, the average Ca^{2+} concentration in the junction ranges from 340 to 2,500 nM depending on the number of open RYR1 (Anim. S1a,b in the **Supplementary Information**). Thus, in this configuration, opening of a single RYR1 allows to reach the experimentally observed microdomain amplitude in the junction. To assess the relative contributions of Ca^{2+} entry through ORAI1 and Ca^{2+} release from the ER through RYR1, we performed the same simulations in the absence of ORAI1 in the PM. As visible in **Figure 2H** in which Ca^{2+} microdomains with and without activated ORAI1 are seen to have nearly the same amplitude, the relative contribution of Ca^{2+} entry is very limited in these conditions. In agreement with this observation, the increase in the amplitude of the Ca^{2+} signal in the microdomain is not related to significant changes in the opening states of ORAI1 (**Figure 2I**). For example, when three RYR1s open simultaneously, all five ORAI1 channels are still in their lower state of activity as in the absence of any RYR1 opening. These computational observations indicate that the geometry depicted in **Figure 1** does not reflect the situation encountered in T-lymphocytes early after TCR activation, since the number of microdomains decreases significantly in $ORAI1^{-/-}$ T cells (Diercks et al., 2018).

Ca^{2+} Microdomains Observed Soon After T Cell Stimulation Are due to the Opening of Type 1 Ryanodine Receptors Localized on Conic ER Around the ER-PM Junction

Because of the small size of the junction, actual Ca^{2+} concentrations are expected to be highly sensitive to the ER-PM distance. Thus, results obtained in the previous section may depend on this junctional depth, which led us to investigate the influence of this distance on the Ca^{2+} profile in the junction. Data indicate that the ER-PM spacing is typically 10–20 nm (Hogan, 2015), but we investigated distances up to 50 nm that might be reached locally. As visible in **Figure 3**, although the amplitude of the Ca^{2+} microdomain is inversely proportional to the height of the junction, the experimental average is reached with less than three RYR1 simultaneously open even for the largest junction considered (50 nm). Moreover, the amplitudes are not much affected by the absence of ORAI1, as visible by the fact that the Ca^{2+} microdomain amplitudes without ORAI1 (dashed lines) are close to those with ORAI1 (plain lines). This does not agree with experimental observations showing that in cells that do not

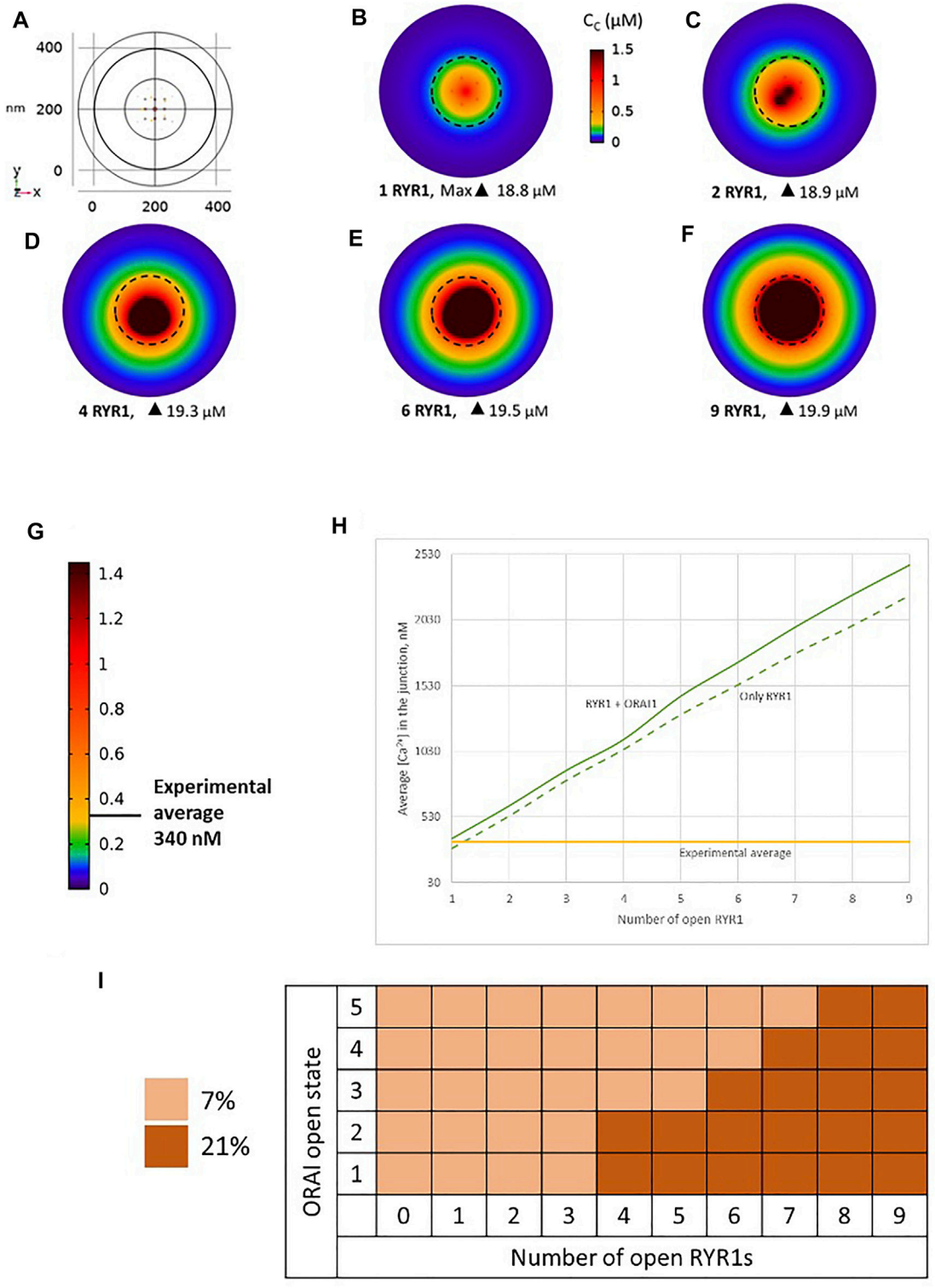


FIGURE 2 | Simulated Ca^{2+} microdomains resulting from the opening of RYR1 inside the junctions, which in turn induces the opening of ORAI1 channels in the junctions as a result of local depletion of ER Ca^{2+} **(A)** Upper view of the arrangement of the ORAI1 channels on the PM of the junction (yellow dots) and of the RYR1 on the ERM (red dots) using COMSOL **(B–F)** Steady-state Ca^{2+} profiles in the junction when opening 1, 2, 4, 6 and 9 RYR1 simultaneously **(B)** to **(F)** respectively. Shown are the profiles 22 ms after opening of the RYR1, but these stabilize very rapidly, after a few ms **(G)** Extended colour code with marking of the average amplitude of a microdomain in unstimulated T cells (Diercks et al., 2018) **(H)** Evolution of the amplitude of the simulated Ca^{2+} microdomains with the number of simultaneously open

(*Continued*)

FIGURE 2 | RYR1 in the junction, showing that experimentally observed microdomains do not agree with the opening of the RYRs inside the junction given the low contribution of the opening of the ORAI1 in conditions of a full ER (see text). Dotted line represents junctional Ca^{2+} concentration reached in the absence of ORAI1 channels **(I)** Individual evolution of 1–5 ORAI1 channels open state (Li et al., 2011) as a result of 0–9 RYR1 opening simultaneously. See **Supplementary Information** and Anim. S1a,b for details.

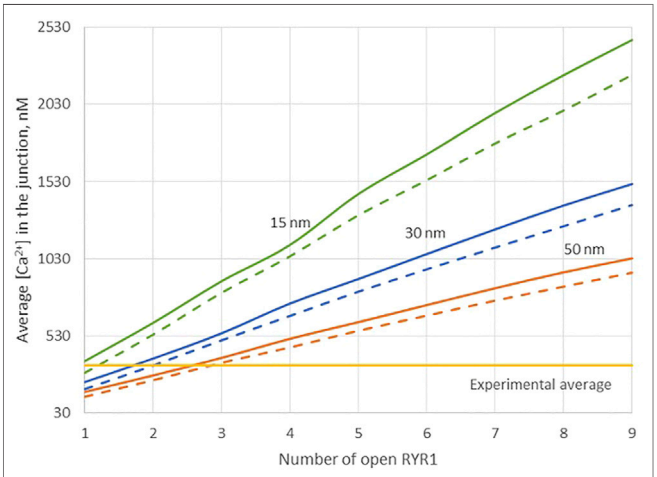


FIGURE 3 | Influence of the value of the distance between the PM and the ERM on the Ca^{2+} microdomains in the ER-PM junction. The green curve (15 nm) corresponds to the situation considered in **Figure 2**. Larger distances, blue curve (30 nm) and orange curve (50 nm) do not influence the low contribution of the opening of the ORAI1 to the Ca^{2+} concentration increase in the junction. Dotted lines represent junctional Ca^{2+} concentration reached in the absence of ORAI1 channels.

express ORAI1, the frequency of occurrence of junctional microdomains is reduced by ~10, while their amplitude is lowered by ~25% (Diercks et al., 2018). We thus concluded that RYR1 located inside the junction, whatever its height, cannot account for experimental observations.

Another possibility would be that RYR1s are located outside the junction, but close to it, in such a way that they affect sub-PM ER Ca^{2+} concentration. Thillaiappan et al. (2017) reported clusters of immobile IP_3Rs surrounding the ER-PM junctions, with the mouths of the IP_3Rs directed towards the PM. In our previous computational study of the IP_3R-dependent, adhesion-induced Ca^{2+} microdomains, we found that simulations based on this configuration agree with experimental observations (Gil et al., 2021). We investigated the possibility that RYR1s are similarly localized around the junction. In this configuration, schematized in **Figure 4**, a ring of RYR1s located in the sub-PM ER membrane and spaced by 90 nm, are releasing Ca^{2+} in the cytosolic space adjacent to the junction. The Ca^{2+} microdomains simulated under this configuration are shown in **Figure 5**, considering 1 (**Figure 5B**, **Supplementary Figure S2**) to 8 (**Figure 5I**) open RYR1. To reach the experimentally observed average amplitude of around 340 nM, seven or eight RYR1 must open simultaneously (**Figure 5J**). This number is slightly affected by the distance between the RYR1 and the junction. If a 45 nm distance is considered, instead of 90 nm as considered in **Figure 5**, opening of five RYR1s simultaneously is sufficient to reach the experimentally observed amplitude in the junction (**Supplementary Figure S4**). Indeed, a larger amount of the Ca^{2+} released by RYR1 can diffuse into the junction in this configuration. In the absence of ORAI1, the increase in Ca^{2+} in the junction due to RYR1 opening is much below the experimentally observed amplitude of the microdomains. Thus, when the RYR1 located in the membrane of the sub-PM ER are not releasing Ca^{2+} directly in the junctions, the model reproduces the experimental observation that NAADP-induced Ca^{2+} microdomains rely on both RYR1 and ORAI1 (Diercks et al., 2018).

Observations in T cells indicate that Ca^{2+} signals in the junction are rather stereotypic, with a relatively constant amplitude (Diercks et al., 2018). It is thus expected that the microdomain characteristics are not very sensitive to the numbers of RYR1 present around the junction. We next investigated the influence of the number of open RYR1 in more detail, considering the possibility that up to 16 RYR1 are located around the junction. This was done in the simulations by considering another ring of eight receptors 90 nm below the first one. As visible in **Figure 6A**, the relation between the amplitude of the simulated microdomains and the number of open RYR1 is non-linear with a marked stepwise behaviour. From one to three open RYR1, the increase in amplitude is linear. All five STIM1/2 bound ORAI1 channels are in their lowest conductance state (**Figure 6B**) and the Ca^{2+} increase in the junction is due to diffusion from the adjacent cytosol. From four open RYR1 on, local ER Ca^{2+} depletion is sufficient to further activate ORAI1 creating changes in the slope of the relation between the amplitude of the Ca^{2+} signal and the number of open receptors. If more than eight RYR1 open simultaneously, additional ones do not activate ORAI1 further. As seen in **Supplementary Figure S1**, the 54% opening state is reached when the value of C_S in the close vicinity of STIM1/2 bound to ORAI1 reaches 260 μM. This would require a Ca^{2+} decrease at the ER lumen close to RYR1 channel that is not reached under localised Ca^{2+} signaling because of fast diffusion-mediated replenishment. From 8 to 16 open RYR1, local depletion is not much affected, with a minimal ER Ca^{2+} concentration that remains around 340 μM as visible in **Figure 7** that shows a cross-sectional view of the Ca^{2+} concentrations in the sub-PM ER and in the cytoplasmic space including the junction (see also **Supplementary Figure S3** for cross-sections). Noticeably, Ca^{2+} concentrations at the cytosolic side of RYR1 slightly decrease with the number of open receptors, from 17.5 to 15.9 μM for 2 and 16 open receptors, respectively. The slight decrease in the Ca^{2+} gradient at the channel pore indeed reduces the flux through the RYR1. Thus, increasing the number of RYR1 leads to a decrease in the ER Ca^{2+} concentration around STIM1/2, but this decrease is not sufficient to provoke the passage of ORAI1 channels to a higher conductance state.

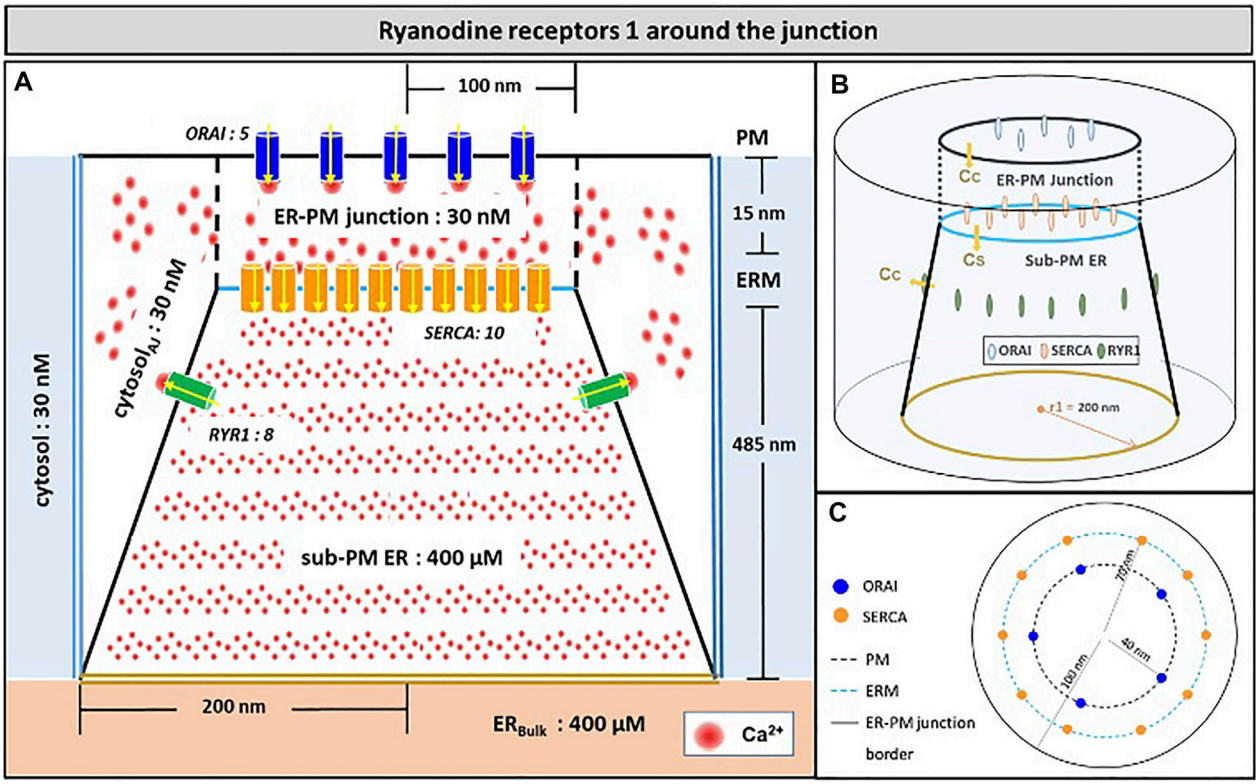


FIGURE 4 | Schematic representation of the model geometry of the ER-PM junction and sub-PM ER used to investigate the origin of the Ca^{2+} microdomains in T cells with eight RYR1 around the junction **(A)** Frontal diagram showing the dimensions of the cone that represents the sub-PM ER, of the junction and of the portion of the cytosol considered in the simulations. ORAI1 channels are in blue, SERCA pumps in orange and RYR1 in green. Plain lines represent membrane boundaries; dashed lines, fictitious limits between the junction and the cytosol and double lines indicate the limits of the simulated system. The resting Ca^{2+} concentrations considered as initial conditions and boundary conditions in the two compartments are indicated. **(B)** 3D view of the model geometry **(C)** Upper view of the positions of the ORAI1 channels on the PM, in blue, and of the SERCA pumps and RYR1 on the ERM, in orange and green respectively. Not to scale. This geometry is based on McIvor et al. (2018). See text for details.

$NAADP$-induced Ca^{2+} microdomains observed early after T cell stimulation thus rely on the simultaneous opening of an average of seven–eight RYR1 located around the ER-PM junction.

Predicted Type 1 Ryanodine Receptors Involvement in the Formation of Microdomains Agree With Type 1 Ryanodine Receptors Open Probabilities in the Presence of NAADP

Simulation results obtained in the previous section indicate that best agreement between modelled and experimentally observed microdomains occur when most of the eight RYR1 located near the preformed ORAI1/STIM1/STIM2 complexes are open simultaneously during 64 ms. This conclusion stems from a direct comparison between the simulated and experimentally observed Ca^{2+} signals. To further validate this result, some reasoning based on RYR1 open probability can be proposed. The 64 ms duration of a Ca^{2+} microdomain corresponds to the average duration of the NAADP-evoked Ca^{2+} signals arising in the first 15 s after TCR activation (Diercks et al., 2018). Interestingly, 64 ms also fits in the range of the reported durations of Ca^{2+} sparks (Jaggar et al., 2000). It is thus likely that in response to the TCR/CD3 stimulation-induced NAADP increase, RYR1 undergo repetitive openings maintained by Ca^{2+}-induced Ca^{2+}-release, a process that generates a small amplitude Ca^{2+} increase in the cytosol, called "spark". The decrease of ER Ca^{2+} that accompanies the spark is in turn responsible for the opening of ORAI1, and thus for the Ca^{2+} microdomain in the junction.

We thus investigated if our conclusions about the molecular mechanism underlying TCR/CD3-induced Ca^{2+} microdomains are compatible with the dynamics of RYR1 during spark-like activity. Upon TCR/CD3 stimulation, global NAADP concentration in T cells increases from 4.1 ± 1.5 nM to 33.6 ± 7.2 nM (Gasser et al., 2006). Because RYR1 are activated by NAADP in T cells (Wolf et al., 2015; Diercks et al., 2018; Roggenkamp et al., 2021), their Ca^{2+}-releasing activity will increase. Indeed, Hohenegger et al. (2002) have shown that the open probability of these receptors is a highly nonlinear function of NAADP concentration, with an EC_{50} of 31.2 ± 6.9 nM. Because NAADP synthesis occurs near the ER-PM junctions (Gu

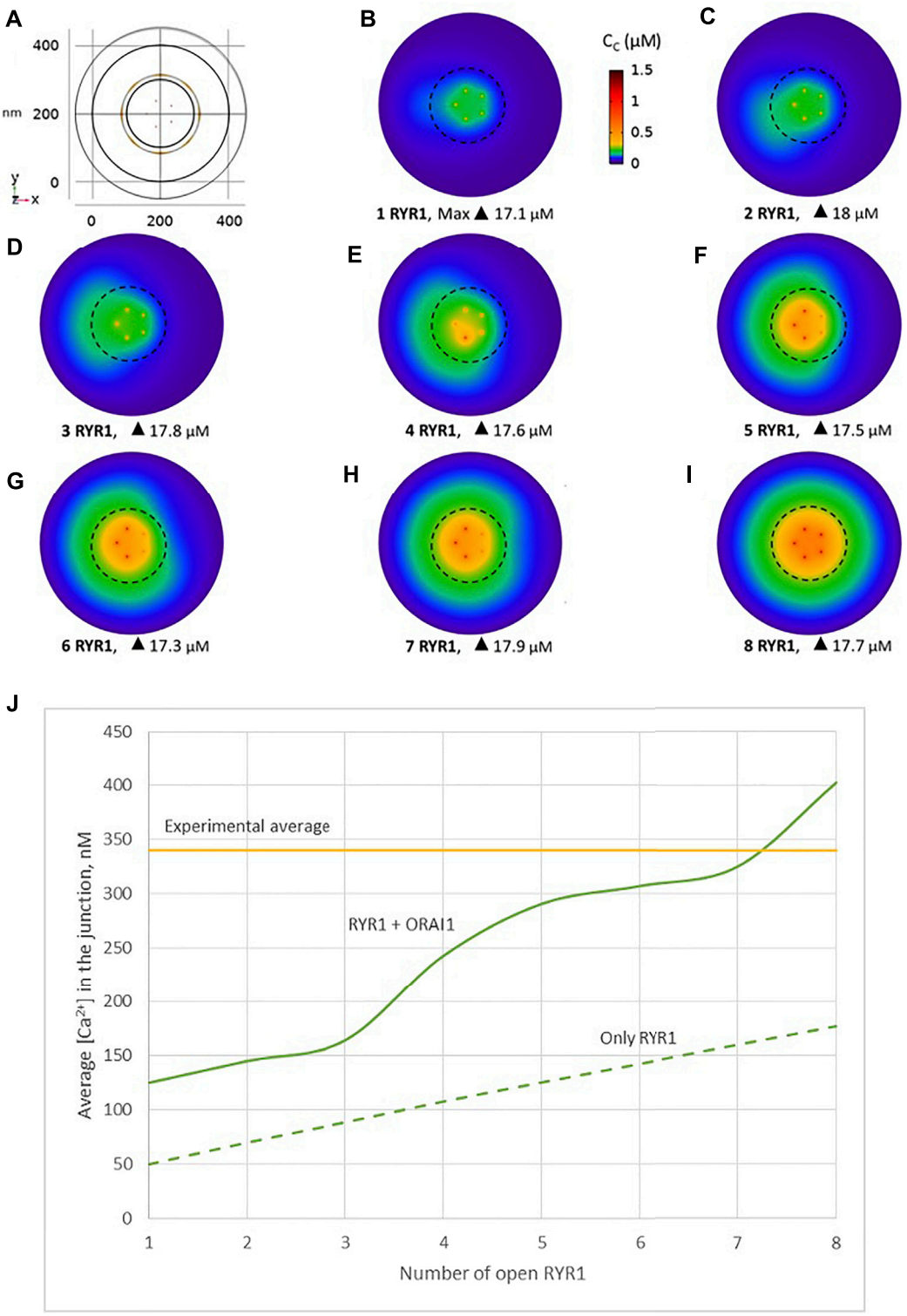


FIGURE 5 | Simulated Ca^{2+} microdomains resulting from the opening of the RYR1 adjacent to the junctions, which in turn induces the opening of ORAI1 channels in the junctions as a result of local depletion of ER Ca^{2+} **(A)** Upper view of the arrangement of the ORAI1 channels on the PM of the junction (red dots) and of the adjacent RYR1 (yellow lines) using COMSOL **(B–I)** Steady-state Ca^{2+} profiles in the junction when opening 1 **(B)** to 8 **(I)** RYR1 simultaneously. Shown are the profiles 22 ms after opening of the RYR1. Upon depletion of local Ca^{2+} in the ER, which is quasi-instantaneous, ORAI1 channels open to an extent that depends on this local concentration, as defined by the function $f_{2/1}$ (see **Supplementary Information**). ORAI1 opening is assumed to occur immediately after depletion because ORAI1-STIM1/2 aggregates are pre-formed (Weiss and Diercks, unpublished results). **(J)** Evolution of the amplitude of the simulated Ca^{2+} microdomains with the number of

(*Continued*)

FIGURE 5 | simultaneously open RYR1 in the junction, showing that experimentally observed microdomains can in principle result from the opening of ORAI1 channels induced by the spontaneous opening of a few RYR1 near the junction, in conditions of a full ER. Dotted line represents junctional Ca^{2+} concentration reached in the absence of ORAI1 channels. See Anim. S2a,b,c,d and S3 in **Supplementary Information**.

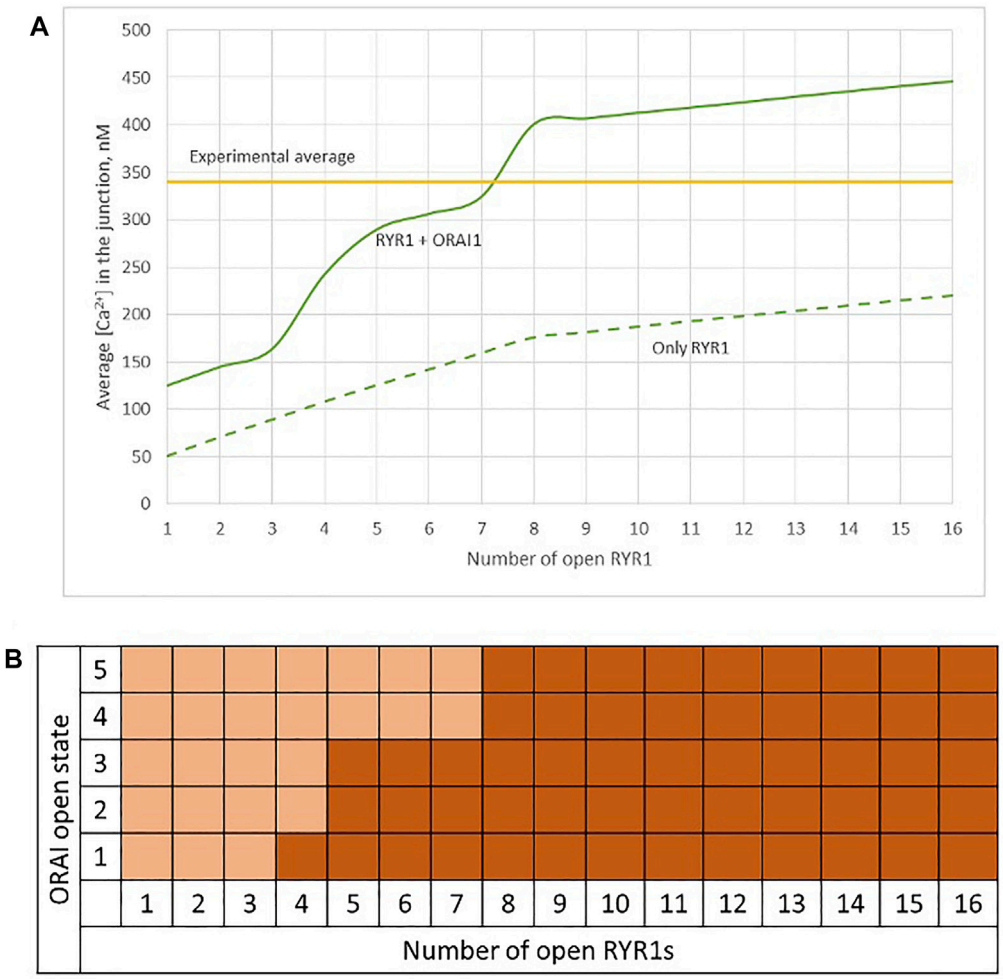


FIGURE 6 | Simulated Ca^{2+} microdomains resulting from the opening of up to 16 RYR1 adjacent to the junctions **(A)** Evolution of the amplitude of the simulated Ca^{2+} microdomains in the presence (green curve) and in the absence of ORAI1s (dotted green curve) in the junction. Both green curves, up to eight simultaneously open RYR1, correspond to the situation considered in **Figure 5**. The theoretical situation of a junction that does not contain ORAI1 channels (dotted green curve) allows to appreciate that the contribution of Ca^{2+} released through the RYR1 to the Ca^{2+} microdomain is linear and rather limited. At eight simultaneously open RYR1 (green curve), the complete cluster of five ORAI1 channels reach their maximum open state possible, in conditions of a full ER **(B)** Individual evolution of 1–5 ORAI1 channels open state (Li et al., 2011) as a result of 1–16 RYR1 opening simultaneously. See **Supplementary Information** for details.

et al., 2021), local concentrations in the vicinity of RYR1 are certainly larger than the average values mentioned above and likely exceed the EC_{50}. Thus, the open probability of RYR1 near the junctions must be of the order of 0.7, which is the maximal value measured at 20 μM Ca^{2+}. Given that the mean open time of RYR1 is ~2 ms (des Georges et al., 2016; Sato & Bers, 2011), the mean closed time in these conditions can be estimated to be 0.86 ms. On the basis of these data, one can obtain a rough approximation of the number of simultaneously open receptors in a spark site when RYR1s are maximally activated by NAADP. Straightforward stochastic simulations of opening and closing of eight RYR1 with average opening and closing times equal to 2 and 0.86 ms, respectively, indicate that most of the time, six receptors are simultaneously open (**Figure 8A**). This result was obtained by performing 64 ms long stochastic simulations of eight independent RYR1 and determining at each time step how many receptors are open. The maximal frequency at six open receptors is in accordance with the three-dimensional spatio-temporal simulations indicating that best agreement between

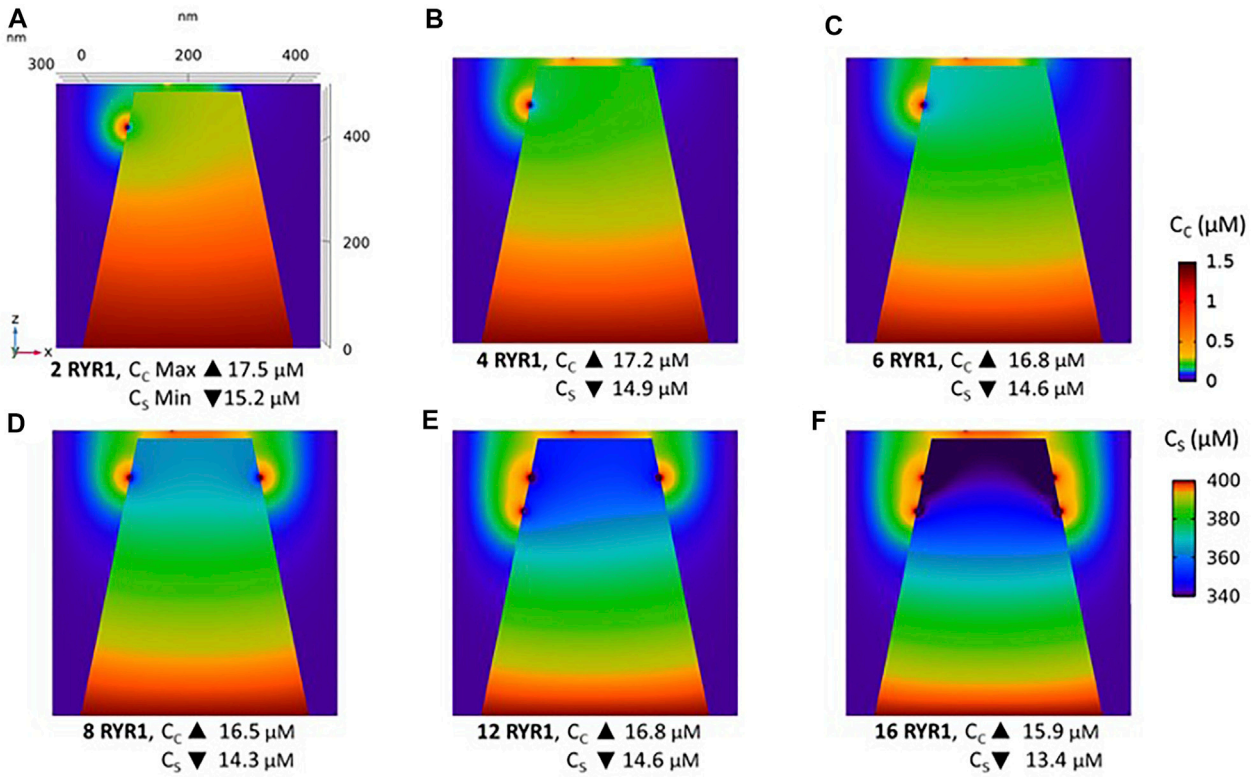


FIGURE 7 | Cross-section of the Ca^{2+} profiles in the junction, in the cytosol adjacent to the junction and in the sub-PM ER during microdomain formation **(A–D)** microdomains created by the opening of 2, 4, 6 and 8 RYR1. Local depletion of ER Ca^{2+} provokes the opening of the nearby ORAI1s. This situation corresponds to the one shown in **Figure 5**. **(E,F)** microdomains created by the opening of 12 and 16 RYR1, respectively. The second cluster of eight RYR1 is located directly underneath the first cluster. Local depletion of ER Ca^{2+} is not enough to provoke additional opening of the nearby ORAI1s. This situation corresponds to the one shown in **Figure 6**. For all panels, the upper right bar indicates the colour code of Ca^{2+} concentration in the cytosol while the lower right bar indicates the colour code of Ca^{2+} concentration in the ER. See Anim. S4.

experimental observations and computational results is obtained when seven to eight RYR1 are simultaneously open during a 64 ms Ca^{2+} spark, considering the existence of pre-formed STIM1/2 and ORAI1 complexes (**Figure 5**). In contrast, the same calculations predict that for basal NAADP concentrations, when the open probability of RYR1 equals 0.4, the highest frequency corresponds to three RYR1 open simultaneously (**Figure 8B**). In the above simulations (**Figure 5**), this corresponds to a Ca^{2+} microdomain with an amplitude well below the experimental average. Similar stochastic simulations (**Supplementary Figure S5** and related text in the SI) indicate that for the IP_3-mediated microdomains corresponding to pre-stimulation conditions, the maximal frequency corresponds to two receptors open simultaneously, in qualitative agreement with our previous results (Gil et al., 2021).

In summary, by combining previous observations about NAADP increase upon TCR/CD3 stimulation of T cells and RYR1 open probability, we conclude that the NAADP-induced increase in the open probability of RYR1 triggers the opening of most of the eight RYR1 located around the ER-PM junction. This is in agreement with the spatio-temporal simulations shown in the previous section, which indicate that Ca^{2+} microdomains observed early after T cell stimulations rely on the simultaneous opening of seven to eight RYR1 located around the junction.

The Isoforms of STIM That Are Bound to ORAI1 Determine the Characteristics of the Ca^{2+} Microdomains

In the previous sections, we found that the microdomains occurring during the first 15 s after TCR/CD3 stimulation involve a larger number of ER Ca^{2+} releasing channels than those observed before stimulation. Indeed, the local depletion induced by the opening of six–seven RYR1 is needed to activate ORAI1 and reproduce experimentally observed NAADP-dependent microdomains (**Figure 5**) while three to six IP_3R are involved in the creation of the adhesion-mediated, IP_3-dependent microdomains (Gil et al., 2021). At first sight, this is paradoxical as the conductance of RYR1 is about 5 times larger than that of IP_3Rs.

We thus studied the characteristics of the Ca^{2+} increases created by the simultaneous opening of either eight IP_3Rs or eight RYR1 (see **Table 1**). For the two Ca^{2+} channel types, we considered two possible preformed ORAI1/STIM complexes:

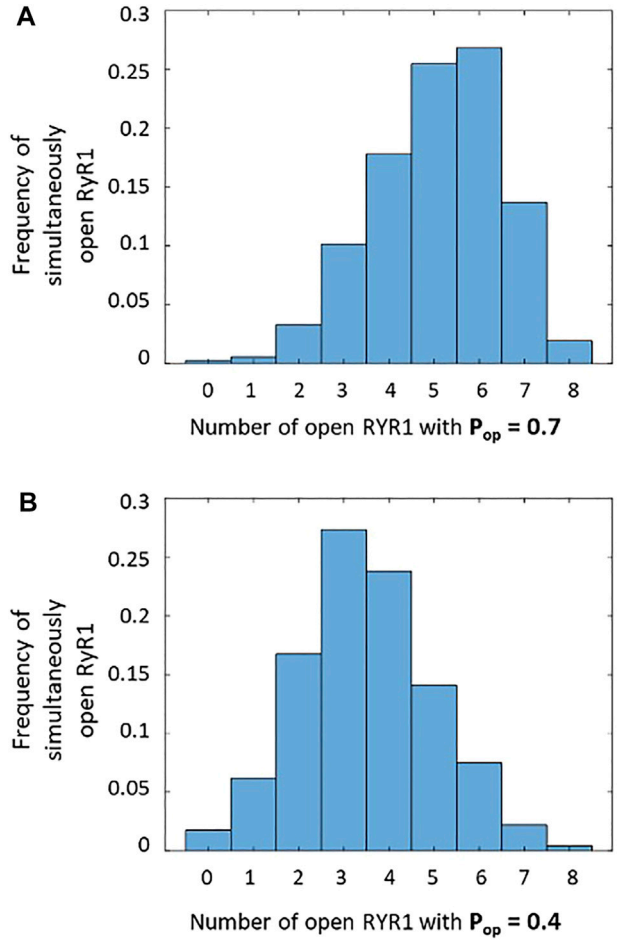


FIGURE 8 | Frequency of simultaneously open receptors in a cluster of eight RYR1 during 64 ms that corresponds to the duration of a Ca^{2+} spark **(A)** During a spark when RYR1 are maximally activated by NAADP, there are most of the time around five to six simultaneously open RYR1, with a single receptor open probability of 0.7 (Hohenegger et al., 2002). This is in agreement with the results seen in **Figure 5**. **(B)** At basal NAADP concentration, the single receptor open probability is around 0.4, and there are continuously around three to four simultaneously open RYR1, which is not enough to reach the experimental Ca^{2+} amplitude. In the two panels, the stochastic simulations of opening and closing of individual receptors are performed during 64 ms.

STIM2 homotetramers (STIM2/2) and STIM1 and STIM2 heterotetramers (STIM1/2). Simulation results indicate that the nature of the ER Ca^{2+} release channel does not much influence the characteristics of the Ca^{2+} microdomains. Indeed, the Ca^{2+} signal in the junction is determined by the opening state of ORAI1, which is the same when Ca^{2+} release from the ER is mediated by IP_3Rs or RYR1. The Ca^{2+} concentration sensed by the ORAI1/STIM complex is nearly the same in the two situations. As shown in **Figure 9C**, once the steady state gradients are established, the fluxes are nearly identical. Because of the slow replenishment around the pore of the receptor channel with D_S = 10 $\mu m^2/s$, the concentration gradient around the two extremities of the pore does not changes drastically and hence the flux remains of the same order for IP_3R and RYR1. In contrast, the nature of the STIM isoforms bound to ORAI1 has a drastic influence on the characteristics of the Ca^{2+} microdomain since it determines their Ca^{2+} sensitivity, and hence the opening state of ORAI1. Because STIM1/2 has a lower sensitivity to ER Ca^{2+} depletion than STIM2/2, the open state of ORAI1 is lower and the increase in Ca^{2+} in the junction has both a smaller amplitude and spatial extent (compare blue to green lines in **Figure 9A,B** to **Figure 6B** and **Table 1**).

Based on these computational observations, the prototypical evolution of Ca^{2+} microdomains from adherent to TCR/CD3 stimulated T cells is proposed to obey the following scenario. Upon adhesion to proteins of the extracellular matrix, integrin evoked IP_3 signaling provokes an increase in the frequency of Ca^{2+} puffs arising from the clusters of IP_3R located near the ER/PM junction. These puffs typically last ~44 ms during which, in average, two IP_3Rs are open simultaneously. On the other hand, five ORAI1 channels are located in the PM of the junction and bound to STIM2/2 homotetramers. In response to the decrease in ER Ca^{2+} created by the puff, the ORAI1 channels open to ~21% of their maximal conductance and create the Ca^{2+} microdomain (**Figures 10A**, 1st row of **Table 2**). After the Ca^{2+} puff, Ca^{2+} is rapidly replenished in the sub-PM ER and ORAI1 channels shift to their lowest conductance state (~7% opening) that corresponds to basal Ca^{2+} entry, but not to a detectable Ca^{2+} microdomain (**Figure 10B**, 2nd row of **Table 2**) when bound to the STIM1/2 isoform, and stay at 21% when bound to STIM2/2. TCR/CD3 stimulation initiates NAADP signaling, which increases the open probability of the RYR1 located around the junction and thus the frequency of Ca^{2+} sparks (**Figure 10C**, 3rd row of **Table 2**). During these events, ~6 RYR1 are open simultaneously. At this stage, ORAI1 channels are preferably bound to STIM1/2 heterotetramers, which decreases their sensitivity to ER Ca^{2+} depletion. Thus, as in the case of the IP_3-dependent microdomains, they open at ~21% of their full capacity. However, because more Ca^{2+} is released in the cytosolic space just around the junction by six RYR1 than by two IP_3R, the Ca^{2+} microdomain in the junction is a bit larger because of diffusion.

DISCUSSION

Activation of T cells is an essential step to start an adaptive immune response. At this particular point a highly important decision is made: whether a T cell stays quiescent or may develop into an effector T cell carrying out immune effector functions to destroy pathogens, or in case of autoimmune reactions, to attack our own body. Among several signaling processes involved, Ca^{2+} signaling is fundamental for T cell activation. In a previous study, we resorted to mathematical modeling to gain insight into the early, small scale Ca^{2+} increases that follow adhesive interactions of T cells, which play a crucial role in T cell migration to inflamed tissue (Weiss

TABLE 1 | Characteristics of the simulated microdomains relying on the simultaneous opening of eight IP_3Rs for 44 ms (first two lines) or of eight RYR1 for 64 ms (lines three and 4). For each case, two situations are considered: the existence of pre-formed clusters of ORAI1 with STIM2/2 homotetramers (lines 1 and 3) and the existence of pre-formed clusters of ORAI1 with STIM2/1 heterotetramers (lines two and 4). Maximal concentrations denote the maximal local concentrations reached in the domains indicated. Spatial extent refers to the area of the junction's portion in which Ca^{2+} concentration exceeds 300 μM. The ER $[Ca^{2+}]$ felt by ORAI is the average local concentration of luminal Ca^{2+} around the mouth of the IP_3R or RYR1, computed in a 108 nm^3 volume.

Eight Open Receptors		Max$[Ca^{2+}]$ in the junction(μM)	Max$[Ca^{2+}]$ around the junction(μM)	Average$[Ca^{2+}]$ in the junction(μM)	Spatial extent (μM^2)	ER$[Ca^{2+}]$ at channel pore(μM)	ER$[Ca^{2+}]$ felt by ORAI(μM)	ORAI mean open state
IP_3R 44 ms	STIM2/2	10	17	0.755	0.049	11	331	54%
	STIM2/1	354	17	0.396	0.024	11	331	21%
RyR 165 ms	STIM2/2	11	18	0.762	0.049	12	328	54%
	STIM2/1	4.57	18	0.402	0.024	12	328	21%

and Diercks, unpublished results; Mezu-Ndubuisi & Maheshwari, 2021). Here, we extend this model to investigate the molecular mechanism underlying the early phase of activation following TCR/CD3 stimulation. Although both types of Ca^{2+} microdomains largely rely on Ca^{2+} entry through preformed ORAI1/STIM complexes, the former response relies on IP_3 signaling while the latter involves NAADP and RYRs. Interestingly, a progressive change in STIM isoforms, from mainly STIM2 homotetramers to STIM1/2 heterotetramers, also accompanies this transition (**Figure 11**). Computational modeling of the spatio-temporal Ca^{2+} dynamics in the ER-PM junctions allowed us to reproduce the observation that the microdomains triggered by cell adhesion or by TCR/CD3 stimulation at its early phase appear rather similar, despite the different underlying mechanisms.

Simulations of the NAADP dependent microdomains quite forwardly predict that RYR1 are most probably located outside the junction, in a region of the ER membrane that is directly adjacent to the junction. This view contrasts with the well-known arrangement of RYR2 in cardiac cells, where they are facing the PM in dyadic clefts (Jones et al., 2018). In principle, a few RYR1 could be located in the ER-PM junction of T cells because the dimensions of its cytoplasmic part (approx. 28 nm × 28 nm × 12 nm, see Lanner et al., 2010) do not exceed the dimensions of the junction. However, Ca^{2+} microdomains simulated with such a spatial arrangement are no longer dependent on ORAI1, since the increase in Ca^{2+} concentration due to the influx mediated by a single RYR into the junction suffices to create a Ca^{2+} signal of the amplitude observed experimentally, which does not agree with experimental results. It is thus most probable that in T cells RYR1 are arranged around the junction, in the same way as IP_3R (Thillaiappan et al., 2017; Taylor and Machaca, 2019).

Computational results indicate that the influx through ORAI1 channels much depends on the STIM isoforms to which it is bound. In the case of local signaling investigated here, the same opening state of ORAI1 is reached after local Ca^{2+} depletion induced by three IP_3R when it is bound to STIM2/2 as after local Ca^{2+} depletion induced by seven RYR1 when it is bound to STIM1/2. Thus, the change in the nature of the ORAI1/STIM complexes that follow TCR/CD3 stimulation is expected to play a crucial role in maintaining Ca^{2+} signaling localized despite the stimulation of RYR by NAADP. Along this line, Ahmad et al. (2021) have recently shown in HEK293 cells that while clusters of STIM2 represent the sites of SOCE initiation, STIM1 molecules are progressively recruited when cells are exposed to low stimulation.

In contrast, the nature of the ER-releasing Ca^{2+} channel that creates the local depletion in the sub-PM ER, subsequently triggering ORAI1 opening does not have a significant effect. This result is *a priori* surprising given that the RYR1 has a conductance ~5 times larger than the IP_3R. Simulations indicate that the flux is limited by the replenishment of ER Ca^{2+} at the mouth of the channel rather than by its conductance. Thus, the extent of local depletion is imposed by the value of the diffusion coefficient of Ca^{2+} in the ER. This computational observation agrees with the major role played by intraluminal diffusion of Ca^{2+} in setting the responsiveness of Purkinje cells to synaptic inputs (Okubo et al., 2015). It should be kept in mind that the peculiar geometry of the junctional ER is expected to play an important role in decreasing the value of the effective Ca^{2+} diffusion coefficient because of the tortuosity of the tubular network of the ER (Schaff et al., 1997; Oloveczky and Verkman, 1998). In our simulations, diffusion is however fast enough to avoid decreases in local ER Ca^{2+} that would trigger the passage of ORAI1 channels in a highly active state. Simulations indicate that during localized Ca^{2+} signaling in T cells, ORAI1 channels never exceed 21% of their maximal activity.

Together with experimental observations (Diercks et al., 2018; Weiss and Diercks, unpublished results), our computational model ascribes the evolution of Ca^{2+} microdomains from the adherent/pre-stimulated to the TCR/CD3 early stimulated state as a passage from puff-to spark-triggered SOCE. Indeed, while the two types of localized Ca^{2+} signaling rely on ORAI1-mediated Ca^{2+} entry, cell adhesion triggers the synthesis of IP_3, and TCR/CD3

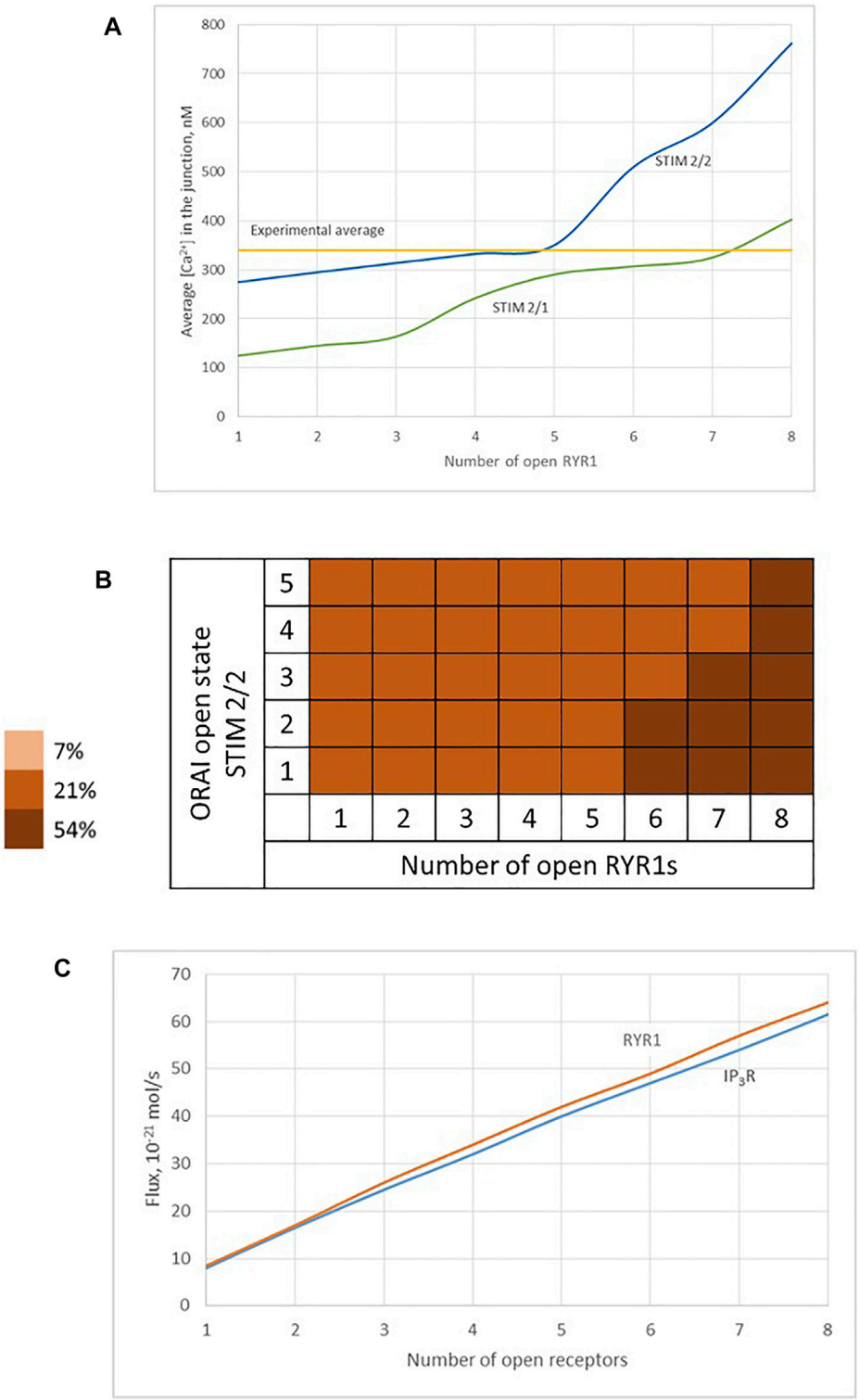


FIGURE 9 | Influence of the nature of the ER Ca^{2+} channel inducing the local depletion of ER Ca^{2+} (IP_3R or RYR1) and of the STIM isoforms bound to ORAI1 on the characteristics of Ca^{2+} microdomains **(A)** Evolution of Ca^{2+} microdomains amplitude with the number of simultaneously open RYR1 in the junction. The microdomains observed in conditions of full ER can be induced by the spontaneous opening of a few RYR1 near the junction that in turn trigger the opening of ORAI1 channels bound to STIM2/2 (blue curve) or to STIM1/2 (green curve). ORAI1 channels open to an extent that depends on local ER Ca^{2+} concentration, as defined by the corresponding function f_2 or $f_{2/1}$ (see **Supplementary Information**) **(B)** Individual evolution between the 5 open states of the ORAI1 (Li et al., 2011), as a result of 1–8 RYR1 opening simultaneously. **(C)** Comparison of Ca^{2+} fluxes through open IP_3Rs and RYR1. Because of the slow replenishment around the pore of the receptor channel with D_S = 10 $\mu m^2/s$, the concentration gradient around the two extremities of the pore does not changes drastically and hence the flux remains of the same order for IP_3Rs and RYR1.

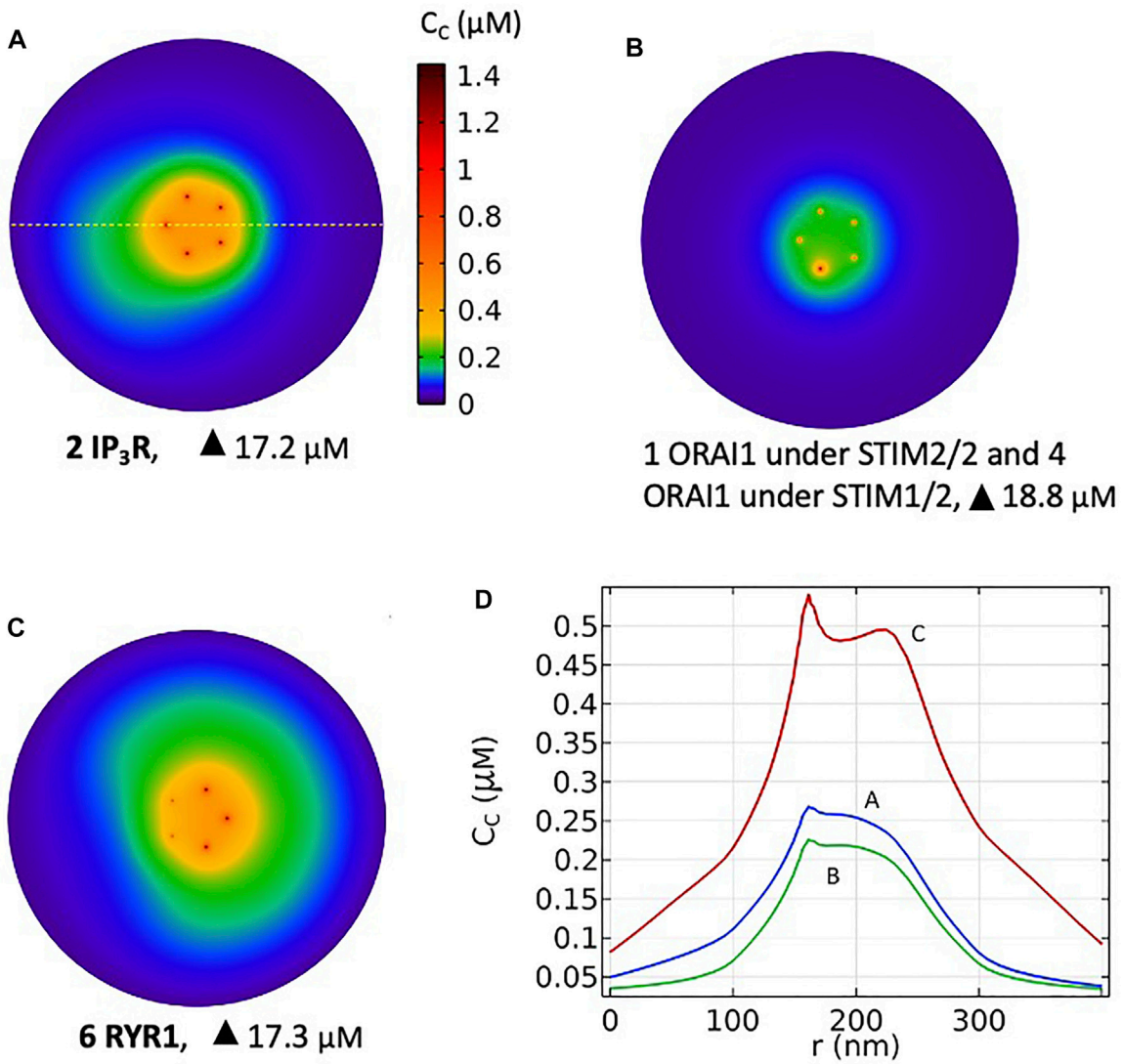


FIGURE 10 | Simulated most probable Ca^{2+} microdomains resulting from a T cell transition between quiescent to early activation. **(A)** Non TCR/CD3-dependent Ca^{2+} microdomains formed by the opening of two IP_3Rs adjacent to the junction and further opening of ORAI1 channels bound to STIM2/2 **(B)** Basal opening of five ORAI1 channels, one inherently co-localized with STIM2/2 and four inherently co-localized with STIM1/2 leading to small microdomains arising from nano-scale [Ca^{2+}] fluctuations in the sub-PM ER. Artificial construction **(C)** TCR/CD3-dependent Ca^{2+} microdomains formed by the opening of six RYR1 adjacent to the junction and further opening of ORAI1 channels bound to STIM1/2. See Anim. S5a,b,c in the **Supplementary Information**. **(D)** Ca^{2+} profiles along the yellow dotted line shown in A that traverses the middle of the junction at a 7.5 nm distance from the PM, corresponding to panels A, B and C.

TABLE 2 | Characteristics of the simulated microdomains corresponding to IP_3-dependent Ca^{2+} signaling stimulated by T cell adhesion (line 1), to a spontaneous opening of ORAI in the absence of stimulation (line 2) or to NAADP-dependent Ca^{2+} signaling in response to TCR/CD3 stimulation (line 3)

—	Average[Ca^{2+}] in the junction(μM)	Spatial extent (μm^2)	ER[Ca^{2+}] felt by ORAI(μM)	ORAI mean open state
Puff,2 IP_3R; STIM2/2	0.293	0.0078	350	21%
ORAI,STIM2/2 to STIM2/1	0.236	—	400	7%
Spark,6 RYR1; STIM2/1	0.307	0.0123	335	21%

stimulation initially that of NAADP. The respective durations of the two types of Ca^{2+} microdomains (44 ± 4 ms and 64 ± 3 ms) are in the ranges of those reported for puffs (Bootman et al., 1997) and sparks (Jaggar et al., 2000), respectively. In the two cases, the local depletion in ER Ca^{2+} created by the puff or the spark can trigger ORAI1 opening, with a resulting simulated Ca^{2+} increase in the ER-PM junction that matches experimental observations. Moreover, straightforward

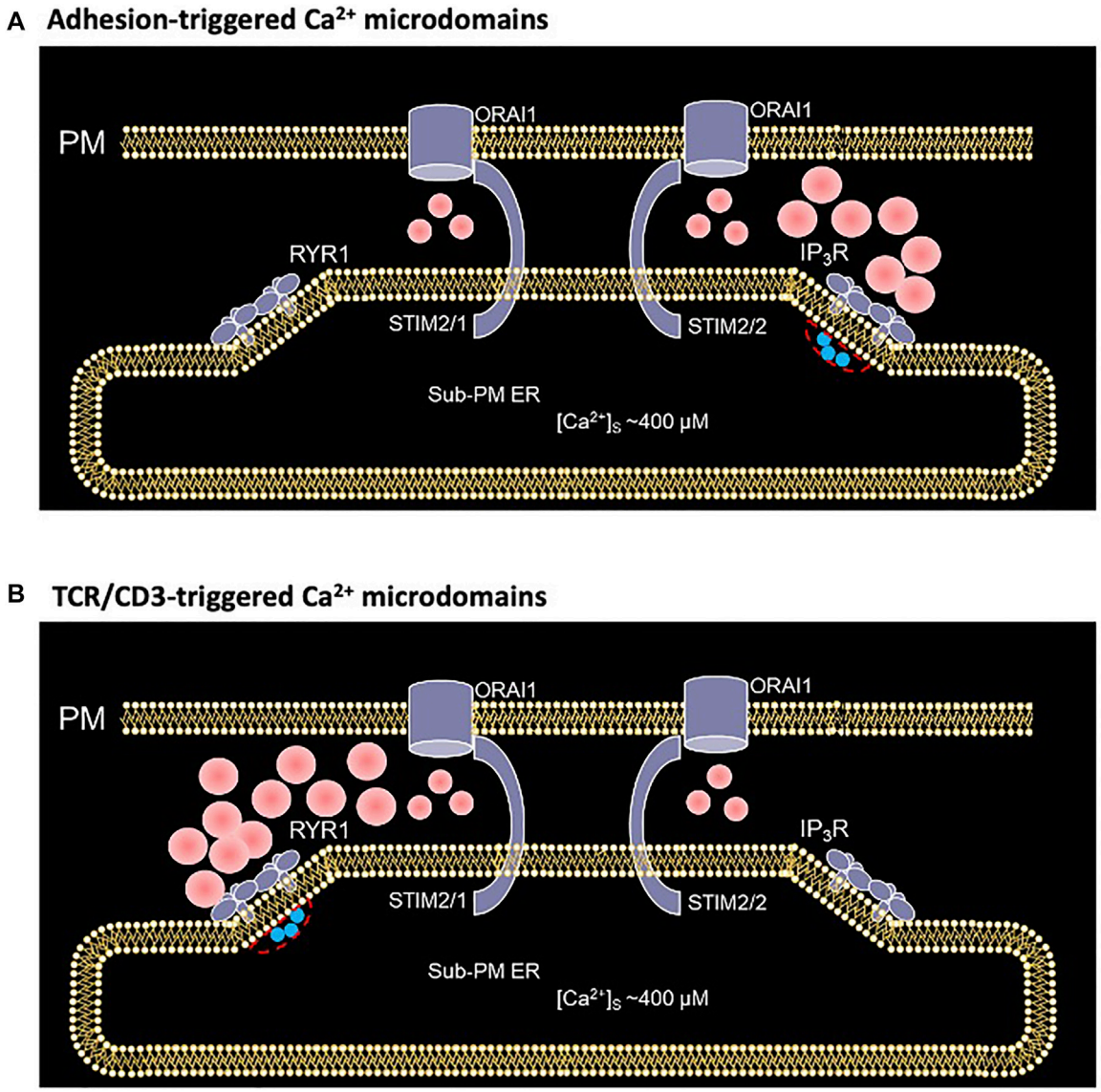


FIGURE 11 | Schematized representation of the proposed mechanism underlying the spontaneous formation of Ca^{2+} microdomains in T cells during its transition from quiescent to early activation **(A)** In an otherwise unstimulated cell, non TCR/CD3 dependent short, spontaneous activation of one or a few IP_3Rs close to the junction, releases Ca^{2+} from the sub-PM ER into the cytosol, leading to further opening of ORAI1 channels most likely bound to STIM2/2 at this stage (Weiss and Diercks, unpublished results) **(B)** During the first 15 s following TCR/CD3 stimulation, and NAADP driven activation of several RYR1 close to the junction, slightly larger amount of Ca^{2+} is released from the sub-PM ER into the cytosol. The resulting local Ca^{2+} depletion close to the RYR1 pore provokes the unbinding of Ca^{2+} from STIM1/2 heterotetramers, which further activates ORAI1 channels (Diercks et al., 2018). Red spots represent Ca^{2+} ions.

stochastic simulations of channel opening and closing taking into account the channels open probabilities in the presence of ligand and high Ca^{2+} concentration indicate a number of simultaneously open receptors matching with the results of the 3D spatiotemporal simulations. In the future, more realistic 3D simulations should be performed to take into account the stochastic nature of puffs and sparks to simulate microdomains, instead of the simplified deterministic description of stereotypic IP_3R- or RYR1-mediated release of ER Ca^{2+} used in the present study.

As another perspective, the present model could be used to investigate how the increase in frequency of RYR1 opening observed ~15 s after TCR/CD3 stimulation affects the spatiotemporal dynamics of junctional Ca^{2+} during the transition of T cells towards full activation. In these longer time scales, additional aspects of SOCE regulation should be considered, such as slow Ca^{2+} dependent inactivation (Dagan and Palty, 2021) or the dynamic nature of the ER-PM junctions (Okeke et al., 2016). The extension of the model to several junctions would enable to investigate how

microdomains spread and interact to propagate Ca^{2+} signals deeper into the cell and promote full activation.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

DG, B-PD, AG and GD contributed to the conceptualization and design of the study. DG, AG and GD contributed to the development and analysis of the model. DG performed all numerical simulations. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) (project number 335447717; SFB1328, project A01 to AG), by the Joachim-Herz-Stiftung (Hamburg), Infectophysics Consortium (project 4; to AG), by NCL-Stiftung Hamburg (to AG), the Hamburg Ministry of Science, Research and Equality (LFF-FV75/0070-134, to AG), and by University Medical Center Hamburg-Eppendorf (M3I consortium, to AG). GD is Research Director at the FNRS.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.811145/full#supplementary-material

REFERENCES

Ahmad, M., Ong, H. L., Saadi, H., Son, G. Y., Shokatian, Z., Terry, L. E., et al. (2021). Functional Communication between IP_3R and STIM2 at Sub-threshold Stimuli Is a Critical Checkpoint for Initiation of SOCE. *Proc. Natl. Acad. Sci. U S A* 119 (3), e2114928118. doi:10.1073/pnas.2114928118

Bootman, M., Niggli, E., Berridge, M., and Lipp, P. (1997). Imaging the Hierarchical Ca2+ Signalling System in HeLa Cells. *J. Physiol.* 499, 307–314. doi:10.1113/jphysiol.1997.sp021928

Brandman, O., Liou, J., Park, W. S., and Meyer, T. (2007). STIM2 Is a Feedback Regulator that Stabilizes Basal Cytosolic and Endoplasmic Reticulum Ca2+ Levels. *Cell* 131 (7), 1327–1339. doi:10.1016/j.cell.2007.11.039

Dagan, I., and Palty, R. (2021). Regulation of Store-Operated Ca2+ Entry by SARAF. *Cells* 10, 1887. doi:10.3390/cells10081887

des Georges, A., Clarke, O. B., Zalk, R., Yuan, Q., Condon, K. J., Grassucci, R. A., et al. (2016). Structural Basis for Gating and Activation of RyR1. *Cell* 167 (1), 145–157. doi:10.1016/j.cell.2016.08.075

Diercks, B. P., Werner, R., Weidemüller, P., Czarniak, F., Hernandez, L., Lehmann, C., et al. (2018). ORAI1, STIM1/2, and RYR1 Shape Subsecond Ca2+ Microdomains upon T Cell Activation. *Sci. Signal.* 11 (561), eaat0358. doi:10.1126/scisignal.aat0358

Feske, S. (2007). Calcium Signalling in Lymphocyte Activation and Disease. *Nat. Rev. Immunol.* 7, 690–702. doi:10.1038/nri2152

Gasser, A., Bruhn, S., and Guse, A. H. (2006). Second Messenger Function of Nicotinic Acid Adenine Dinucleotide Phosphate Revealed by an Improved Enzymatic Cycling Assay. *J. Biol. Chem.* 281 (25), 16906–16913. doi:10.1074/jbc.m601347200

Gil, D., Guse, A. H., and Dupont, G. (2021). Three-Dimensional Model of Sub-plasmalemmal Ca2+ Microdomains Evoked by the Interplay between ORAI1 and InsP3 Receptors. *Front. Immunol.* 12, 659790. doi:10.3389/fimmu.2021.659790

Gu, F., Krüger, A., Roggenkamp, H., Alpers, R., Lodygin, D., Jaquet, V., et al. (2021). DUOX2 Synthesizes NAADP in the Early Phase of T Cell Activation. *Sci. signaling* 14, eabe3800. doi:10.1126/scisignal.abe3800

Guo, T., Gillespie, D., and Fill, M. (2012). Ryanodine Receptor Current Amplitude Controls Ca 2+ Sparks in Cardiac Muscle. *Circ. Res.* 111 (1), 28–36. doi:10.1161/circresaha.112.265652

Guse, A. H., Gil Montoya, D. C., and Diercks, B.-P. (2021). Mechanisms and Functions of Calcium Microdomains Produced by ORAI Channels, D-Myo-Inositol 1,4,5-trisphosphate Receptors, or Ryanodine Receptors. *Pharmacol. Ther.* 223, 107804. doi:10.1016/j.pharmthera.2021.107804

Hogan, P. G. (2015). The STIM1-ORAI1 Microdomain. *Cell calcium* 58 (4), 357–367. doi:10.1016/j.ceca.2015.07.001

Hohenegger, M., Suko, J., Gscheidlinger, R., Drobny, H., and Zidar, A. (2002). Nicotinic Acid-Adenine Dinucleotide Phosphate Activates the Skeletal Muscle Ryanodine Receptor. *Biochem. J.* 367 (Pt 2), 423–431. doi:10.1042/BJ20020584

Jaggar, J. H., Porter, V. A., Lederer, W. J., and Nelson, M. T. (2000). Calcium sparks in Smooth Muscle. *Am. J. Physiology-Cell Physiol.* 278, C235–C256. doi:10.1152/ajpcell.2000.278.2.c235

Jayasinghe, I., Clowsley, A. H., Lin, R., Lutz, T., Harrison, C., Green, E., et al. (2018). True Molecular Scale Visualization of Variable Clustering Properties of Ryanodine Receptors. *Cel Rep.* 22 (2), 557–567. doi:10.1016/j.celrep.2017.12.045

Jones, P. P., MacQuaide, N., and Louch, W. E. (2018). Dyadic Plasticity in Cardiomyocytes. *Front. Physiol.* 9, 1773. doi:10.3389/fphys.2018.01773

Lanner, J. T., Georgiou, D. K., Joshi, A. D., and Hamilton, S. L. (2010). Ryanodine Receptors: Structure, Expression, Molecular Details, and Function in Calcium Release. *Cold Spring Harbor Perspect. Biol.* 2, a003996. doi:10.1101/cshperspect.a003996

Lewis, R. S. (2011). Store-operated Calcium Channels: New Perspectives on Mechanism and Function. *Cold Spring Harbor Perspect. Biol.* 3 (12), a003970. doi:10.1101/cshperspect.a003970

Li, Z., Liu, L., Deng, Y., Ji, W., Du, W., Xu, P., et al. (2011). Graded Activation of CRAC Channel by Binding of Different Numbers of STIM1 to Orai1 Subunits. *Cell Res* 21 (2), 305–315. doi:10.1038/cr.2010.131

Luik, R. M., Wang, B., Prakriya, M., Wu, M. M., and Lewis, R. S. (2008). Oligomerization of STIM1 Couples ER Calcium Depletion to CRAC Channel Activation. *Nature* 454 (7203), 538–542. doi:10.1038/nature07065

Mazel, T., Raymond, R., Raymond-Stintz, M., Jett, S., and Wilson, B. S. (2009). Stochastic Modeling of Calcium in 3D Geometry. *Biophysical J.* 96, 1691–1706. doi:10.1016/j.bpj.2008.10.066

McIvor, E., Coombes, S., and Thul, R. (2018). Three-dimensional Spatio-Temporal Modelling of Store Operated Ca2+ Entry: Insights into ER Refilling and the Spatial Signature of Ca2+ Signals. *Cell calcium* 73, 11–24. doi:10.1016/j.ceca.2018.03.006

Mezu-Ndubuisi, O. J., and Maheshwari, A. (2021). The Role of Integrins in Inflammation and Angiogenesis. *Pediatr. Res.* 89 (7), 1619–1626. doi:10.1038/s41390-020-01177-9

Okeke, E., Dingsdale, H., Parker, T., Voronina, S., and Tepikin, A. V. (2016). Endoplasmic Reticulum-Plasma Membrane Junctions: Structure, Function and Dynamics. *J. Physiol.* 594, 2837–2847. doi:10.1113/jp271142

Okubo, Y., Suzuki, J., Kanemaru, K., Nakamura, N., Shibata, T., and Iino, M. (2015). Visualization of Ca2+ Filling Mechanisms upon Synaptic Inputs in the Endoplasmic Reticulum of Cerebellar Purkinje Cells. *J. Neurosci.* 35, 15837–15846. doi:10.1523/JNEUROSCI.3487-15.2015

Oloveczky, B., and Verkman, A. (1998). Monte Carlo Analysis of Obstructed Diffusion in Three Dimensions: Applications to Molecular Diffusion in Organelles. *Biophysical J.* 74, 2722–2730.

Putney, J. W. (2009). Capacitative Calcium Entry: from Concept to Molecules. *Immunological Rev.* 231 (1), 10–22. doi:10.1111/j.1600-065x.2009.00810.x

Roggenkamp, H. G., Khansahib, I., Hernandez C, L. C., Zhang, Y., Lodygin, D., Krüger, A., et al. (2021). HN1L/JPT2: A Signaling Protein that Connects NAADP Generation to Ca^{2+} Microdomain Formation. *Sci. signaling* 14 (675), eabd5647. doi:10.1126/scisignal.abd5647

Rückl, M., and Rüdiger, S. (2016). Calcium Waves in a Grid of Clustered Channels with Synchronous IP_3 Binding and Unbinding. *The Eur. Phys. J. E, Soft matter* 39 (11), 108. doi:10.1140/epje/i2016-16108-4

Sato, D., and Bers, D. M. (2011). How Does Stochastic Ryanodine Receptor-Mediated Ca Leak Fail to Initiate a Ca Spark. *Biophysical J.* 101 (10), 2370–2379. doi:10.1016/j.bpj.2011.10.017

Schaff, J., Fink, C. C., Slepchenko, B., Carson, J. H., and Loew, L. M. (1997). A General Computational Framework for Modeling Cellular Structure and Function. *Biophysical J.* 73, 1135–1146. doi:10.1016/s0006-3495(97)78146-3

Solovey, G., Fraiman, D., Pando, B., and Ponce Dawson, S. (2008). Simplified Model of Cytosolic Ca2+ Dynamics in the Presence of One or Several Clusters of Ca2+ -release Channels. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 78 (4 Pt 1), 041915. doi:10.1103/PhysRevE.78.041915

Stathopulos, P. B., Li, G.-Y., Plevin, M. J., Ames, J. B., and Ikura, M. (2006). Stored Ca2+ Depletion-Induced Oligomerization of Stromal Interaction Molecule 1 (STIM1) via the EF-SAM Region. *J. Biol. Chem.* 281 (47), 35855–35862. doi:10.1074/jbc.m608247200

Streb, H., Irvine, R. F., Berridge, M. J., and Schulz, I. (1983). Release of Ca2+ from a Nonmitochondrial Intracellular Store in Pancreatic Acinar Cells by Inositol-1,4,5-Trisphosphate. *Nature* 306 (5938), 67–69. doi:10.1038/306067a0

Swaminathan, D., Ullah, G., and Jung, P. (2009). A Simple Sequential-Binding Model for Calcium Puffs. *Chaos* 19 (3), 037109. doi:10.1063/1.3152227

Taylor, C. W., and Machaca, K. (2019). IP3 Receptors and Store-Operated Ca2+ Entry: a License to Fill. *Curr. Opin. Cel Biol.* 57, 1–7. doi:10.1016/j.ceb.2018.10.001

Thillaiappan, N. B., Chavda, A. P., Tovey, S. C., Prole, D. L., and Taylor, C. W. (2017). Ca2+ Signals Initiate at Immobile IP3 Receptors Adjacent to ER-Plasma Membrane Junctions. *Nat. Commun.* 8 (1), 1505. doi:10.1038/s41467-017-01644-8

Thul, R., Thurley, K., and Falcke, M. (2009). Toward a Predictive Model of Ca2+ Puffs. *Chaos* 19 (3), 037108. doi:10.1063/1.3183809

Trebak, M., and Kinet, J.-P. (2019). Calcium Signalling in T Cells. *Nat. Rev. Immunolimmunology* 19 (3), 154–169. doi:10.1038/s41577-018-0110-7

Walker, M. A., Gurev, V., Rice, J. J., Greenstein, J. L., and Winslow, R. L. (2017). Estimating the Probabilities of Rare Arrhythmic Events in Multiscale Computational Models of Cardiac Cells and Tissue. *Plos Comput. Biol.* 13 (11), e1005783. doi:10.1371/journal.pcbi.1005783

Wolf, I. M., Diercks, B. P., Gattkowski, E., Czarniak, F., Kempski, J., Werner, R., et al. (2015). Frontrunners of T Cell Activation: Initial, Localized Ca2+ Signals Mediated by NAADP and the Type 1 Ryanodine Receptor. *Sci. Signal.* 8 (398), ra102. doi:10.1126/scisignal.aab0863

Wu, M. M., Buchanan, J., Luik, R. M., and Lewis, R. S. (2006). Ca2+ Store Depletion Causes STIM1 to Accumulate in ER Regions Closely Associated with the Plasma Membrane. *J. Cel. Biol.* 174 (6), 803–813. doi:10.1083/jcb.200604014

Yin, C.-C., D'Cruz, L. G., and Lai, F. A. (2008). Ryanodine Receptor Arrays: Not Just a Pretty Pattern. *Trends Cell Biology* 18 (4), 149–156. doi:10.1016/j.tcb.2008.02.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Gil, Diercks, Guse and Dupont. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Molecular Biosciences

METHODS
published: 23 February 2022
doi: 10.3389/fmolb.2022.800856

Optimal Experimental Design Based on Two-Dimensional Likelihood Profiles

Tim Litwin[1,2,3]*, *Jens Timmer*[2,3,4] *and Clemens Kreutz*[1,2,4]

[1]*Institute of Medical Biometry and Statistics (IMBI), Faculty of Medicine and Medical Center, University of Freiburg, Freiburg, Germany,* [2]*Freiburg Center for Data Analysis and Modelling (FDM), University of Freiburg, Freiburg, Germany,* [3]*Institute of Physics, University of Freiburg, Freiburg, Germany,* [4]*Centre for Integrative Biological Signalling Studies (CIBSS), University of Freiburg, Freiburg, Germany*

Dynamic behavior of biological systems is commonly represented by non-linear models such as ordinary differential equations. A frequently encountered task in such systems is the estimation of model parameters based on measurement of biochemical compounds. Non-linear models require special techniques to estimate the uncertainty of the obtained model parameters and predictions, e.g. by exploiting the concept of the profile likelihood. Model parameters with significant uncertainty associated with their estimates hinder the interpretation of model results. Informing these model parameters by optimal experimental design minimizes the additional amount of data and therefore resources required in experiments. However, existing techniques of experimental design either require prior parameter distributions in Bayesian approaches or do not adequately deal with the non-linearity of the system in frequentist approaches. For identification of optimal experimental designs, we propose a two-dimensional profile likelihood approach, providing a design criterion which meaningfully represents the expected parameter uncertainty after measuring data for a specified experimental condition. The described approach is implemented into the open source toolbox Data2Dynamics in Matlab. The applicability of the method is demonstrated on an established systems biology model. For this demonstration, available data has been censored to simulate a setting in which parameters are not yet well determined. After determining the optimal experimental condition from the censored ones, a realistic evaluation was possible by re-introducing the censored data point corresponding to the optimal experimental condition. This provided a validation that our method is feasible in real-world applications. The approach applies to, but is not limited to, models in systems biology.

Keywords: experimental design, profile likelihood, systems biology, mathematical model, parameter uncertainty, prediction uncertainty, confidence distribution

OPEN ACCESS

Edited by:
Ernesto Perez-Rueda, Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Juan Guillermo Diaz Ochoa, PERMEDIQ GmbH, Germany
Luis Diambra, National University of La Plata, Argentina

***Correspondence:**
Tim Litwin
litwin@imbi.uni-freiburg.de

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *24 October 2021*
Accepted: *07 January 2022*
Published: *23 February 2022*

Citation:
Litwin T, Timmer J and Kreutz C (2022) Optimal Experimental Design Based on Two-Dimensional Likelihood Profiles. Front. Mol. Biosci. 9:800856. doi: 10.3389/fmolb.2022.800856

1 INTRODUCTION

With Fisher's pioneering work on optimizing the design of agricultural experiments lying a century in the past, the design of informative experiments has long since become a foundation for most quantitative sciences. While there are undeniably practical aspects of conducting an experiment to generate the data used for analysis, planning a successful experiment requires consideration of statistical concepts even before any data is collected as this can help to develop the "logic of experimentation" (Bishop et al., 1982).

In systems biology, the underlying models used for analyses become increasingly complex. This is due to the fields aspiration to provide holistic descriptions of biological systems which are able to capture not only static properties of a system but the dynamic interactions of the system's components (Kitano, 2002; Nurse and Hayles, 2011). For these systems, mathematical models are established to reduce the complexity of the biological components to their relevant features. The process of "building a model" is an intertwined process of finding a model which adequately describes the observed dynamics given the existing biological knowledge and providing the quantitative inputs for this model through experimentation (Kreutz and Timmer, 2009). The aim of systems biology is to construct "useful" models (Wieland et al., 2021), i.e. models that yield biological insights. Assessing whether a model is useful can be "notoriously difficult" (Liepe et al., 2013), even more so if the data obtained from experiments is insufficient to inform the model. Therefore, close cooperation of experimenters and theoreticians throughout the process increases the chance of generating data that is suitable for this task.

Biochemical processes can often be represented as ordinary differential equations (Nurse and Hayles, 2011; Liepe et al., 2013; Raue et al., 2013) which are often adequate representations of molecular dynamics. In general, this means that the observed biochemical compounds will be non-linearly related to the model parameters. Although such models are able to describe the system realistically, non-linearity proves to be a challenge in the analysis of the models properties. One consequence of non-linearity is the frequent absence of analytical solutions to the differential equations which determine the time-evolution of the biological states involved. Consequently, estimation of model parameters by optimization of the objective function, which measures the deviation of the model predictions to the measured data, is limited to numerical approaches (Raue et al., 2013). The difficulty of this "inverse problem" (Liepe et al., 2013) of determining the model parameters which describe the observed data the best is exacerbated in biological systems. Characterization of these systems can lead to a model with many parameters and biological states with the available data being noisy (Kreutz and Timmer, 2009). Additionally, the system is generally only partially observable, i.e. not all biochemical compounds in the model can be measured (Raue et al., 2009).

A major task in developing experiments in this defined setting is to propose practically feasible experiments which decrease the uncertainty about the value of parameters of interest. A well-known result from the classical theory of non-linear experimental design is that the optimal design depends on the "true model parameters," i.e. the parameters that govern the true evolution of the system (Busetto et al., 2013), e.g. illustrated for the setting of Fisher's dilution series experiments (Cochran, 1973). However, we are interested in inferring exactly these unknown parameters. A solid initial guess about the parameter values would solve the problem, but given the complex nature of the modeled systems, prior knowledge is usually sparse (Kreutz and Timmer, 2009; Bazil et al., 2012). A natural approach is then to design experiments sequentially (Cochran, 1973; Ford et al., 1989), i.e. measure the data in batches, updating the knowledge about the initial parameter values for each experimental design iteration.

Much of the classical literature on designing the optimal experiment is based on the Fisher information matrix (Ford et al., 1989; Atkinson and Donev, 1992; Fedorov, 2010). This is a natural approach in linear systems, as the inverted Fisher information matrix determines the covariance matrix of the estimated model parameters. Appropriate characteristics of this covariance matrix are then optimized by a suitable experimental design (Atkinson and Donev, 1992; Faller et al., 2003). However, application of the Fisher information matrix is known to be troublesome in non-linear systems if the amount of data is limited and statistical properties are far from asymptotic. The Wald confidence intervals implied by the Fisher information matrix might then only crudely reflect the existing uncertainty. Confidence intervals generated by the profile likelihood approach have more desirable properties in the finite sample case (Meeker and Escobar, 1995) and allow for the conceptional and operational definition of practical identifiability (Raue et al., 2009). Experimental planning in frequentist statistics should therefore make use of this powerful concept of quantifying parameter uncertainty and identifiability.

Approaches to the experimental design problem have also been developed in a Bayesian framework. The conceptual foundation of updating prior parameter knowledge given the newly measured data in Bayesian statistics provides natural solutions to the problem of experimental design. The information gain of an experiment can be reasonably quantified by means of the Shannon information (Lindley, 1956) and application of this theory to Bayesian experimental design provides a tool to plan optimal experiments for parameter inference (Huan and Marzouk, 2013; Liepe et al., 2013) and model discrimination (Busetto et al., 2013). However, we focus on a frequentist approach as it is usually not feasible to provide reasonable priors for all model parameter in the systems biology context.

There exist frequentist methods for experimental design if it is infeasible to provide prior information for all model parameters. If sets of parameters which are compatible with existing data about the system were known, the corresponding set of model trajectories would indicate for which observables and for which time points the model prediction is not yet reasonably constrained; such experimental conditions would then be "experimentally distinguishable" (Bazil et al., 2012). This set was previously constructed from efficient sampling of the parameter space (Bazil et al., 2012) or exploring the parameters along the likelihood profiles (Steiert et al., 2012). The latter method was applied in the DREAM6-Challenge (Dialogue for Reverse Engineering Assessments and Methods) and has been awarded as the best performing approach (Steiert et al., 2012). However, this approach assesses the impact of different sets of model parameters on the model predictions. In order to optimally design experiments which decrease parameter uncertainty, the logic of the design scheme has to be reversed: Instead of assessing the impact of different model parameters on the model prediction, the impact of different measurement outcomes on the parameter estimate of interest has to be assessed.

We propose a frequentist approach for optimal experimental design which realizes the full potential of the profile likelihood approach by extending the previously best-performing method (Steiert et al., 2012). For a specified experimental condition, we quantify the expected uncertainty of a targeted parameter of interest after a possible measurement. The parameter uncertainty after any specific measurement outcome is determined by the respective profile likelihood, effectively yielding a two-dimensional likelihood profile when accounting for different possible measurement outcomes. The range of reasonable measurement outcomes given the current data available before the measurement is quantified via the concept of validation profiles (Kreutz et al., 2012). The two-dimensional likelihood profile provides both the range of reasonable measurement outcomes of an intended experiment and their impact on the parameter likelihood profile. Hence, this allows for a definition of a design criterion which represents the expected average width of the confidence interval after measuring data for a certain experimental condition. The two-dimensional likelihood profiles therefore provide quantitative information usable for sequential experimental design and additionally serve as an intuitive tool to visualize the impact of an experiment on the uncertainty of the parameter of interest.

2 MATERIALS AND METHODS

2.1 Mathematical Model

We introduce the concept of ordinary differential equation models, because they are frequently used for modeling of the dynamics of biological systems. However, we want to emphasize that the introduced method for experimental design is generic and only requires specification of a suitable likelihood function.

Biological quantities such as the concentration of a molecular compound are represented by mathematical states $x(t)$ and are assumed to follow a set of ordinary differential equations

$$\dot{x}(t) = f(x, p, u) \tag{1}$$

which generates the trajectories according to the unknown underlying true dynamic model parameters p_0. The function f is typically defined by translating biochemical interactions, e.g. by the rate equation approach. The trajectories depend on the specific experiment conducted which is denoted by the experimental perturbations u, representing interventions such as external stimulation of the system or knockout of specific genes. The set of model parameters will usually include the initial values x_0 of the model states.

Estimation of the true parameters typically requires measurement of time-resolved data on these states. However, some states in the considered system might not be observable at all or only indirectly accessible, e.g. if only a sum of different states can be observed (Raue et al., 2009). Additionally, measured data will usually be subject to random errors. Therefore, the set of observables

$$y(t) = g(x(t), s_{obs}) + \epsilon \tag{2}$$

defines the types of data that can be measured. In this equation, ϵ describes the random error of the measurement which is usually assumed to be normally distributed, i.e. $\mathcal{N}(0, \sigma^2(x(t), s_{err}))$, but not necessarily homogeneous across measurements, i.e. the magnitude of the noise might depend on parameters s_{err}. Random variations in biological systems usually occur on a relative scale (Limpert et al., 2001) and are thus proportional to the current value of the state. This implies that errors are frequently normally distributed if the observables are considered on a logarithmic scale. The observation function g determines how the states are mapped unto the observables. This mapping will on many occasions introduce new unknown parameters s_{obs} such as scale parameters. The set of all parameters is denoted by $\theta = \{p, x_0, s_{obs}, s_{err}\}$.

The measured data of the system provides a set of scalar values y_i which each corresponds to an experimental condition D_i containing all information necessary to interpret the value y_i. The experimental condition is uniquely defined as the measured observable, the time point of measurement and the corresponding experimental perturbation.

The objective function which indicates the agreement of experimental data with the model prediction given some parameters θ and measured data $Y = \{y_1, \ldots, y_n\}$ is the likelihood function

$$L(\theta|Y) = \prod_i \rho(y_i|D_i, \theta) \tag{3}$$

with ρ indicating the probability density for the considered data point. Maximizing this likelihood leads to the maximum likelihood estimate $\hat{\theta}(Y)$ which indicates the parameters for which the fit between data and model predictions is optimal. Numerical optimization of this function is preferably performed by minimizing the monotonously transformed function $LL = -2\ln(L)$ to improve numerical stability. If the data is independently normally distributed and variances are known, this transformation has the advantageous properties that the optimization of the objective function is equivalent to least squares optimization.

2.2 Profile Likelihood

2.2.1 Parameter Profile Likelihood

The task of parameter inference is not completed with the identification of the maximum likelihood estimate. In general, other parameter estimates may provide other model trajectories which might fit similarly well to the given data. Additionally, replications of the same experiment will lead to different measurement results and therefore also different parameter estimates due to variance in the biological samples and the measurement process. From a frequentist standpoint, methods are required to construct confidence intervals for either individual parameters or multiple parameters jointly, which have a pre-defined coverage probability of containing the true parameter value if the experiment were to be replicated. Within the context of this paper, we focus exclusively on confidence intervals for individual parameters.

The commonly encountered Wald confidence intervals are based on a quadratic approximation of the likelihood and fail if

the model features non-linear dynamics (Meeker and Escobar, 1995; Raue et al., 2009). The quadratic approximation of the likelihood depends on the parametrization of the model, may not respect boundaries of the parameter space and cannot capture global behavior such as the existence of local optima.

A more refined tool which reduces the high-dimensional likelihood onto the one-dimensional parameter of interest p_i is the profile likelihood

$$PL_{p_i}(\beta|Y) := -2\ln\left(\frac{L(\beta, \hat{\omega}(\beta)|Y)}{L(\hat{\beta}, \hat{\omega}|Y)}\right) \tag{4}$$

with the parameter vector $\theta = \{\beta, \omega\}$ being split into the parameter of interest $p_i = \beta$ and the nuisance parameters $p_{i\neq j} = \omega$. The hats indicate maximum likelihood estimates, i.e.

$$\hat{\omega}(\beta) = \arg\max_{\omega} L(\beta, \omega|Y) \tag{5}$$

are the nuisance parameters which maximize the likelihood if β is fixed to a specific value. The parameter profile likelihood is invariant under one-to-one parameter transformations and can accurately reduce complex shapes of the underlying likelihood function to an adequate one-dimensional representation. Confidence intervals can be constructed from the parameter profile by Wilks' Theorem (Wilks, 1938) and take the form

$$CI_\alpha = \{\beta | PL_{p_i}(\beta|Y) < icdf(\chi^2_1, \alpha)\}. \tag{6}$$

with *icdf* representing the inverse cumulative distribution function. Note that high values of the profile likelihood defined in **Eq. 4** correspond to lower values of the likelihood. This implies that parameter values β associated with a large profile likelihood value $PL_{p_i}(\beta|Y)$ are less likely to correspond to the true parameter value. Therefore, only parameter values with a profile likelihood value below a certain confidence threshold are included in the corresponding confidence interval.

Informally, Wilks' theorem implies that asymptotically, these confidence intervals will attain the correct coverage probability α as they become equivalent to the Wald approximation. However, the finite sample properties of the profile likelihood intervals are superior. The notion of parameter profiles allows identifiability analyses on the parameters (Raue et al., 2009). Parameters can be: 1) Identifiable, in which case the width of the defined confidence interval is finite. 2) Structurally non-identifiable, in which case the profile likelihood is flat. This implies that any change of the parameter of interest can be compensated by changing other model parameters. 3) Practically non-identifiable, in which case the profile likelihood is not completely flat, but does not cross the confidence threshold to both sides such that the size of the confidence interval is infinite. While structural non-identifiability can only be resolved either by reparametrization of the model or qualitatively new experiments, practically non-identifiability can usually be resolved by providing higher quality data from similar experiments. Identifiability is distinct from the frequently encountered concept of sloppiness (Chis et al., 2016) which plays no role for the experimental design as discussed within this study. Due to the advantageous theoretical as well as practical properties of the profile likelihood, parameter uncertainties in this study are exclusively discussed in terms of their corresponding likelihood profile.

2.2.2 Validation Profile Likelihood

The parameter profile likelihood allows for the evaluation of the uncertainty of parameters given the current data. For some applications, assessing the "prediction uncertainty," i.e. the uncertainty about the outcome of measuring at a certain experimental condition, might be more relevant. In a frequentist setting, one can readily extend the concept of the parameter profile likelihood to this setting in the form of the "validation profile likelihood" (Kreutz et al., 2012), also called "predictive profile likelihood" (Bjornstad, 1990), in which case the likelihood is reduced to the dimension of the measurement outcome of interest. Formally, this profile is defined by

$$VPL(z|y) := -2\ln\left(\frac{L(\hat{\theta}(Y,z)|Y,z)}{L(\hat{\theta}(Y)|Y, F(\hat{\theta}(Y)|D_z))}\right) \tag{7}$$

with z defined as the outcome of measuring at experimental condition D_z and $F(\hat{\theta}(Y)|D_z)$ defined as the respective model prediction given parameters $\hat{\theta}(Y)$. The interpretation of this validation profile likelihood is completely analogous to the parameter profile likelihood since the same coverage statement as in **Eq. 6** holds (Kreutz et al., 2012) if the random variable z is normally distributed. It should be remarked that the statement about the coverage is slightly adapted in the sense that the coverage probability is true if Y and z are repeatedly drawn. Just as parameter uncertainty is reasonably quantified by the parameter profile likelihood, uncertainty of the measurement outcome for an experimental condition of interest is specified by the validation profile likelihood.

2.3 Experimental Design Task

Understanding the task of designing an informative experiment requires clarification. We start by introducing the common terminology of the theory of optimal experimental design. The design region $\mathcal{K}$ is the set of all experiments that can be conducted and a design point (or experimental condition) D within this design region labels a possible experiment which returns a data point (Fedorov, 2010). In our context, the design region $\mathcal{K}$ is designed as the set of all admissible triples of measurable observables g, possible time points of the experiment t as well as all external perturbations u considered. A design point D is then defined as the triple $D = \{g, t, u\}$ (Kreutz et al., 2012). The collective of all conducted experiments can therefore be represented as the set of corresponding design points which is called the design of the experiment. This design is more commonly represented by a probability measure ξ on the design space $\mathcal{K}$ which compactly specifies all design points for which measurements outcomes are available. (Ford et al., 1989; Fedorov, 2010).

The problem we are concerned with is the reduction of uncertainty for a single parameter by conducting an experiment at an informative design point. This means that given a set of admissible design points, we want to decide which of these experimental designs will best reduce the

existing uncertainty about a pre-specified parameter. To put this into a more formal framework, we are looking for a design criterion $\Phi: \mathcal{K} \to \mathbb{R}$ which quantifies the most informative experiment in this context. Note that the optimal experimental design will usually depend on the unknown true parameters of the system. Additionally, we want to emphasize that we are concerned only with a single best experiment and not a batch of parallel experiments conducted at the same time.

2.4 Measuring Parameter Uncertainty

2.4.1 Classical Theory

The optimal experimental design depends on the choice of a reasonable design criterion. Classical design theory solves this problem by applying the Fisher information matrix as the appropriate measure of information and establishing design criteria based on this matrix $M(\theta, \xi)$, i.e. the design criterion takes the form $\Phi(M(\theta, \xi))$ (Ford et al., 1989; Atkinson and Donev, 1992; Fedorov, 2010). The Fisher information matrix is concerned with the local behavior of the likelihood function around a specified parameter, which in application usually means in the neighborhood of the current maximum likelihood estimate of the parameters (Faller et al., 2003). For the same reasons discussed earlier, we propose that it is more adequate to utilize the profile likelihood of the parameter of interest p_i to construct a measure of information which we can use to design an optimal experiment.

2.4.2 Confidence Distribution

There is no unique way to define the information available in the likelihood profile $\{(\beta, PL_{p_i}(\beta))|\beta \in\}$ of the parameter of interest p_i. Instead of thinking in terms of the available information about the parameter, it is instructive to think in terms of existing uncertainty which we want to minimize by the experiment. In practical applications it is common that profile likelihoods are evaluated to obtain the respective 95%-confidence intervals $CI_{0.95}(p_i)$ which serve as a measure of existing uncertainty. This comes with two notable issues: The 95%-interval might not be finite, which complicates the interpretation of existing uncertainty. This can be resolved in the definition of the model's parameter space, which is constrained by parameter boundaries which span orders of magnitude and thus only weakly constrain the possible parameter values. On a more conceptual level, working with arbitrary fixed confidence levels is discouraged (Wasserstein and Lazar, 2016) and uncertainty is more comprehensively assessed if all confidence levels are considered simultaneously.

This issue can be resolved by confidence distributions (Xie and Singh, 2013) which can be thought of simultaneously containing the information about the confidence intervals to all levels. This concept allows the construction of an object that has the form of a distribution estimator for the parameter of interest p_i in the realm of frequentist statistics. The corresponding confidence density $\rho_{p_i}^{par}(\beta)$ has the property that each interval $[\beta_1, \beta_2]$ which satisfies $\int_{\beta_1}^{\beta_2} \rho_{p_i}^{par}(\beta)\mathrm{d}\beta = \alpha$ is an α%-confidence interval for the parameter p_i. Conceptually, we can derive the confidence density $\rho_{p_i}^{par}(\beta)$ for a parameter from the set of confidence intervals $\{CI_\alpha(p_i)|\alpha \in [0, 1]\}$ obtained from its likelihood profile. However, we will not use the concept explicitly and we remark that in the case of finite sample size, the obtained confidence distributions are not exact. By associating the parameter of interest p_i with their corresponding confidence distribution $\rho_{p_i}^{par}$, we have a theoretically well-founded quantity on which the uncertainty of the parameter can be quantified.

2.4.3 Uncertainty as a Scalar Quantity

Ranking different experiments by their information content requires a way to order their corresponding design criterion values. A necessary step to achieve this is to reduce the confidence distribution of a parameter to a scalar value. We suggest utilizing the average confidence interval width

$$\bar{w}(p_i|Y, z) = \int_0^1 w(CI_\alpha(p_i|Y, z))d\alpha \tag{8}$$

to summarize the information content of the confidence distribution. The function w assigns the width to the corresponding confidence interval. Different confidence interval widths are averaged by weighting with their respective confidence measure $d\alpha$. The measure $d\alpha$ specifies the confidence that the true parameter value is covered by the interval $CI_{\alpha+\mathrm{d}\alpha}(p_i)$, but not by $CI_\alpha(p_i)$. Evaluation of this average confidence interval width does not require the explicit confidence distribution but only the individual confidence intervals. Thus, it can be directly calculated from the profile likelihood. In practice, we will only consider confidence intervals up to the 95%-level to ensure practical feasibility.

2.5 Two-Dimensional Profile Likelihood as a Design Criterion

In the previous sections, we proposed to quantify parameter uncertainty via the profile likelihood approach by definition of an average profile width in **Eq. 8**, which summarizes the existing uncertainty about the parameter of interest. Optimal experimental design aims at minimizing this measure of uncertainty by choosing an experimental condition for the next measurement which optimizes a suitable design criterion. However, for a given experimental condition D it is a priori unknown which value will result from a future measurement. This implies that the parameter profile likelihood $PL_{p_i}(\beta|Y, z)$ and therefore the average profile width $\bar{w}(p_i|Y, z)$ depends on the measurement outcomes z. In **Figure 1A**, the original parameter profile before the measurement (black line) is practically non-identifiable at the 95%-level for the parameter of interest. Depending on the measured data point z_i, the uncertainty about the parameter of interest is reduced to different degrees as indicated by the corresponding parameter profiles (blue lines). Since the measurement outcome is unknown, it is unclear what the uncertainty will be after the measurement.

The plausibility of different possible measurement outcomes can be accounted for by weighting the average profile widths for different measurement outcomes by their likelihood of occurrence. As discussed in **Section 2.2.2**, this plausibility measure is implied by the validation profile likelihood. **Figure 1B** shows the validation profile for the specified design: z_2 corresponds to the current

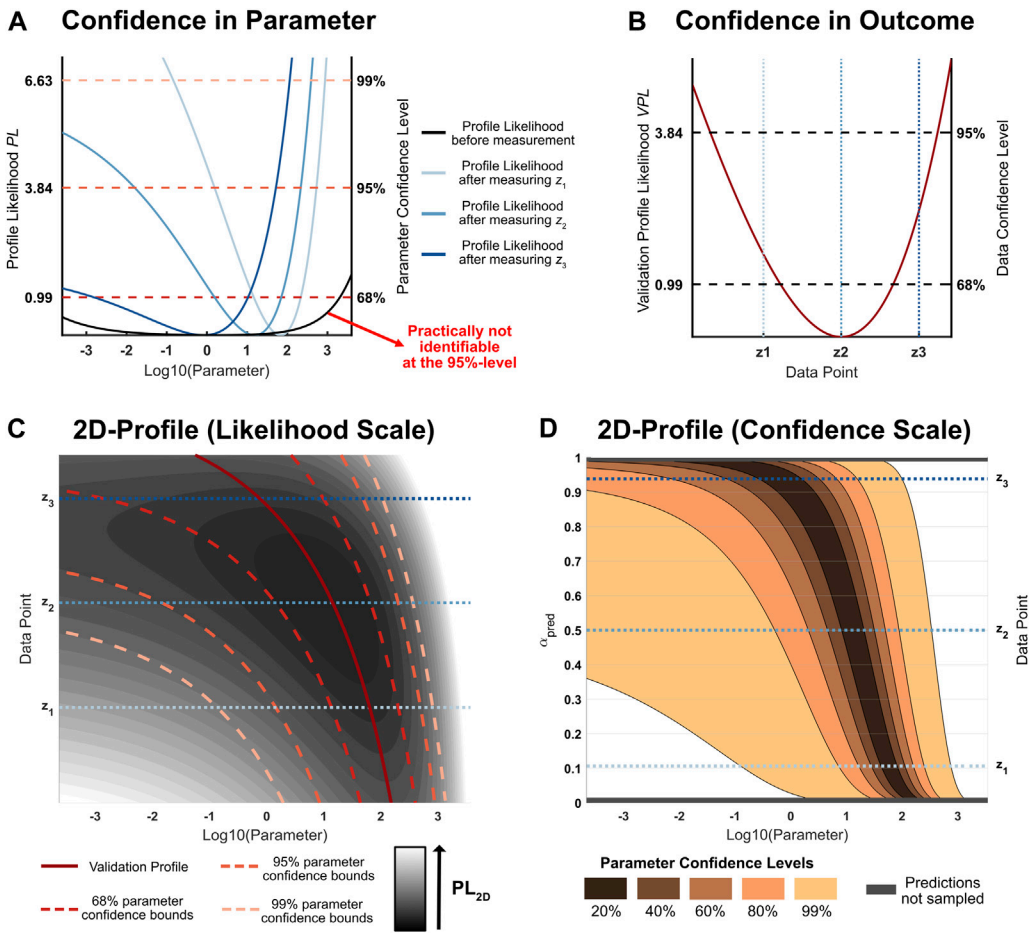


FIGURE 1 | (A): Likelihood profiles of a hypothetical parameter of interest. Different measurement outcomes z_1, z_2, z_3 for the same experimental condition lead to different updated parameter profiles which assess uncertainty about a parameter of interest. **(B)**: Validation profile of the considered hypothetical experimental condition. This profile assesses the likelihood of a new measurement: The smaller the validation profile value, the more likely the respective outcome **(C)**: 2D-Likelihood profile for the parameter of interest under some given experimental condition. The vertical axis corresponds to different possible measurement outcomes. If the outcome on the vertical axis would be observed, the profile likelihood after the measurement is given by the corresponding horizontal cross-section through the two-dimensional profile. In this example, lower values of the measurement outcome lead to narrow parameter confidence intervals after the measurement. **(D)**: 2D-Likelihood profile on a confidence scale. Intervals of the same size on the *y*-axis hold equal confidence that a measurement will yield a data point in the corresponding interval. The prediction confidence levels on the vertical axis illustrate that the sampled two-dimensional likelihood profile covers most of the plausible measurement outcomes.

maximum likelihood prediction for this experimental condition and is therefore the most likely measurement outcome given the current evidence, while z_3 has a higher validation profile value than z_1 and is therefore less likely. Therefore, the validation profile likelihood implies a predictive distribution which can be defined in analogy to the confidence distribution derived from the parameter profile likelihood. The corresponding predictive density $\rho^{pred}(z|Y)$ associates different measurement outcomes with our confidence that the specific outcomes occur.

The concept of summarizing parameter uncertainty for a fixed measurement outcome and subsequent aggregation of different possible measurement outcomes based on the predictive density can be combined to construct a design criterion for an experimental condition of interest. To this end, each expected parameter profile width $\bar{w}(p_i|Y,z)$ is weighted with the certainty $\rho^{pred}(z|Y)$ of observing measurement outcome z and the expected average profile width follows as

$$W(p_i|Y,D) = \int \bar{w}(p_i|Y,z)\rho^{pred}(z|Y)\mathrm{d}z. \tag{9}$$

$W(p_i|Y, D)$ exclusively depends on the given data Y and the experimental condition D of a subsequent experiment and thus by definition constitutes a design criterion. This quantity can be interpreted as the expected average profile width after measuring at the experimental condition D, where the average is taken over different parameter confidence levels and the expectation is taken over different possible measurement outcomes, weighted by their predicted plausibility. Given a set of experimental conditions $\{D_1, \ldots, D_n\}$, the optimal experiment $D^\star$ to inform parameter p_i is the one which minimizes $W(p_i|Y, D)$ given the current data Y, i.e.

$$D^{\star} = \underset{D\in\{D_1,\ldots,D_n\}}{\arg\min}\ W(p_i|Y,D). \tag{10}$$

The information necessary to evaluate the design criterion in **Eq. 9** is summarized by a two-dimensional likelihood profile, defined as

$$PL_{2D}(z,\beta|D_z) := -2\ln\left(\frac{L(\beta,\hat{\omega}(\beta|Y,z)|Y,z)}{L(\hat{\theta}(Y)|Y,F(\hat{\theta}(Y)|D_z))}\right). \tag{11}$$

For any fixed measurement outcome z, the resulting parameter profile likelihood can be extracted from this quantity. Simultaneously, the two-dimensional likelihood profile contains information about the plausibility of different measurement outcomes. **Figure 1C** illustrates this relationship: The profiles in **Figure 1A** are horizontal cross-sections from the two-dimensional likelihood profile (blue lines). The minimal profile value of each horizontal cross-section defines the path of the validation profile in **Figure 1B** (solid red line). The confidence intervals (dashed red lines) depend on the different possible measurement outcomes: Some measurement outcomes lead to more information about the parameter of interest than others as indicated by narrower confidence intervals.

The process of averaging confidence interval widths over the various parameter confidence levels and taking the expected value over the possible plausible data realizations is visualized in **Figure 1D**. The displayed two-dimensional profile is based on the same data as depicted in **Figure 1C**, but it has been transformed onto a different scale. The minimum of each horizontal cross section is shifted to the common null value, but still represents the trajectory of the validation profile. For the transformed two-dimensional likelihood profile, the vertical axis is proportional to the prediction confidence levels, i.e. intervals with the same length correspond to an equal confidence of yielding a measurement value in the given intervals. This transformation reveals that the interval $[z_1, z_3]$ is a 83%-prediction interval for a future measurement outcome given the experimental condition. The horizontal gray patches at the top and the bottom of **Figure 1D** correspond to all the measurement outcomes for which the original two-dimensional profile likelihood was not sampled, because they are unlikely to occur. The trend of different parameter confidence intervals as a function of different data points is illustrated for five discrete confidence levels (shades of red). On this scale, the expected average profile width $W(\beta|Y, D)$ is equal to the average of all the colored areas, where the smaller confidence intervals are included in the larger ones.

2.6 Experimental Design Workflow

Utilization of two-dimensional likelihood profiles as a tool for experimental design requires a ready-to-use workflow in applications. We provide an example for this workflow in a fully sequential experimental design scheme to put the previous definitions into a more practical context. **Figure 2** shows a flowchart of the steps involved in this workflow. Starting from an initial data set, the model parameters are estimated and the profile likelihood is calculated for all model parameters to obtain information about existing parameter uncertainties. The likelihood profiles are calculated by numerical evaluation of **Eq. 4** for a finite set of profile parameters. If there are non-identifiable parameters, the biologically most relevant parameter is targeted for improvement by the experimental design scheme.

After a representative set of experimental conditions has been defined, the design criterion in **Eq. 9** needs to be evaluated for each of the experimental conditions by the following steps. First, a validation profile is calculated for the experimental condition. This validation profile provides the range of relevant measurement outcomes for the respective experimental condition. Therefore, the space on which the two-dimensional likelihood profile needs to be sampled is finite. This space is sampled by evaluating the parameter profile likelihood for a representative set of measurement outcomes. The expected average profile width is calculated from the two-dimensional likelihood profile by employing the discrete counterparts of all expressions appearing in **Eq. 9**. At this point the details are more of technical than conceptual relevance and we want to emphasize that an automated implementation of this algorithm is available and referred to at the end of this manuscript. The final step of the workflow is now to choose the experimental condition which provides the minimal value for the design criterion as the target for the next measurement. This workflow can be repeated after a new data point has been generated to determine a sequence of informative measurements.

3 RESULTS

We illustrate the process of choosing the best experimental design for a parameter of interest by two examples. The first example is based on simulated data for a simple model with two consecutive reactions in which compound A is converted to compound B which is then converted to compound C and is therefore termed as ABC model in the following. This example will serve to illustrate the interpretation of the two-dimensional profile likelihood. The second example is based on the published experimental data for a model of erythropoietin (EPO) degradation (Becker et al., 2010) for which data has been censored in order to mimic a setting in which experimental design can be applied. This example serves to explain the full workflow of the sequential experimental design scheme in an application setting and illustrates the practical feasibility of our approach.

3.1 Experimental Design in the ABC-Model

The ABC model describes a simple case of a model in which the model predictions non-linearly depend on the model parameters. The reactions are illustrated in **Figure 3A**: State A is converted to state B with the rate p_1 and B is subsequently converted into compound C with rate p_2. In a biochemical setting, these three states might represent three conformations of three activation states in terms of different phosphorylations. The dynamics of this system are determined by the following differential equations:

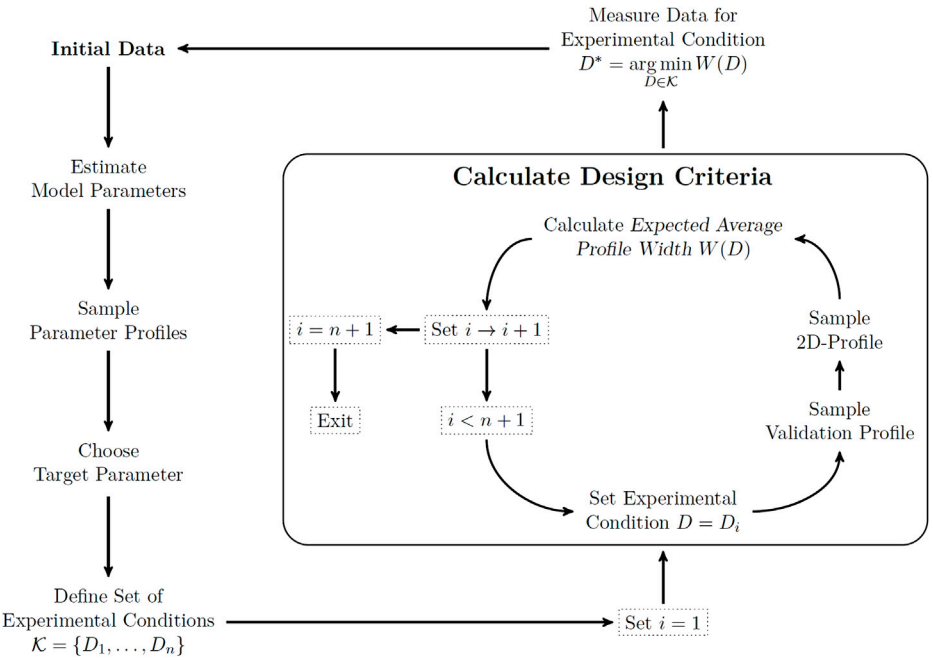


FIGURE 2 | Workflow for the sequential experimental design scheme. Starting from the current data set (top left), the target parameter is chosen and relevant experimental conditions are specified. Calculating the two-dimensional profile likelihood and evaluating the expected average profile width for each experimental condition (box) reveals the optimal condition for the next measurement. Dotted rectangles specify the state of the *for loop*, while text without rectangles correspond to the experimental design steps involved.

$$\dot{A} = -p_1 A$$
$$\dot{B} = p_1 A - p_2 B$$
$$\dot{C} = p_2 B.$$

In order to illustrate the two-dimensional likelihood profile approach on a simple model for which we know the true underlying parameters, we defined the true model parameters and simulated data from this model. In this true model, the initial concentrations of state B and state C were set to zero which was assumed to be known for inference, such that the system is characterized by the three parameters $\{p_1, p_2, A_0\}$ which were assumed to be unknown and have to be estimated from data. The data set simulated from the true model parameters is sparse: state A has been assumed to not be observable and the two data points available for each state B and C have been generated with an initial concentration $\log(A_0) = 0$ from log-normal distributions with a standard deviation $\sigma_{\log} = 0.2$ After simulation of the data, the model parameters $\{p_1, p_2, A_0\}$ are optimized to estimate their values and the corresponding state trajectories.

The true as well as the estimated state trajectories are illustrated in **Figure 3B**: While the model predictions fit the data well, there is still considerable disagreement of the underlying true model and the best model fit. This is especially true for state A, considering that the differences between trajectories are analyzed on a log-scale, which measures relative differences. An analysis of parameter uncertainty reveals that parameter p_2 is identifiable, as information for state B and C suffices to inform this rate. By similar reasoning, there is less information available for parameter p_1 and the corresponding profile likelihood reveals that the parameter is practically non-identifiable over the whole considered parameter space, as illustrated in **Figure 3C**.

In our example, we want to inform this practically non-identifiable parameter p_1 by choosing a measurement out of three possible experimental conditions. For demonstration purposes, the three experimental conditions of measuring either state A, B or C at the time point $t = 40$ are considered. The corresponding two-dimensional likelihood profiles are illustrated in **Figure 3D**. If it was possible to measure observable A, this would be highly informative and in fact guarantees that the parameter p_1 is identifiable no matter the outcome of the measurement. This is intuitive, since the measurement of the yet unobserved quantity A highly constrains the possible dynamics. It should be noted that possible outcomes for the observable A vary across orders of magnitudes which can be attributed to the fact that the dynamics for A are poorly constrained given the current data set. The two-dimensional profiles associated with observable B and C reveal that the parameter p_1 will likely not be identifiable even after the measurement. However, the magnitude of outcomes will yield at least an upper or a lower bound for the parameter of interest: Large values of B and C put an upper limit on p_1, as this means that the reaction can not be arbitrarily fast, while low values of B and C put a lower limit of p_1 because the reaction can not be arbitrarily slow. A not immediately obvious result from the two-dimensional profiles is that measuring observable B is more informative than measuring observable C as seen from the calculated design criterion. This example illustrates that two-

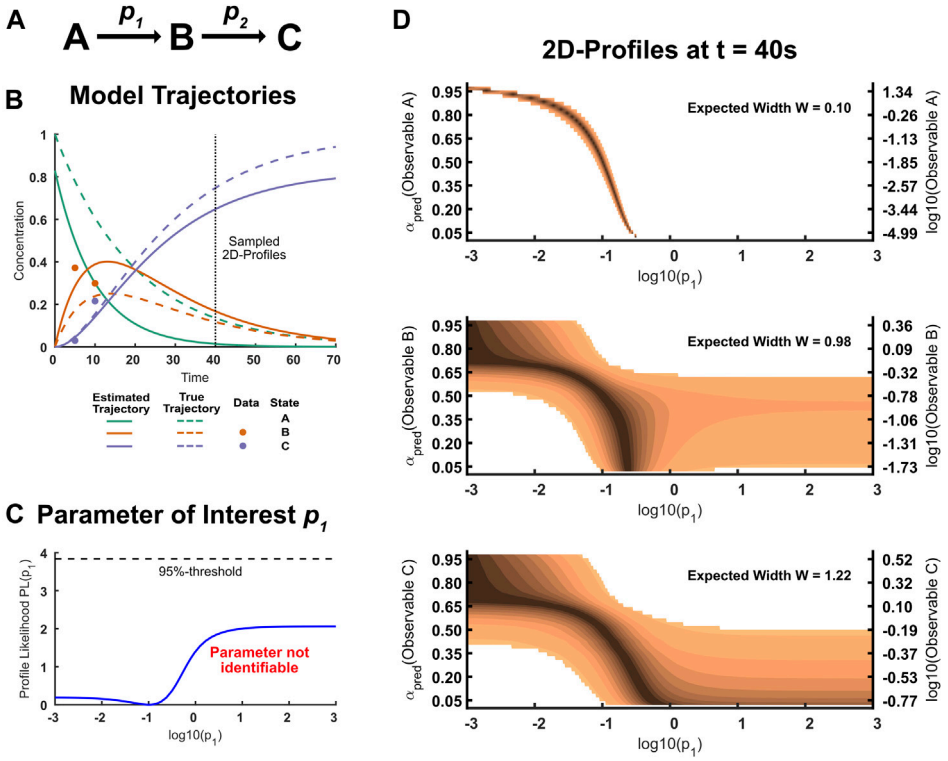


FIGURE 3 | (A): States and parameters in the ABC-model. The model has three parameters: Two rate constants p_1 and p_2, and the initial concentration A_0. The initial concentrations of B and C are assumed to be zero. **(B)**: Trajectories of the ABC-Model. The dots correspond to the sparse data simulated from the true model. In this example, state B and C were assumed to be observable, but have only been observed at early time points. The true trajectory of state A yet differs considerably from the estimated trajectory. **(C)**: Likelihood profile of the practically non-identifiable parameter p_1. Because the initial concentration of state A is unknown, this parameter is difficult to estimate without information about state A. **(D)**: 2D-Likelihood profiles for the three states A, B and C if measured at time point t = 40. The illustrated profiles are presented on a confidence scale according to **Figure 1D**. If state A was observable, the finite width of the 2D-profile to the 95% level indicates that any measurement outcome will make the parameter p_1 identifiable. Note that possible values for A scatter across six orders of magnitude because the predictions for A are barely informed. Measuring state B or C will likely put an lower or upper limit on the parameter p_1.

dimensional likelihood profiles provide qualitative as well as quantitative information about how experiments impact parameter uncertainty.

3.2 Experimental Design in the Erythropoietin Degradation Model

The modeled system for the degradation of erythropoietin (EPO) (Becker et al., 2010) is an example of a non-linear model with intertwined reactions of biochemical states. EPO acts as a ligand by binding to the corresponding cell receptor to form a complex. This complex is internalized and then EPO is degraded. The mathematical model provided the insight that a combination of EPO receptor turnover and recycling guarantees that biochemical response to a broad range of ligand concentrations is possible (Becker et al., 2010).

A scheme of reactions in the biological system is illustrated in **Figure 4A**. The model features six dynamic states: EPO (Epo) and degraded EPO ($dEpo_e$) outside of the cell, EPO receptors ($EpoR$) and EPO–EPO receptor complexes ($EpoEpoR$) on the cell membrane, and internalized EPO–EPO receptor complexes ($EpoEpoR_i$) and degraded EPO ($dEpo_i$) inside of the cell. The reactions illustrated in the figure can be translated into the following set of differential equations (Becker et al., 2010):

$$\begin{aligned}
\frac{d}{dt}EpoR &= k_t \cdot (EpoR_0 - EpoR) - k_{on} \cdot Epo \cdot EpoR + k_{off} \cdot EpoEpoR + k_{ex} \cdot EpoEpoR_i \\
\frac{d}{dt}Epo &= -k_{on} \cdot Epo \cdot EpoR + k_{off} \cdot EpoEpoR + k_{ex} \cdot EpoEpoR_i \\
\frac{d}{dt}EpoEpoR &= k_{on} \cdot Epo \cdot EpoR - k_{off} \cdot EpoEpoR - k_e \cdot EpoEpoR \\
\frac{d}{dt}EpoEpoR_i &= k_e \cdot EpoEpoR - k_{de} \cdot EpoEpoR_i - k_{di} \cdot EpoEpoR_i - k_{ex} \cdot EpoEpoR_i \\
\frac{d}{dt}dEpo_i &= k_{di} \cdot EpoEpoR_i \\
\frac{d}{dt}dEpo_e &= k_{de} \cdot EpoEpoR_i.
\end{aligned}$$

There are seven dynamic parameters (k_t, k_{on}, k_{off}, k_{ex}, k_e, k_{de}, k_{di}) and two unknown initial conditions (Epo_0, $EpoR_0$) in the model which are biologically interpretable as well as six further parameters which appear only in the observation function and not in the dynamic model. Because EPO can be traced with a radioactive marker, the concentration of EPO can be measured outside of the cell ($EPO_{external}$), on the cell membrane ($EPO_{membrane}$) and inside of the cell ($EPO_{internal}$). This provides us with three different observables for the six dynamic states, i.e. the model is only partially observable.

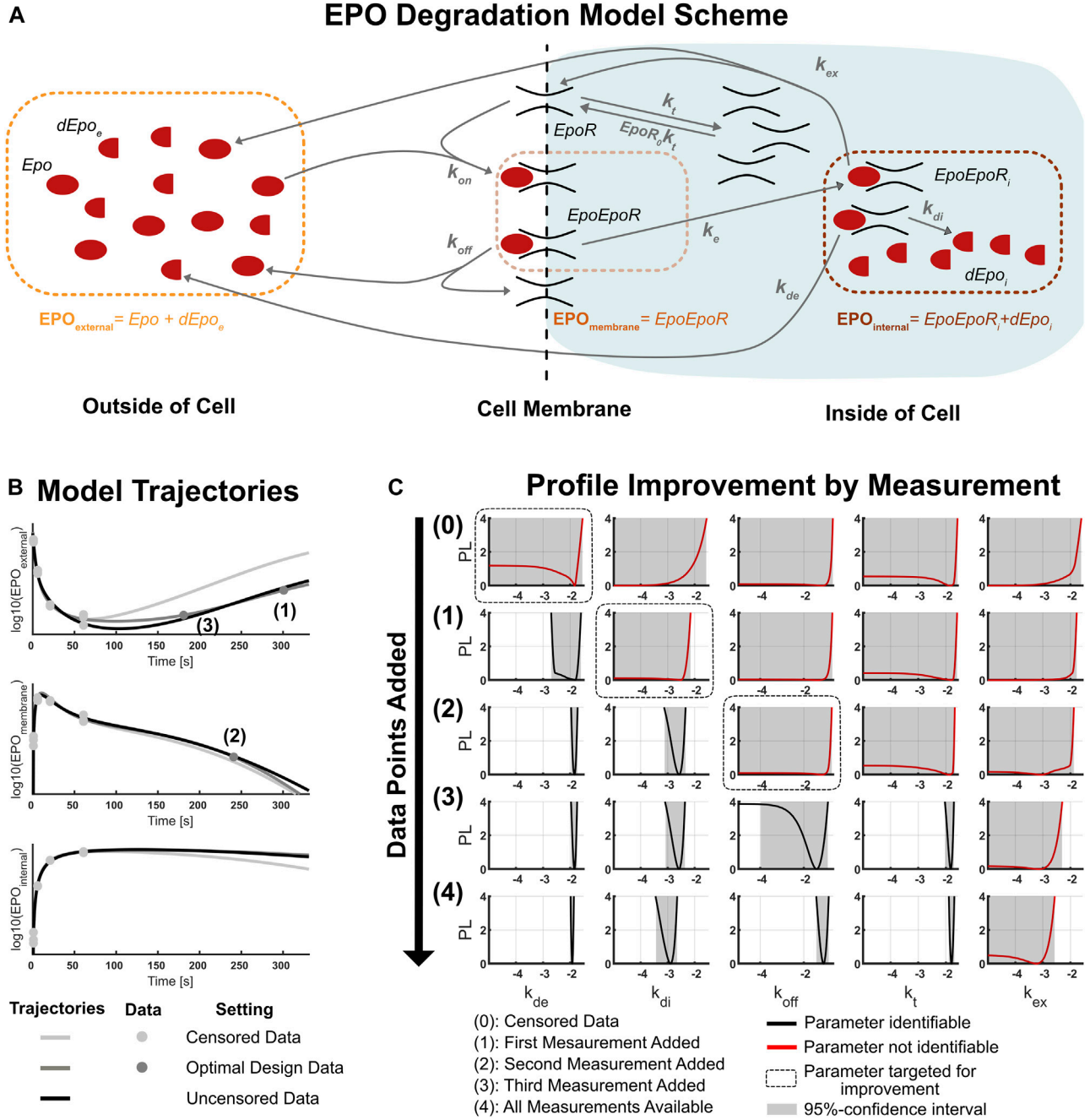


FIGURE 4 | (A): Scheme of the biological dynamics in the EPO degradation model (Becker et al., 2010). There are six model states (black text) which interact through different biological reactions (gray arrows) and three observables (colored text). EPO is transported into the cell and degraded there. **(B)**: Model trajectories for the observables of the EPO-model. The plotted curves are the best fit trajectories for three different data sets: the censored data set used at the start of the experimental design analysis, the data set after adding three sequentially proposed data points, and the uncensored published data set. The numbers indicate the order of the sequentially measured data points. **(C)**: Change of parameter likelihood profiles during the sequential experimental design procedure. The targeted parameter always became identifiable after data for the optimal experimental condition proposed by the two-dimensional likelihood approach was added into the model. Incorporating the three optimal data points into the model already produces results of similar accuracy compared to the published data set with 36 additional data points.

The parameters of the model are identifiable except for one parameter given the complete data set of the study. In order to illustrate experimental design considerations, a model which is not yet well informed by data is required. Thus, we censored one half of the complete data set for all observables which respectively correspond to the later stages of the dynamics. This serves two purposes: First, we reduced the information content of the data, thus creating non-identifiabilities for some parameters. Second,

we gain access to biological data for 3 (observables) x 4 (time points) = 12 experimental conditions which can be used to mimic real measurements.

The best fit model trajectories for the three observables are illustrated in **Figure 4B** for three sets of data (shades of gray). One set of trajectories corresponds to the censored data set, the next set corresponds to an optimal sequential experimental design with three additional measurements and the last set of trajectories corresponds to the original full data set with 3 (observables) x 4 (time points) x 3 (replications) = 36 additional data points. The predictions change significantly when adding the three optimal data points to the censored data set, while adding the rest of the data only changes the model trajectories slightly.

The three optimal data points were determined by applying the workflow for the sequential experimental design scheme shown in **Figure 2**. The iterative improvement of the likelihood profiles of the non-identifiable parameters by this workflow is illustrated in **Figure 4C**. Starting with the censored data set, five parameters are non-identifiable. This comprises the external and internal EPO degradation rate k_{de} and k_{di}, the complex dissociation constant k_{off}, the receptor turnover rate k_t and the complex recycling rate k_{ex}.

The internal EPO degradation rate k_{de} was targeted by the first experiment, and has been made identifiable after measuring $EPO_{external}$ at a late stage of the dynamics. Note that retrospectively, this choice was highly tailored to the identification of k_{de}, as the profile likelihood for the other parameters only changed slightly. This underlines that experiments proposed by our approach aim specifically at improving the knowledge about the targeted parameter of interest.

In the second experiment, the internal EPO degradation rate k_{di} was targeted. The corresponding optimal experiment is a measurement of $EPO_{membrane}$ at a late time point. Because this design is optimal at an earlier time point than the first measurement, this suggests that the first measurement of $EPO_{external}$ already carries information which could have been obtained from measuring $EPO_{membrane}$ at the same time point, highlighting that model dynamics are highly intertwined. Imitating the measurement for the proposed experimental design again shows that the targeted parameter is identifiable after the experiment, while the others are still practically non-identifiable.

The third iteration of experimental design targeted the complex dissociation constant k_{off} and revealed that measuring $EPO_{external}$ at an earlier time point is now more informative than measuring the observable $EPO_{internal}$, for which late time measurements are still not available. This highlights the fact that determination of the optimal experimental design is difficult by intuitive considerations and experimental design approaches provide non-trivial insights. This measurement removed the non-identifiability of both the targeted parameter k_{off} and also the turnover rate k_t which was not considered when planning the experiment.

A fourth iteration of the sequential experimental design was not conducted because the two-dimensional likelihood profiles for the last non-identifiable parameter k_{ex} indicate that a single additional data point for any of the remaining experimental conditions does not provide enough information to make the parameter identifiable. In fact, this is in line with the results of the final model with all data available, as the parameter is still practically non-identifiable given the complete data set. The two-dimensional likelihood profiles corresponding to the four experimental design iterations are illustrated in the **Supplementary Figures S1–S4**.

The comparison of parameter likelihood profiles for the design with three optimally chosen measurements with the full data set design of 36 new data points is shown in the last two rows of **Figure 4C**. The similarity of all profiles across all parameters indicates that three optimally chosen experimental conditions already yield much of the information contained in the set of all 36 data points. This underlines the ability of optimal experimental design to reduce the amount of data needed to remove non-identifiabilities for the parameters of interest. Therefore, application of the optimal sequential experimental design on a realistic biological model demonstrated the feasibility and merits of the two-dimensional likelihood profiles as an approach for experimental design.

4 DISCUSSION

4.1 Experimental Design by Two-Dimensional Likelihood Profiles

A well-planned experiment can save time and resources. Therefore, optimal experimental design aimed at reducing the amount of data needed to inform the model is desirable in any context, but this task is often non-trivial for complex models such as those encountered in systems biology. We established a method for optimal experimental design aiming at reducing parameter uncertainty for a single parameter of interest in a frequentist setting. To this end, we define two-dimensional likelihood profiles which contain information about the likely parameter uncertainty after a measurement. Our approach for experimental design employs the theoretically appealing concept of likelihood profiles, which can serve as a measure for uncertainty in parameter estimates but also for a measure of uncertainty of measurement outcomes. These measures can be conceptually understood to imply confidence densities for parameters or predictive densities for measurement outcomes with strictly frequentist concepts. The presented approach allows for the evaluation of the impact of an experiment in a qualitative as well as in a quantitative manner.

The two-dimensional profile likelihood approach for experimental design was employed in two examples to illustrate its properties and establish feasibility of the method. The ABC reaction model features a non-linear relationship between model states and parameters and served to illustrate the features of two-dimensional likelihood profiles. In order to show practical feasibility of the approach in a realistic setting, an established erythropoietin degradation model (Becker et al., 2010) was investigated. To this end, half of the full data set has been censored to simulate a realistic setting for experimental design in which some model parameters were practically

non-identifiable. A fully sequential experimental design procedure indicated that only 3 of the 36 censored data points were required to successfully remove all possible parameter non-identifiabilities.

4.2 Implementation and Limitations

The numerical implementation is provided as part of the *Data2Dynamics* (Steiert et al., 2019) modeling environment in *MATLAB*. The algorithm exploits the existing one dimensional profile likelihood calculation in order to construct the two-dimensional profile likelihood. Computationally, this amounts to about ~1,000 local optimizations per two-dimensional likelihood profile, where local optimization is to be understood as deterministic optimization from a good initial guess for the parameters. Robustness of these fits is generally easier to obtain if the available data is appropriate for the size of the model, such that the model dynamics are constrained to some degree.

Problems associated with limited data availability go beyond numerical issues and are rooted in the structure of our approach. As a frequentist method, all information used in our experimental design scheme must stem from the data already measured. We have not assessed how much data needs to be initially available before a systematic experimental design procedure is practically feasible. However, the issue of lacking prior knowledge is not exclusive to our approach and a more general theme in non-linear experimental design. For the application in systems biology, initial data is often needed in proposing a suitable model, such that there will usually be data to start off with.

A practical limitation induced by insufficient data occurs if the range of reasonable measurement outcomes can not be predicted by the model, i.e. the validation profile reveals a practical non-identifiability of the model prediction. The existence of this non-identifiability complicates the estimation of the expected parameter uncertainty in the two-dimensional profile likelihood approach, because it relies on the prediction of the measurement outcomes given the model and current data. On the one hand, this fully utilizes the information available in the model, but on the other hand this constrains the applicability of the approach if the prediction for the measurement outcome is insufficiently constrained by the available model data. In case the model prediction of interest is not identifiable, a weak quadratic prior can be added to the validation profile in order to guarantee a finite sample space. This heuristic approach increases the scope of possible application settings. We emphasize that our experimental design procedure works best from a computational as well as methodical point of view if enough data is available such that model predictions are at least loosely constrained.

The usual assumption of the correctness of the model structure is especially important in our proposed method because it utilizes the model for predicting likely outcomes of the experiment and for calculating existing parameter uncertainties. This assumption is usually implicitly contained in any design strategy, but we emphasize that the full exploitation of the likelihood in our approach implies that the proposed experimental design will benefit greatly from solid prior knowledge about the model structure. This does not apply to prior knowledge about model parameters, because likelihood profiles account for parameter uncertainties.

The relationship between confidence intervals and likelihood profiles critically depends on the distributional assumption for the corresponding likelihood profile in **Eq. 6**. The implicit assumption that these likelihood ratios are χ^2_1 distributed for the true parameter set in general holds only asymptotically. However, as this is general practice in the interpretation of likelihood profiles, we follow this procedure and underline that improving upon this assumption offers opportunities for improving the assessment of parameter uncertainties.

4.3 Comparison to Existing Methods

There are two conceptually different methods in the literature which we want to discuss, neglecting approaches based on the Fisher information matrix as reasoned before. One branch of methods deals with a Bayesian approach to experimental design which utilizes the Shannon information of the posterior distribution to plan optimal experiments. The other branch of methods discusses the concepts of frequentist approaches which find experimental designs by sampling relevant regions of the parameter space in order to assess the sensitivity of model predictions with respect to these parameters.

The Bayesian approach (Busetto et al., 2013; Huan and Marzouk, 2013; Liepe et al., 2013) is conceptually similar to our approach, but only applicable if suitable prior parameter distributions are available. The posterior parameter distribution after a possible measurement depends on unknown measurement outcomes which can be resolved by averaging the posterior distribution over the Bayesian predictive density. Similarly, our proposed frequentist method utilizes a predictive density for the measurement outcomes and a confidence density for the parameter estimates, eliminating the need for prior distributions. These "distributional estimators" (Xie and Singh, 2013) are implicitly derived from the likelihood profiles. This theoretical framework suggests the use of confidence and predictive densities in quantifying the confidence that an interval of parameter values or measurement outcomes contains the true parameter value or, respectively, a future measurement outcome.

Our method explicitly determines the impact of different plausible measurement outcomes of an experimental design on the parameter estimate of interest in order to derive a design criterion. This is different to existing frequentist approaches (Bazil et al., 2012; Steiert et al., 2012) which consider the sensitivity of model predictions to the different parameters which are consistent with the current data. Predictions which largely vary under these acceptable parameters indicate experimental conditions which are likely informative as they constrain the set of possible model dynamics. This approach has been awarded as best performing in the DREAM6 challenge (Steiert et al., 2012), although the feedback of the possible measurement results on the model parameter is not considered directly. This hinders intuitive interpretation of how a possible experiment feeds back into the parameter of interest and lacks a quantitative assessment of what constitutes

a large variation of model predictions. Reversing the logic of this approach by considering the impact of likely model predictions on the parameter of interest leads to our refined approach, although this requires a higher computational cost.

4.4 Implications for Research

Our proposed approach can be used to select the most informative experimental design for a targeted parameter of interest. This is often relevant if there are certain biological parameters of interest which are not identifiable given the current data. We want to emphasize that although we discussed reduction of uncertainty for a single target parameter of interest, generalization to reducing the uncertainty for a model prediction, i.e. for a function of model parameters, is straightforward. The detailed quantitative and qualitative information gain by comparing two-dimensional profiles for the different experimental conditions comes with a higher computational cost compared to other approaches. As such, the detailed information provided by our method might be especially useful if experimental measurements require considerable time and resources and as such accuracy is favored over computational efficiency.

The experimental design approach only requires the existence of a suitable likelihood function and is therefore applicable in a broad spectrum of applications. We emphasize the novelty of our approach in employing confidence and predictive distributions as frequentist distributional measures for the confidence in parameter and measurement outcomes, which serve a similar function as Bayesian probabilities. Exploring the interaction of these concepts provides a point of interest for further research in frequentist experimental design.

4.5 CONCLUSION

To summarize, we established an experimental design procedure which aims at reducing the uncertainty for a parameter of interest. This design procedure reduces the likelihood function to a two-dimensional likelihood profile: One dimension informs our confidence of observing a certain measurement outcome for the given experimental condition, while the other dimension informs our confidence in the model parameter corresponding to the underlying true parameter. Testing our experimental design procedure on a simple model with simulated data and on a real model with experimental data revealed that our approach accurately predicted relevant experimental designs. Our method provides detailed information about possible experimental conditions on an easily interpretable quantitative as well as qualitative level.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: doi.org/10.6084/m9.figshare.16863559, https://github.com/Data2Dynamics/d2d.

AUTHOR CONTRIBUTIONS

CK proposed the initial idea of exploiting two-dimensional likelihood profiles for experimental design. TL conceptualized the methods and formalized the presented approach with CK and JT providing input during the process. TL wrote the numerical implementation of the approach and created the results presented in this work. TL wrote a first manuscript draft and all authors revised the manuscript and provided contributions.

FUNDING

This work has been funded by the Federal Ministry of Education and Research of Germany (BMBF) (EA:Sys,FKZ031L0080), by the Federal Ministry of Education and Research of Germany (BMBF) within the Liver Systems Medicine network (LiSyM, Project number 031L0048), by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project number 272 983 813-TRR 179 and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germanys Excellence Strategy (CIBSS-EXC-2189-2100 249 960-390 939 984). The article processing charge was funded by the Baden-Württemberg Ministry of Science, Research and Art and the University of Freiburg in the funding programme Open Access Publishing.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.800856/full#supplementary-material

REFERENCES

Atkinson, A., and Donev, A. (1992). *Optimum Experimental Designs*. New York: Oxford University Press.

Bazil, J. N., Buzzard, G. T., and Rundell, A. E. (2012). A Global Parallel Model Based Design of Experiments Method to Minimize Model Output Uncertainty. *Bull. Math. Biol.* 74, 688–716. doi:10.1007/s11538-011-9686-9

Becker, V., Schilling, M., Bachmann, J., Baumann, U., Raue, A., Maiwald, T., et al. (2010). Covering a Broad Dynamic Range: Information Processing at the Erythropoietin Receptor. *Science* 328, 1404–1408. doi:10.1126/science.1184913

Bishop, T., Petersen, B., and Trayser, D. (1982). Another Look at the Statistician's Role in Experimental Planning and Design. *Am. Stat.* 36, 387–389. doi:10.1080/00031305.1982.10483058

Bjornstad, J. F. (1990). Predictive Likelihood: A Review. *Stat. Sci.* 5, 242–254. doi:10.1214/ss/1177012175

Busetto, A. G., Hauser, A., Krummenacher, G., Sunnåker, M., Dimopoulos, S., Ong, C. S., et al. (2013). Near-optimal Experimental Design for Model Selection in Systems Biology. *Bioinformatics* 29, 2625–2632. doi:10.1093/bioinformatics/btt436

Chis, O.-T., Villaverde, A. F., Banga, J. R., and Balsa-Canto, E. (2016). On the Relationship between Sloppiness and Identifiability. *Math. Bioscie.* 282, 147–161. doi:10.1016/j.mbs.2016.10.009

Cochran, W. G. (1973). Experiments for Nonlinear Functions (R.A. Fisher Memorial Lecture). *J. Am. Stat. Assoc.* 68, 771–781. doi:10.1080/01621459.1973.10481423

Faller, D., Klingmüller, U., and Timmer, J. (2003). Simulation Methods for Optimal Experimental Design in Systems Biology. *Simulation* 79, 717–725. doi:10.1177/0037549703040937

Fedorov, V. (2010). Optimal Experimental Design. *WIREs Comp. Stat.* 2, 581–589. doi:10.1002/wics.100

Ford, I., Titterington, D. M., and Kitsos, C. P. (1989). Recent Advances in Nonlinear Experimental Design. *Technometrics* 31, 49–60x. doi:10.1080/00401706.1989.10488475

Huan, X., and Marzouk, Y. M. (2013). Simulation-based Optimal Bayesian Experimental Design for Nonlinear Systems. *J. Comput. Phys.* 232, 288–317. doi:10.1016/j.jcp.2012.08.013

Kitano, H. (2002). Systems Biology: A Brief Overview. *Science* 295, 1662–1664. doi:10.1126/science.1069492

Kreutz, C., and Timmer, J. (2009). Systems Biology: Experimental Design. *FEBS J.* 276, 923–942. doi:10.1111/j.1742-4658.2008.06843.x

Kreutz, C., Raue, A., and Timmer, J. (2012). Likelihood Based Observability Analysis and Confidence Intervals for Predictions of Dynamic Models. *BMC Syst. Biol.* 6, 120. doi:10.1186/1752-0509-6-120

Liepe, J., Filippi, S., Komorowski, M., and Stumpf, M. P. H. (2013). Maximizing the Information Content of Experiments in Systems Biology. *Plos Comput. Biol.* 9, e1002888. doi:10.1371/journal.pcbi.1002888

Limpert, E., Stahel, W. A., and Abbt, M. (2001). Log-normal Distributions across the Sciences: Keys and Clues. *BioScience* 51, 341–352. doi:10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

Lindley, D. V. (1956). On a Measure of the Information provided by an experiment. *Ann. Math. Statist.* 27, 986–1005. doi:10.1214/aoms/1177728069

Meeker, W. Q., and Escobar, L. A. (1995). Teaching about Approximate Confidence Regions Based on Maximum Likelihood Estimation. *Am. Stat.* 49, 48–53. doi:10.1080/00031305.1995.10476112

Nurse, P., and Hayles, J. (2011). The Cell in an Era of Systems Biology. *Cell* 144, 850–854. doi:10.1016/j.cell.2011.02.045

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., et al. (2009). Structural and Practical Identifiability Analysis of Partially Observed Dynamical Models by Exploiting the Profile Likelihood. *Bioinformatics* 25, 1923–1929. doi:10.1093/bioinformatics/btp358

Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelke, M., Kaschek, D., et al. (2013). Lessons Learned from Quantitative Dynamical Modeling in Systems Biology. *PloS ONE* 8, e74335. doi:10.1371/journal.pone.0074335

Steiert, B., Raue, A., Timmer, J., and Kreutz, C. (2012). Experimental Design for Parameter Estimation of Gene Regulatory Networks. *PloS one* 7, e40052. doi:10.1371/journal.pone.0040052

Steiert, B., Kreutz, C., Raue, A., and Timmer, J. (2019). "Recipes for Analysis of Molecular Networks Using the Data2dynamics Modeling Environment," in *Modeling Biomolecular Site Dynamics: Methods and Protocols*. Editor W. S. Hlavacek (New York, NY: Springer New York), 341–362. doi:10.1007/978-1-4939-9102-0_16

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA Statement on P-Values: Context, Process, and Purpose. *Am. Stat.* 70, 129–133. doi:10.1080/00031305.2016.1154108

Wieland, F.-G., Hauber, A. L., Rosenblatt, M., Tönsing, C., and Timmer, J. (2021). On Structural and Practical Identifiability. *Curr. Opin. Syst. Biol.* 25, 60–69. doi:10.1016/j.coisb.2021.03.005

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Ann. Math. Statist.* 9, 60–62. doi:10.1214/aoms/1177732360

Xie, M.-g., and Singh, K. (2013). Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review. *Int. Stat. Rev.* 81, 3–39. doi:10.1111/insr.12000

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Litwin, Timmer and Kreutz. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Molecular Biosciences

METHODS
published: 02 March 2022
doi: 10.3389/fmolb.2022.800152

UPMaBoSS: A Novel Framework for Dynamic Cell Population Modeling

Gautier Stoll[1,2], *Aurélien Naldi*[3,4], *Vincent Noël*[5,6,7], *Eric Viara*[8], *Emmanuel Barillot*[5,6,7], *Guido Kroemer*[1,2,9], *Denis Thieffry*[3,4] *and Laurence Calzone*[5,6,7]*

[1]Equipe Labellisée Par La Ligue Contre Le Cancer, Université de Paris, Sorbonne Université, INSERM UMR1138, Centre de Recherche des Cordeliers, Paris, France, [2]Metabolomics and Cell Biology Platforms, Gustave Roussy Cancer Campus, Université Paris Saclay, Villejuif, France, [3]Institut de Biologie de L'ENS (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, Paris, France, [4]Lifeware Group, Inria Saclay-Ile de France, Palaiseau, France, [5]Institut Curie, PSL Research University, Paris, France, [6]INSERM U900, Paris, France, [7]MINES ParisTech, CBIO-Centre for Computational Biology, PSL Research University, Paris, France, [8]Sysra, Yerres, France, [9]Pôle de Biologie, Hôpital européen Georges Pompidou, AP-HP, Paris, France

Mathematical modeling aims at understanding the effects of biological perturbations, suggesting ways to intervene and to reestablish proper cell functioning in diseases such as cancer or in autoimmune disorders. This is a difficult task for obvious reasons: the level of details needed to describe the intra-cellular processes involved, the numerous interactions between cells and cell types, and the complex dynamical properties of such populations where cells die, divide and interact constantly, to cite a few. Another important difficulty comes from the spatial distribution of these cells, their diffusion and motility. All of these aspects cannot be easily resolved in a unique mathematical model or with a unique formalism. To cope with some of these issues, we introduce here a novel framework, UPMaBoSS (for Update Population MaBoSS), dedicated to modeling dynamic populations of interacting cells. We rely on the preexisting tool MaBoSS, which enables probabilistic simulations of cellular networks. A novel software layer is added to account for cell interactions and population dynamics, but without considering the spatial dimension. This modeling approach can be seen as an intermediate step towards more complex spatial descriptions. We illustrate our methodology by means of a case study dealing with TNF-induced cell death. Interestingly, the simulation of cell population dynamics with UPMaBoSS reveals a mechanism of resistance triggered by TNF treatment. Relatively easy to encode, UPMaBoSS simulations require only moderate computational power and execution time. To ease the reproduction of simulations, we provide several Jupyter notebooks that can be accessed within the CoLoMoTo Docker image, which contains all software and models used for this study.

Keywords: stochastic simulation, cell interactions, heterogeneous cell population, logical model, pathway modeling

OPEN ACCESS

Edited by:
Alberto Jesus Martin, Universidad Mayor, Chile

Reviewed by:
Paola Lecca, Free University of Bozen-Bolzano, Italy
Robert Sinkovits, University of California, United States

***Correspondence:**
Laurence Calzone
laurence.calzone@curie.fr

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *22 October 2021*
Accepted: *21 January 2022*
Published: *02 March 2022*

Citation:
Stoll G, Naldi A, Noël V, Viara E, Barillot E, Kroemer G, Thieffry D and Calzone L (2022) UPMaBoSS: A Novel Framework for Dynamic Cell Population Modeling. Front. Mol. Biosci. 9:800152. doi: 10.3389/fmolb.2022.800152

1 INTRODUCTION

One of the key challenges in cell biology or in biochemistry is to understand how perturbations of genes, proteins or metabolites affect cellular behavior. For that, the construction and analysis of mathematical models constitute a powerful approach, which can help formalize and reason on the complex phenomena governing the functioning of the cell. At the cellular level, the relations between single entities can be described as signaling, biochemical, or metabolic pathways, and transcribed into

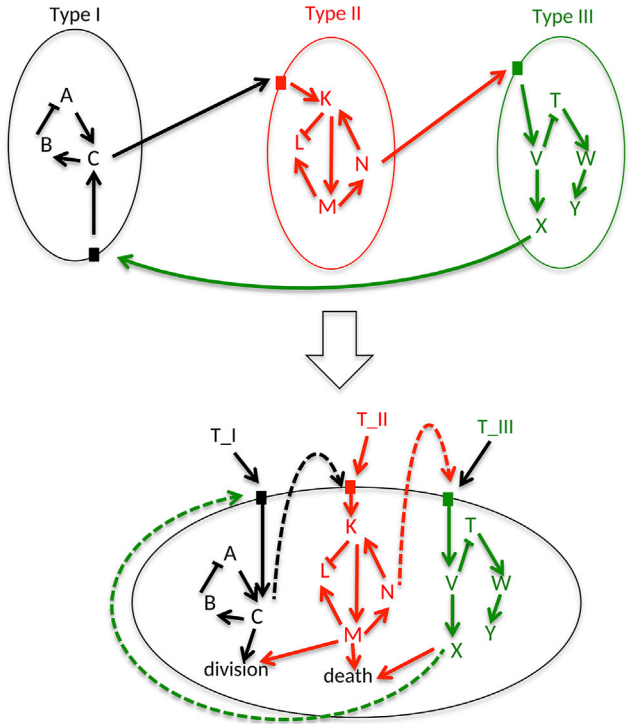


FIGURE 1 | Generic cell. A generic cellular network is constructed by assembling all the signaling pathways that can be activated in different cell types (here 3 cell types are considered: Type I, II, and III), including ligand-receptor interactions. Cells can die, divide or interact through ligand-receptor interactions.

mathematical terms to predict the impact of specific perturbations on cellular processes. The difficulty grows when considering inter-cellular signals and their effect at the organ level or the role of the micro-environment on the cell fate. Ideally, the mathematical model should include not only detailed pathway descriptions for every cell, but also key events occurring at the population level, e.g., extra-cellular diffusion, cell motility, inter-cellular communications, death or division of cells.

Multi-cellular systems models have already been studied in developmental biology (Cartwright et al., 2009). Most often, cell populations are considered as a large set of single entities (cells), diffusing and moving throughout the environment, giving rise to reproducible spatial organizations. Formal frameworks borrowed from physics are often used, such as partial differential equations. These approaches allow detailed and accurate temporal and spatial descriptions of collective cellular behaviors (Cowan et al., 2012).

In the field of cancer and immunology, many published models considered a generic cell (e.g., an epithelial cell, a T lymphocyte cell, a macrophage, etc.) to represent the behavior of cell populations. An extension of this simplistic view relies on stochastic simulations to estimate the evolution of desynchronized cell populations. Although such studies do not consider explicitly individual cells (Shmulevich et al., 2002; Albert et al., 2008; Stoll et al., 2017), the outputs of these simulations can be interpreted as the composition of a population of non-interacting cells.

In both cancer (Anderson et al., 2006) or auto-immune diseases (El-Badri et al., 2007), there are different cell types that need to be considered. In this respect, agent-based approaches associate an agent with each cell, which activity depends on that of its neighbors (Bonabeau, 2002; Altinok et al., 2011). Powerful tools have been developed to define and analyze such models, including CellSys (Drasdo and Hoehme, 2010), CompuCell3D (Swat et al., 2012) or PhysiCell (Ghaffarizadeh et al., 2018).

However, these models usually do not explicitly take into account intra-cellular signaling pathways, or the specific deregulations that may occur in these signaling pathways for good reasons. First, adding some internal dynamics to each agent increases considerably the level of complexity and the number of parameters to tune. Second, it can be computationally costly depending on the formalism used for modeling the signaling pathways inside each agent, such as ordinary differential equations (ODEs) (Clairambault, 2006) or logical formulae (Letort et al., 2018; Varela et al., 2019).

Integrating a proper description of intra-cellular pathways to agent-based models is a challenge, but it becomes crucial when studying the response to a drug that targets specific pathways, even though the effects are often observed at the level of the population (e.g., survival). Indeed, signaling pathways are organized in complex networks encompassing numerous cross-talks and feedbacks. Hence, the deregulation of one specific pathway often leads to non-intuitive effects at the population level.

Mathematical modeling of such complex and intricate networks can help understand and predict experimental results (Cohen et al., 2013; Ferrell, 2015; Kolch et al., 2015; Remy et al., 2015; Abou-Jaoudé et al., 2016). However, the choice of the most appropriate mathematical formalism to model such intra-cellular processes depends on the biological question and the available data (Le Novère, 2015).

Here, we present UPMaBoSS (Update Population MaBoSS), a modeling framework focusing on the dynamics of populations of interacting cells and based on stochastic simulations of a discrete model.

As MaBoSS (Stoll et al., 2012; Stoll et al., 2017)), UPMaBoSS relies on a logical formalism applied to signaling pathways and regulatory circuits. Logical models can be viewed as a coarse grain approximation of more refined and precise modeling approaches, and it proved to be efficient in several studies for which details of the chemical reactions is poorly known (Collombet et al., 2017; Eduati et al., 2017). In this framework, each cell type has its own dynamics and relies on a qualitative model of intra-cellular signaling networks using MaBoSS grammar, thereby enabling probabilistic simulations of cellular models. UPMaBoSS alternates the simulation of intra-cellular models with regular updates of estimates of cell population sizes and environmental signals. These updates are based on the values of key network nodes (receptors and ligands) and processes accounting for cell division and cell death. The whole population can be considered

as an heterogeneous cell population, with different cell types, or with cells in different states.

The extension of pathway models constructed in MaBoSS is easy, and the additional computational cost of simulations remains comparable to that of simulations of original MaBoSS models. UPMaBoSS enables the description of complex cellular networks encompassing relatively large numbers of components (up to a few hundreds), interacting through positive and negative influences. Some previous works have already developed a similar algorithm, but based on chemical kinetics (Charlebois et al., 2011; Charlebois and Kærn, 2013).

UPMaBoSS can be used to address biological questions involving interactions between cell types, when signaling pathways are known. As for many approaches considering intra-cellular details, the main application of UPMaBoSS is to explore the effects of perturbations that occur at the level of individual cells in order to understand how these perturbations impact cell populations. Such explorations constitute a first step towards the use of multi-scale modeling for clinical applications (Wolkenhauer et al., 2014; Viceconti and Hunter, 2016).

This modeling approach is illustrated with a simple toy model of cell-cell interaction, and with a case study dealing with the effect of the Tumor Necrosis Factor (TNF) on the cell fate decisions triggered by the engagement of death receptors.

To foster the reproducibility of our results, we provide several notebooks in a dedicated GitHub at https://github.com/sysbio-curie/UPMaBoSS-docker, with examples of models in MaBoSS language, as well as in the more standard SBML-Qual format (Chaouiya et al., 2013). Noteworthy, published logical models of cellular networks (Helikar et al., 2012; Naldi et al., 2018) can be easily adapted and extended by adding cell death, division, and cell-cell interactions. The definition of an UPMaBoSS model then enables the integration of these effects at the population level.

2 MATERIALS AND METHODS

MaBoSS is a tool to simulate continuous time Markov processes on Boolean networks (Stoll et al., 2012; Stoll et al., 2017). It was built as a middle term between the detailed yet complex description of signaling pathways using a chemical kinetics approach and the simpler coarse-grain description using a discrete formalism. To compare these three views (ODE, logical and MaBoSS), models for two network motifs are provided in the **Supplementary Material** (**section 4**).

Models of single cell types can each be built using MaBoSS. However, creating a comprehensive model of a dynamic population of these cell types requires following a protocol that can be summarized in two steps: 1) the construction of a single network that encompasses all pathways in all cell types, and 2) the implementation of this network within UPMaBoSS framework. For the latter part, details on the algorithm of UPMaBoSS are given in the **Supplementary Section S1.2**). For the practical aspects of the UPMaBoSS model construction, we describe the procedure below.

2.1 Connecting Cell Population Pathways Into a Single Influence Network

The starting point for an UPMaBoSS model is a set of influence networks that each represents signaling pathways of the cell type they describe. These pathways should contain receptors that are activated by ligands to model properly the interactions between the cell types. The translation of these networks into an UPMaBoSS model requires the following steps, recapitulated in **Figure 1**:

1 Collect all networks of each cell type and integrate them into a single network (lower panel in **Figure 1**). Note that it is possible that the obtained integrated network is disconnected. In the case that two pathways belonging to 2 cell types have identical entities, e.g., TGFb, there are two possibilities: either the two networks are merged through the entity or the entity is renamed to specify in which cell type it belongs (i.e., TGFb_Tumor and TGFb_TCell, see **Supplementary Section S1.6** for a detailed explanation).
2 Add a node for each cell type (T-Cell, Tumor, etc.), and connect them to the appropriate entities in their associated pathways. For instance, for a protein A activated by B and C in tumor, the network will have ProtB → ProtA, ProtC → ProtA and also Tumor → ProtA.
3 Add two nodes, one that represents cell division and another one that accounts for cell death. The readouts of the model that are related to these phenotypes can then be connected to these new nodes; for example, CyclinB can be connected to division and caspase three to death.

As a result, all possible signaling pathways described in all cell types are gathered into a unique UPMaBoSS model, where the extracellular interactions are represented a feedback loops (**Figure 1**). This unique network could be understood either as an undifferentiated cell ready to be differentiated into different cell types, or as a population composed of different cell types. This formalism is very flexible and allows the characterization of various cell types present in the micro-environment, as illustrated in the examples in **section 3** of the **Supplementary Material**.

2.2 Constructing a Model in UPMaBoSS From an Influence Network

UPMaBoSS relies on MaBoSS software and extends it by adding some important functionalities to model interacting cell populations (Stoll et al., 2012; Stoll et al., 2017).

Every population state is represented by a set of Boolean values associated with the network nodes (including the *division* node, the *death* node, the ligand and the receptors, cell type nodes, etc.). The UPMaBoSS framework enables the computation of the time-dependent probabilities of states, allowing an interpretation of the dynamics at the cell population level, considering that:

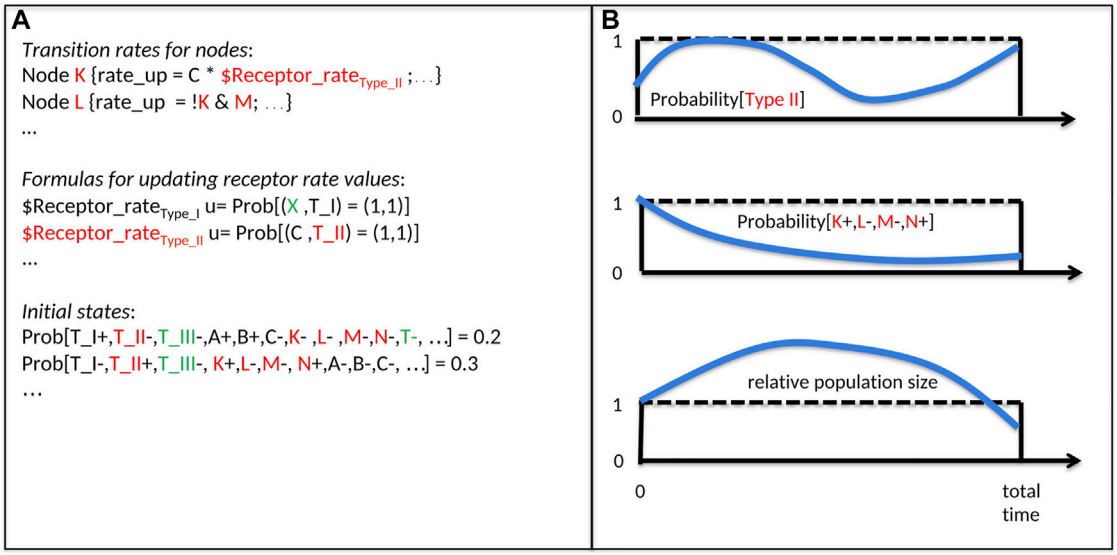


FIGURE 2 | Inputs and Outputs of an UPMaBoSS model. The notation is related to **Figure 1**: A, B, C, K, L, M represent genes/proteins; T_I, T_II, T_III represent cell types. **(A)** Inputs of UPMaBoSS: Transition rates for nodes: for each node (here K and L of **Figure 1**), a logical rule, the rate up and rate down are written; Formulas for updating receptors rates values: the update rules, starting by u = . . ., depend on the population state and regulate the value of the external variable $Receptor_rate of cell type I and II; and initial states: they can be defined such that cell types, proteins, etc. can be characterized as present (+) or absent (–) with a probability for this model state to be active initially. Colors correspond to cell types of **Figure 1**. Note that names starting with a $ correspond to external variables, specific to MaBoSS/UPMaBoSS, listed in bnd file, set up in cfg file and updated in upp file. **(B)** Ouputs of UPMaBoSS: time-dependent probabilities of cell types (upper panel, example of cell type II from **Figure 1**), with the corresponding model states (middle panel), and the time-dependent population size (lower panel).

$$\text{(Number of cells in state S)} = \text{(Number of cells)} \times \text{(Probability of S)}$$

The definition of the model components assumes that each node of the network represents a gene, protein, complex, phenotype or cell type. The UPMaBoSS model gathers the pathways and entities potentially active in the different cell types. It is derived from an influence network by applying the following procedure:

- For every node, except receptors, two transition rates are defined, rate_up for activation and rate_down for inhibition, with triggering rules formulated in the language of MaBoSS (see the Results section for an example of the grammar used for writing logical rules and transition rates).
- For receptors, the rates associated with the update of their state must contain term(s) that depend on the population state probabilities.
- The initial conditions for all the entities of the model need to be set with the same formalism as in MaBoSS: a probability can be associated with each node of the model (e.g., [A]. istate = 0.2 [0], 0.8 [1] means that the node A will start with 80% of the trajectories with A in state 1) or to a vector of nodes (e.g., [A,B]. istate = 0.2 [0,0], 0.5 [0,1], 0.1 [1,0], 0.2 [1,1] means that 20% of the trajectories will start with both nodes at 0, 50% with B active only, 10% with A active only and 20% with both nodes active).

For each time step, UPMaBoSS computes the relative population size (with respect to the initial size), and the distribution of state probabilities in the population (including the status of *death*, *division*, and cell type nodes) (**Figure 2**).

UPMaBoSS launches several consecutive MaBoSS runs (the number of runs being defined by the user). At the end of each run, the population is updated synchronously: new model states are produced according to the parameters influencing the population status (death, division, receptor activity), setting a novel initial condition for the next MaBoSS run (see **section 1** of **Supplementary Material** for more details).

2.3 Tuning Parameters

An UPMaBoSS model contains a set of parameters that need to be calibrated. They can be separated into two families: parameters with a biological interpretation versus those modulating the simulations.

The first family of parameters include the rates of activation or inactivation of a variable, which can be derived from experiments. They can correspond to the mean time necessary to achieve transcription, (de)phosphorylation, synthesis or degradation. These transition rates can be separated into fast or slow variables. If such information is not available, the default value 1.0 is used. Other parameters can account for the initial conditions or the duration of the experiment (total time of simulation is reached when (total time) = (number of steps) × (length of MaBoSS run)).

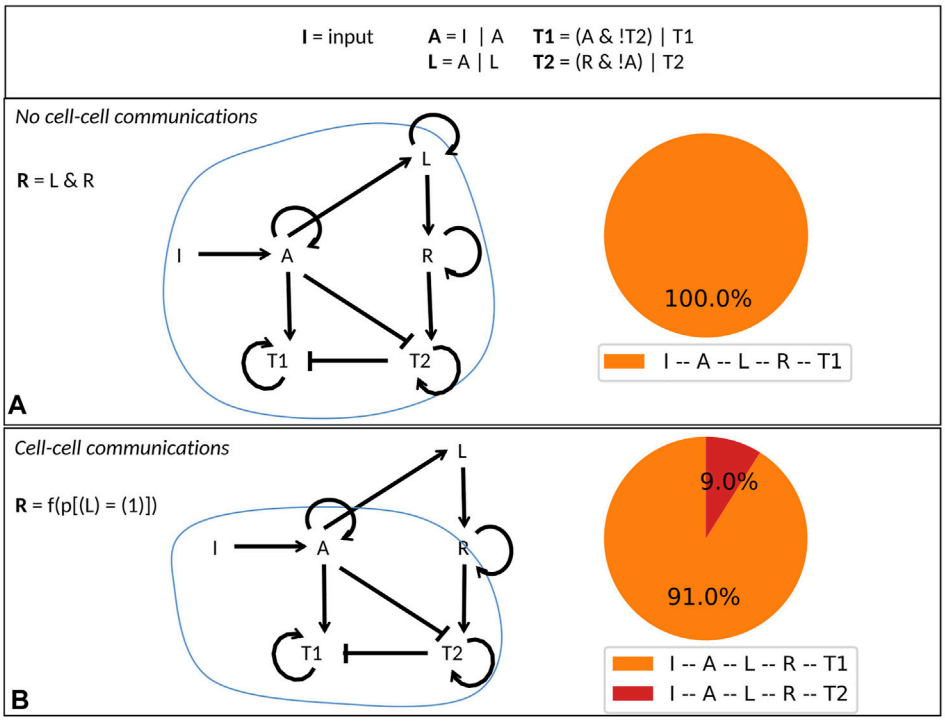


FIGURE 3 | Simple cell differentiation model. **(A)** Definition of the toy model with logical rules (upper panel) and conditional rule for R depending on the value of the external parameter innerOn. If innerOn is equal to 1, then A is able to activate L in all cells (middle panel). If innerOn is set to 0, then the value of R will depend on the population status of L (lower panel). **(B)** Model simulations of the two cases: when innerOn = 1, only T1 cell type can be reached; when innerOn = 0, a proportion of cells can differentiate into T2 cell type.

The second family of parameters corresponds to the number of trajectories to include in the computation, and the length of one MaBoSS run (ensuring that transitory behaviors are observable, see an example below).

An exhaustive list of these parameters is provided in the **Supplementary Section S1.5**, including default values and guidelines for choosing their values. In some cases, a sensitivity analysis may be needed to select the appropriate range of parameter values. We provide an example with a variation of the length of one MaBoSS run (*max_time*) and analyze its impact on the expected results in a python notebook included in the GitHub folder (https://github.com/sysbio-curie/UPMaBoSS-docker/tree/master/CellFate).

3 RESULTS

We illustrate the use of UPMaBoSS with two examples, the first one focusing on cell-cell interactions, and the second one extending a published model of cell fate decision in response to the activation of death receptors (Calzone et al., 2010).

3.1 Toy Model

First, we present a simplified model to highlight important dynamical differences occurring when considering a unique cell model to represent a population of non interacting cells (MaBoSS simulation) versus considering a dynamic population of interacting cells (UPMaBoSS simulation). This simple model implements a differentiation mechanism leading to two cell types, $T1$ and $T2$. This differentiation process is initiated by a trigger I. When I activates A, A is able to drive the $T1$ cell type differentiation. At the same time, A is also able to activate the ligand L, which itself activates a receptor R leading to $T2$ cell type differentiation, but only in the absence of A. The activation of $T2$ is then irreversible (no degradation rate).

We compare the behavior of a population of independent cells with that of a population of interacting cells (through ligands and receptors). The difference between these two situations lies in the logical rule associated with the receptor variable R ($R = innerOn\ ?\ L : outerL$ in the model file, which reads as if $innerOn = 1$ then L else $outerL$). In the case of a single cell model, when cell-cell interaction is not considered, R is activated by L ($innerOn = 1$ in the parameter file) (**Figure 3A**). In the case of a population model, when cells can interact, R is activated by a function of the probability for L to be active ($innerOn = 0$ and $outerL = 5{\star}p[(L) = (1)]$ defined in the population file) (**Figure 3B**).

When a population of independent cells is considered, A is always present when R is active, which continuously inhibit $T2$, ultimately leading to the $T1$ phenotype (**Figure 3A**).

When a population of interacting cells is considered, R is updated according to the population state of L. Consequently, R can be activated in some cells independently of the activity of A,

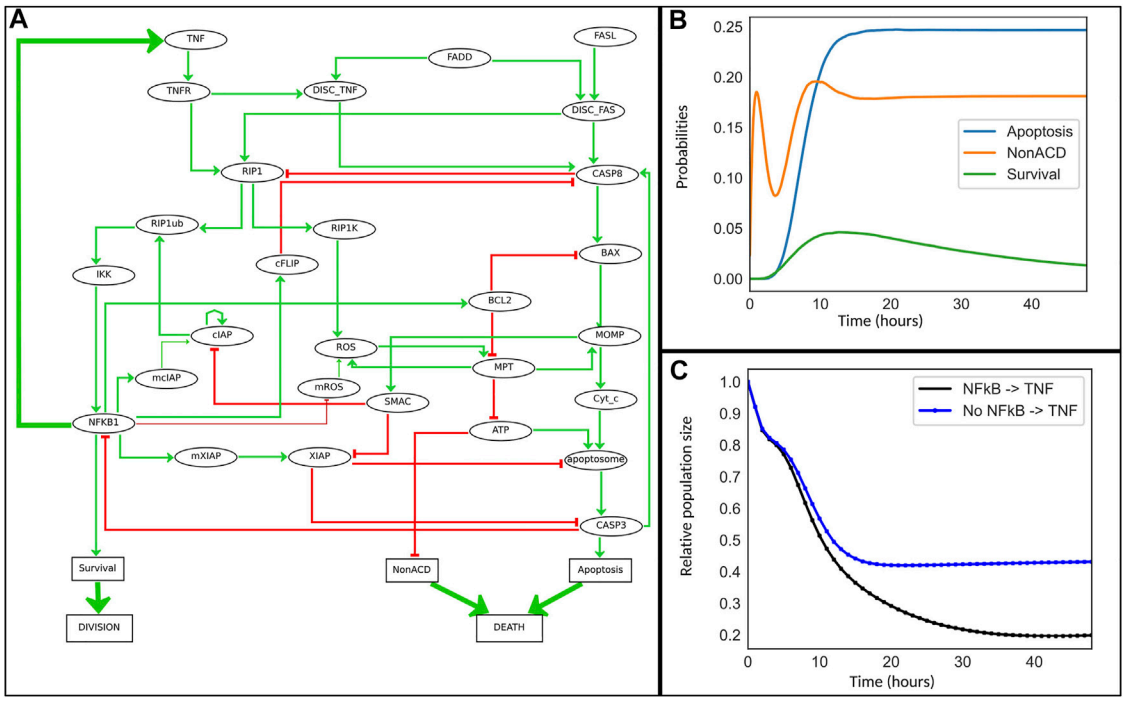


FIGURE 4 | Cell fate model for TNFα resistance. **(A)** This model is an extension of the model reported in (Calzone et al., 2010). Some nodes representing the mRNA of cIAP, ROS and XIAP family members have been added. The ellipsoid nodes represent genes, mRNA, proteins, or complexes, while the rectangular nodes denote phenotypes. Green and red arrows represent positive and negative influences, respectively. The thick green arrows denote activating interactions added to the initial model: a feedback from NFκB to TNFα encodes the ligand-receptor activation, while the "Division" and "Death" nodes have been introduced specifically for UPMaBoSS population updates. **(B)** Simulation of the cell fate model with MaBoSS for 48 h. **(C)** Simulation of the model with UPMaBoSS: temporal evolution of population sizes with (black) and without (blue) the TNF paracrine signaling.

allowing differentiation in the *T*2 cell type with a non-zero probability (**Figure 3B**). In other words, with *I* active, we obtain about 87% of T1 cells (where *I*, *A*, *L*, *R*, and *T*1 are active) and 13% of T2 cells (where *I*, *A*, *L*, *R*, and *T*2 are active). *A* can eventually reactivate in those cells, but we assumed the T2 phenotype to be irreversible (self loop).

This simple example highlights a mechanism occurring only when cells are allowed to interact. When integrating data into a model, fitting parameters, or constructing a Boolean model from experimental data, these considerations might be of importance.

We provide a Jupyter notebook for this example, including the UPMaBoSS model (with the three corresponding files) and an explanation of how to build an interacting cell population model from a standard cell model [defined with the bnet format (Müssel et al., 2010)] (see GitHub at https://github.com/sysbio-curie/UPMaBoSS-docker/tree/master/ToyModelUP).

3.2 A Model of Cell Fate Decision

This case study provides an example of the use of UPMaBoSS to model the response of a cell population to different drug treatments. In this respect, the model integrates pathways controlling cell proliferation and death, which has a direct impact on the size of the cell population.

3.2.1 Description of the Model

We start with a model initially built to understand how the same signal can lead to three different cell phenotypes depending on the cellular context. The activation of the death receptors TNFR or Fas by their respective ligands can trigger a cascade of events leading to either survival with the activation of NFκB pathway, or to non-apoptotic cell death (referring to necrosis or to a programmed necrosis called necroptosis) with the loss of ATP, or to apoptosis with the cleavage of caspase 3 (Calzone et al., 2010). This generic model was built on the basis of data collected from the literature, focusing on the main components influencing the cell fate decision between death and survival. The model can be found in the GINsim repository: http://ginsim.org/node/227. The original analysis explored which components contribute to each phenotype, as well as the cross-talks between the three pathways, enforcing mutual exclusion of the three alternative fates.

In the present study, we extend this analysis by considering the impact of the timing and duration of TNF treatments. To this end, a feedback from NFκB pathway to TNFα was added to the model (**Figure 4A**). Indeed, it has been showed that TNFα is a target of NFκB, and that constitutive activation of NFκB leads to systemic inflammation through TNFα activation (Shakhov et al., 1990; Drouet et al., 1991; Liu et al., 2000; Dong et al., 2010). We

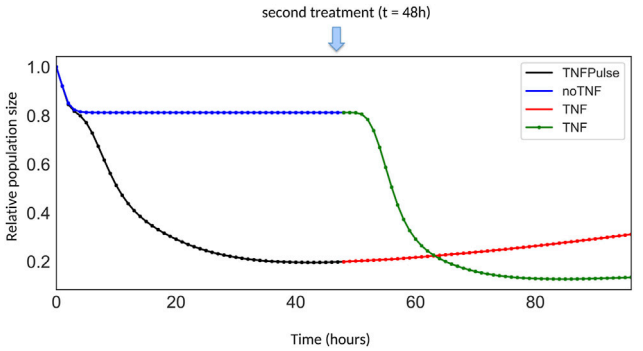


FIGURE 5 | Growth curves for different TNF treatment scenarios. The first scenario corresponds to the simulation of cells initially treated by a pulse of TNF (black segment), followed by a constitutive TNF treatment at $t = 48$ h (red segment). The other scenario corresponds to the simulation of cells initially untreated (blue segment), but receiving a constitutive TNF treatment at $t = 48$ h (green segment).

further decided to focus on the role of TNFα and thus kept Fas OFF for all our simulations.

We explored the effect of sequential treatments of TNFα at a cell population level. Interestingly, several studies showed that prolonged treatments of low doses of TNF can lead to resistance in prostate cancer patients (Smyth et al., 2004), and that TNF exhibits a dual impact on tumor progression: at low doses, it triggers angiogenesis (Wang et al., 2017), whereas at high doses, it induces cell death, mainly through necrotic effects (Bertazza and Mocellin, 2010).

3.2.2 Biological Questions

In a first *in silico* experiment, we simulated the model for a period corresponding to 48 h of cell culture (unit time set to 1 hour). Experimentally, there is no consensus for time duration of TNF effect *in vitro*; nevertheless, key events are known to require over 24 h to occur (Udommethaporn et al., 2016), which justifies the choice of 48 h. Indeed, in our simulations (**Figure 4**), stability is reached after 48 h. For a transient treatment of TNF, we considered a TNF half-life of 4 h (degradation rate of 1/4). Although TNF degradation rate varies extensively depending on experimental conditions, 4 h seems a reasonably small interval compared to 48 h.

We first simulated the model with MaBoSS, which is particularly important in order to define the time step for each MaBoSS run when using UPMaBoSS. Indeed, the chosen time window (i.e., max_time) must be such that the population is in a transient state. The best value for the parameter max_time turned out to be around 1 h (just before the peak of activation of **Figure 4B**). To simulate the population dynamics, we proceeded to compute 48 MaBoSS runs with max_time equal to one for each run. We further studied the sensitivity to this parameter and included online (Jupyter notebook *TimeStepDependency.ipynb*, https://github.com/sysbio-curie/UPMaBoSS-docker/tree/master/CellFate).

In the MaBoSS simulation, we noticed that non-apoptotic cell death first decreases, before increasing to reach a steady state solution after $t = 15$ h. This dynamics is due to the activity of ATP, itself dependent on that of RIP1. RIP1 increases until CASP8 is activated and able to inhibit it. It takes longer to activate CASP8 than RIP1. This behavior is typical of incoherent feedforward loops (Jin, 2013).

For this biological application, we focused on two questions:

1 What is the effect of the feedback from NFκB to TNFα at the population level when treated by a transient activation of TNFα?
2 What is the effect of TNFα sequential treatments on the population dynamics?

To address the first question, two model variants were considered: with and without the NFκB → TNF paracrine loop, with a transient TNF treatment (**Figure 4C**). For the second question, we selected the model with the paracrine loop and further studied the two following scenarios in which cells are treated with: 1) a transient TNF treatment at time 0 ("TNF Pulse"), followed by a constant TNF treatment at 48 h ("TNF") and 2) no TNF treatment at time 0 ("NoTNF"), followed by a constant TNF treatment at 48 h ("TNF") (**Figure 5**).

3.2.3 TNF Treatments in Wild Type Conditions

Simulations of the temporal evolution of cell populations are displayed in **Figure 4C** in presence or absence of the feedback and in **Figure 5** for two different TNF treatment scenarios.

In **Figure 4C**, following a pulse of TNF at $t = 0$, the comparison of population growth curves in the absence (blue curve) or in the presence (black curve) of the feedback from NFκB to TNFα indicates that the TNF paracrine loop leads to a decrease of the population size (from 43% to 20% of the initial population size of 100%) confirming the stronger effect of the feedback in TNF-treated cells.

Remarkably, when a sustained treatment is applied, the impact on the population size differs depending on whether the population has already been treated or not, in a non-intuitive way. Indeed, as shown in **Figure 5**, after 48 h, the population initially untreated (blue + green curves) decreases faster than the cell population initially treated with a pulse of TNF (black + red curves). This difference could be interpreted as a resistance mechanism: cells that have already been exposed to TNF, even transiently, do not respond as well to a second TNF treatment compared to cells that have never been treated with TNF. Noteworthy, this "resistance" results purely from network dynamics, in the absence of any genetic modifications and could be related to network motifs, as proposed in (Charlebois et al., 2014; Camellato et al., 2019). Intuitively, this behavior arises from the mutual inhibitory connections between survival and death pathways. Indeed, the first TNF treatment selects the cells that have activated genes of the survival pathway, and therefore cannot activate their death pathway upon a second TNF treatment. In the following section, we further investigate which parts of the network contribute to drug resistance by comparing wild-type and various mutant simulations.

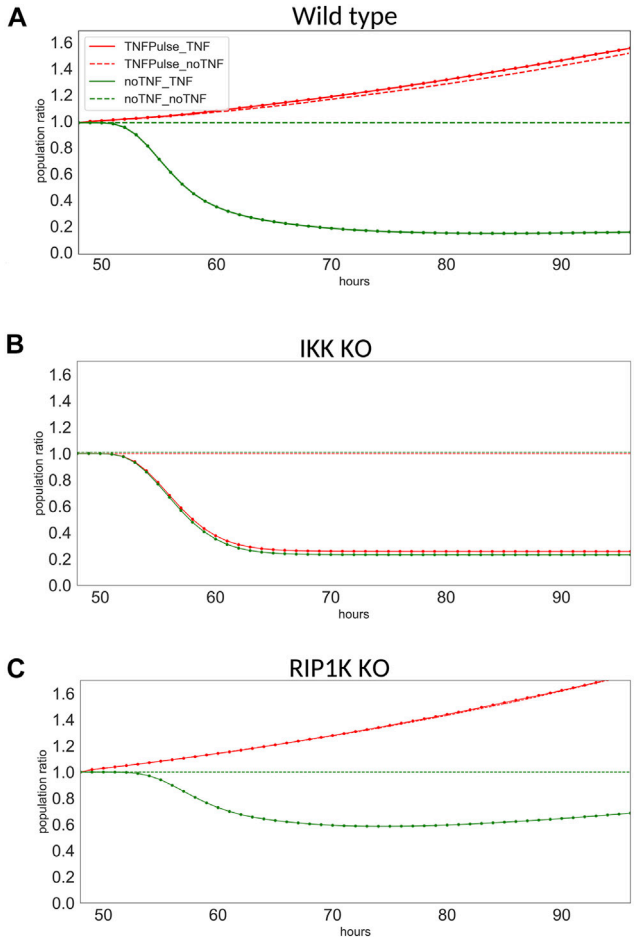


FIGURE 6 | Population ratio at from t = 48–96 h for three models. Population ratios for the four conditions in **(A)** wild type, **(B)** IKK knock-down, and **(C)** RIP1K knock-down.

3.2.4 TNF Treatments in Mutant Conditions

In the clinics, the various mutations found in patients may affect the efficacy of the response to treatments. To explore the potential roles of the different model components in the observed TNF resistance, we simulated the effects of all possible single mutants (Jupyter notebook *CellFateModel_upmaboss* at https://github.com/sysbio-curie/UPMaBoSS-docker/tree/master/CellFate).

For each single mutant, we measured the population ratio at t = 96 h for four possible scenarios: "No TNF" + "No TNF" (control, dashed green curve), "No TNF" + "TNF" (late treatment, plain green curve), "Pulse of TNF" + "No TNF" (single early treatment, dashed red curve), and "Pulse of TNF" + "TNF" (consecutive treatments, plain red curve).

Here, we focus on the results obtained for three genetic backgrounds: wild type, IKK knock-down, and RIP1K knock-down (**Figure 6**). For sake of simplicity, we focus on the response after t = 48 h, and to ease the comparison, we normalized the population ratio found in **Figure 5**. Note that two conditions were added to those shown on **Figure 5**, where only two of the four scenarios were simulated (i.e., we added the cases when there is no treatment after 48 h, no matter what the cells receive at t = 0, dashed lines).

The wild type model clearly exhibits a resistance: when the cells have received a first treatment, they do not respond, whether they receive a second treatment or not (**Figure 6A**, for TNFPulse_TNF and TNFPulse_noTNF). The resistance effect is lost for IKK knock-down (**Figure 6B**), as the mutated cells respond to the treatment whether they received a first treatment or not (plain and dashed lines coincide). In the case of the RIP1K knock-down, the decrease of the population size after TNF treatments is milder than for wild type model, but the resistance mechanism is still observed (**Figure 6C**).

IKK belongs to the NFκB pathway that induces survival. This pathway introduces a positive feedback at the cell population level, which may explain why IKK knock-down shuts down the resistance to TNF by blocking this feedback loop. RIP1K induces non-apoptotic cell death by blocking ATP, which explains the reduction of the apoptotic effect of TNF observed for the RIP1K knock-down.

The effects of double mutations can also be simulated. However, for our example, the double mutant simulations do not result into additional insight because the single mutations are already informative.

All results and figures of this analysis can be reproduced with the Jupyter notebook provided online in the folder *usecases* in the CoLoMoTo Docker image (https://colomoto.github.io/colomoto-docker/), as well as on github: https://github.com/sysbio-curie/UPMaBoSS-docker/tree/master/CellFate).

4 DISCUSSION

UPMaBoSS is a novel modeling framework enabling the exploration of cell population dynamics. It considers the division, the death and interactions of cell populations and relies on stochastic simulations with regular synchronous updating of cell populations.

Using a simple toy model, we first showed that the results of the simulations are different if we consider a homogeneous, non-interacting cell population compared to a dynamic population of interacting cells. We further applied our approach to a model of TNF-induced cell fate. We show that the paracrine loop involving NFκB enhances cell death. Surprisingly, our simulation revealed an intriguing resistance mechanism: once the cells have been treated by TNF transiently, they can resist to a second treatment, an effect not attributable to genetic selection. One limitation of our approach lies in the setting of population update time. However, the case study presented here suggests that results are moderately sensitive to changes of this parameter within a reasonable interval. Finally, an application of this modeling framework to a concrete example describing events of the immunogenic cell death was recently published (Checcoli et al., 2020). In this study, the main steps of the immunogenic cycle and the relative timing of the events were reproduced: the parameters were chosen so that some processes such as the migration of the dendritic cells to the tumor microenvironment would take the expected time (around 12 h).

Simulations and predictions obtained with UPMaBoSS could be validated experimentally using different techniques: the probabilities of the nodes corresponding to proteins/mRNA can be compared to

their experimental relative concentrations (measured by western blot, immunofluorescence, qPCR, etc.). This can be done by considering positive and negative controls, since what we are most interested in with this approach is to compute the change of probabilities between two experiments or 2 cell conditions rather than compare exactly the model probabilities with quantitative experimental measures. This comparison can be facilitated by normalizing experimental measures between 0 and 1. The probabilities of the network states could be validated using flow cytometry or microscopic images, using thresholds that separate active/inactive states. Single-cell transcriptomic (with CRISPR/Cas9 genetic engineering) and flow cytometry (on cell population markers), in principle, can be used to estimate the sizes of different cell sub-populations: once thresholds have been applied on the quantitative markers defining cell populations, the experimental data can be translated in terms of cell sub-populations.

There are, of course, a number of limitations with our framework. First, UPMaBoSS is not meant to answer precise biological questions on diffusion, localization, or drug dosage. More quantitative modeling frameworks would be more appropriate to address such questions. However, previous studies suggested that the presence, rather the precise localization, of immune infiltrates in tumors constitute good prognostic biomarkers (Baxevanis et al., 2019), which could be modeled with UPMaBoSS.

Another limitation is that UPMaBoSS is not optimal to model metabolic networks, which involve reactions consuming reactants. In practice, metabolites could be included in UPMaBoSS models, in particular if they play a role in signaling pathways. Metabolites are then represented by Boolean nodes, implying and requiring a proper discretization of their levels (multilevel variables can then be encoded by multiple Boolean nodes).

We think that this approach is appropriate to model many processes in cell biology, including cell differentiation, innate/adaptive immune system activation, cancer micro-environment, and tissue homeostasis. The fact that the spatial dimension is not taken into account might appear as a limiting factor for clinical applications. In this respect, it is possible to extend the model using PhysiBoSS, a multi-agent modeling tool in which each agent is a cell running an intracellular MaBoSS model, but which requires to tune more parameters (Letort et al., 2018).

In the future, we plan to apply UPMaBoSS to model the effect of the micro-environment on cancer cell fate. UPMaBoSS simulations are based on consecutive runs of MaBoSS. While MaBoSS generates tens of thousands trajectories, UPMaBoSS parses and writes probability distributions for each population updates. Consequently, computational cost increases moderately (less than an order of magnitude) compared to prior MaBoSS models. Moreover, because UPMaBoSS can be run directly on a conda environment or within a simple pair of python and C++ scripts, simulations can be launched on High Performance Computers, with a natural parallelization for tasks such as parameter variations and gene mutations.

Finally, on the practical side, the implementation of UPMaBoSS in a docker image and in a conda package should greatly facilitate its use, whereas the reproducibility of the analyses can be enforced by the use of Jupyter notebooks.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, and on a GitHub repository (https://github.com/sysbiocurie/UPMaBoSS-docker). Further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

All authors have read and approved the manuscript. GS and EV conceived the algorithm and the tool. VN, EV, and AN participated in the development of the tool. GS, AN, DT, and LC worked on the examples. DT, EB, GK, GS, and LC designed and supervised the study. GS and LC wrote the article but all co-authors contributed to the writing and agree on the content.

FUNDING

VN and LC were partially funded by ANR-FNR project "AlgoReCell" (ANR-16-CE12-0034) and VN was partially funded by the European Union's Horizon 2020 Programme under the PerMedCoE Project (www.permedcoe.eu), grant agreement n° 951773. The project received financial support from Inserm Cancer (ModICeD, ITMO MIC 2020).

ACKNOWLEDGMENTS

We would like to thank I. Martins. GK is supported by the Ligue contre le Cancer (équipe labellisée); Agence National de la Recherche (ANR)—Projets blancs; ANR under the frame of E-Rare-2, the ERA-Net for Research on Rare Diseases; AMMICa US23/CNRS UMS3655; Association pour la recherche sur le cancer (ARC); Association "Le Cancer du Sein, Parlons-en!"; Cancéropôle Ile-de-France; Chancelerie des universités de Paris (Legs Poix), Fondation pour la Recherche Médicale (FRM); a donation by Elior; European Research Area Network on Cardiovascular Diseases (ERA-CVD, MINOTAUR); Gustave Roussy Odyssea, the European Union Horizon 2020 Project Oncobiome; Fondation Carrefour; Institut National du Cancer (INCa); Inserm (HTE); Institut Universitaire de France; LeDucq Foundation; the LabEx Immuno-Oncology (ANR-18-IDEX-0001); the RHU Torino Lumière; the Seerave Foundation; the SIRIC Stratified Oncology Cell DNA Repair and Tumor Immune Elimination (SOCRATE); and the SIRIC Cancer Research and Personalized Medicine (CARPEM).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.800152/full#supplementary-material

REFERENCES

Abou-Jaoudé, W., Traynard, P., Monteiro, P. T., Saez-Rodriguez, J., Helikar, T., Thieffry, D., et al. (2016). Logical Modeling and Dynamical Analysis of Cellular Networks. *Front. Genet.* 7, 94. doi:10.3389/fgene.2016.00094

Albert, I., Thakar, J., Li, S., Zhang, R., and Albert, R. (2008). Boolean Network Simulations for Life Scientists. *Source Code Biol. Med.* 3, 16. doi:10.1186/1751-0473-3-16

Altinok, A., Gonze, D., Lévi, F., and Goldbeter, A. (2011). An Automaton Model for the Cell Cycle. *Interf. Focus.* 1, 36–47. doi:10.1098/rsfs.2010.0009

Anderson, A. R. A., Weaver, A. M., Cummings, P. T., and Quaranta, V. (2006). Tumor Morphology and Phenotypic Evolution Driven by Selective Pressure from the Microenvironment. *Cell* 127, 905–915. doi:10.1016/j.cell.2006.09.042

Baxevanis, C. N., Sofopoulos, M., Fortis, S. P., and Perez, S. A. (2019). The Role of Immune Infiltrates as Prognostic Biomarkers in Patients with Breast Cancer. *Cancer Immunol. Immunother.* 68, 1671–1680. doi:10.1007/s00262-019-02327-7

Bertazza, L., and Mocellin, S. (2010). The Dual Role of Tumor Necrosis Factor (TNF) in Cancer Biology. *Curr. Med. Chem.* 17, 3337–3352. doi:10.2174/092986710793176339

Bonabeau, E. (2002). Agent-Based Modeling: Methods and Techniques for Simulating Human Systems. *Proc. Natl. Acad. Sci.* 99, 7280–7287. doi:10.1073/pnas.082080899

Calzone, L., Tournier, L., Fourquet, S., Thieffry, D., Zhivotovsky, B., Barillot, E., et al. (2010). Mathematical Modelling of Cell-Fate Decision in Response to Death Receptor Engagement. *PLoS Comput. Biol.* 6, e1000702. doi:10.1371/journal.pcbi.1000702

Camellato, B., Roney, I. J., Azizi, A., Charlebois, D., and Kaern, M. (2019). Engineered Gene Networks Enable Non-Genetic Drug Resistance and Enhanced Cellular Robustness. *Eng. Biol.* 3, 72–79. doi:10.1049/enb.2019.0009

Cartwright, J. H. E., Piro, O., and Tuval, I. (2009). Fluid Dynamics in Developmental Biology: Moving Fluids that Shape Ontogeny. *HFSP J.* 3, 77–93. doi:10.2976/1.3043738

Chaouiya, C., Berenguier, D., Keating, S. M., Naldi, A., van Iersel, M. P., Rodriguez, N., et al. (2013). SBML Qualitative Models: A Model Representation Format and Infrastructure to foster Interactions between Qualitative Modelling Formalisms and Tools. *BMC Syst. Biol.* 7, 135. doi:10.1186/1752-0509-7-135

Charlebois, D. A., Balázsi, G., and Kærn, M. (2014). Coherent Feedforward Transcriptional Regulatory Motifs Enhance Drug Resistance. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 89, 052708. doi:10.1103/PhysRevE.89.052708

Charlebois, D. A., Intosalmi, J., Fraser, D., and Kærn, M. (2011). An Algorithm for the Stochastic Simulation of Gene Expression and Heterogeneous Population Dynamics. *Commun. Comput. Phys.* 9, 89–112. doi:10.4208/cicp.280110.070510a

Charlebois, D. A., and Kærn, M. (2013). An Accelerated Method for Simulating Population Dynamics. *Commun. Comput. Phys.* 14, 461–476. doi:10.4208/cicp.130612.121012a

Checcoli, A., Pol, J. G., Naldi, A., Noël, V., Barillot, E., Kroemer, G., et al. (2020). Dynamical Boolean Modeling of Immunogenic Cell Death. *Front. Physiol.* 11, 1320. doi:10.3389/fphys.2020.590479

Clairambault, J. (2006). Physiologically Based Modelling of Circadian Control on Cell Proliferation. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2006, 173–176. doi:10.1109/IEMBS.2006.260855

Cohen, D., Kuperstein, I., Barillot, E., Zinovyev, A., and Calzone, L. (2013). From a Biological Hypothesis to the Construction of a Mathematical Model. *Methods Mol. Biol.* 1021, 107–125. doi:10.1007/978-1-62703-450-0_6

Collombet, S., van Oevelen, C., Ortega, J. L. S., Abou-Jaoude, W., Di Stefano, B., Thomas-Chollier, M., et al. (2017). Logical Modeling of Lymphoid and Myeloid Cell Specification and Transdifferentiation. *Proc. Natl. Acad. Sci. U S A.* 114, 5792–5799. doi:10.1073/pnas.1610622114

Cowan, A. E., Moraru, I. I., Schaff, J. C., Slepchenko, B. M., and Loew, L. M. (2012). Spatial Modeling of Cell Signaling Networks. *Methods Cel. Biol.* 110, 195–221. doi:10.1016/B978-0-12-388403-9.00008-4

Dong, J., Jimi, E., Zeiss, C., Hayden, M. S., and Ghosh, S. (2010). Constitutively Active NF-κB Triggers Systemic TNFα-Dependent Inflammation and Localized TNFα-Independent Inflammatory Disease. *Genes Dev.* 24, 1709–1717. doi:10.1101/gad.1958410

Drouet, C., Shakhov, A. N., and Jongeneel, C. V. (1991). Enhancers and Transcription Factors Controlling the Inducibility of the Tumor Necrosis Factor-Alpha Promoter in Primary Macrophages. *J. Immunol.* 147, 1694–1700.

Eduati, F., Dolde-Martelli, V., Klinger, B., Cokelaer, T., Sieber, A., Kogera, F., et al. (2017). Drug Resistance Mechanisms in Colorectal Cancer Dissected with Cell Type-specific Dynamic Logic Models. *Cancer Res.* 77, 3364–3375. doi:10.1158/0008-5472.CAN-17-0078

El-Badri, N. S., Hakki, A., Ferrari, A., Shamekh, R., and Good, R. A. (2007). Autoimmune Disease: Is it a Disorder of the Microenvironment? *Immunol. Res.* 41, 79–86. doi:10.1007/s12026-007-0053-8

Ferrell, J. E. (2015). Bistability, Bifurcations, and Waddington's Epigenetic Landscape. *Curr. Biol.* 22, 11. doi:10.1016/j.cub.2012.03.045

Ghaffarizadeh, A., Heiland, R., Friedman, S. H., Mumenthaler, S. M., and Macklin, P. (2018). Physicell: An Open Source Physics-Based Cell Simulator for 3-d Multicellular Systems. *Plos Comput. Biol.* 14, e1005991. doi:10.1371/journal.pcbi.1005991

Helikar, T., Kowal, B., McClenathan, S., Bruckner, M., Rowley, T., Madrahimov, A., et al. (2012). The Cell Collective: Toward an Open and Collaborative Approach to Systems Biology. *BMC Syst. Biol.* 6, 96. doi:10.1186/1752-0509-6-96

Hoehme, S., and Drasdo, D. (2010). A Cell-Based Simulation Software for Multi-Cellular Systems. *Bioinformatics* 26, 2641–2642. doi:10.1093/bioinformatics/btq437

Jin, G. (2013). *Feed Forward Loop*. New York, NY: Springer New York, 737–738. doi:10.1007/978-1-4419-9863-7_463

Kolch, W., Halasz, M., Granovskaya, M., and Kholodenko, B. N. (2015). The Dynamic Control of Signal Transduction Networks in Cancer Cells. *Nat. Rev. Cancer* 15, 515–527. doi:10.1038/nrc3983

Le Novère, N. (2015). Quantitative and Logic Modelling of Molecular and Gene Networks. *Nat. Rev. Genet.* 16, 146–158. doi:10.1038/nrg3885

Letort, G., Montagud, A., Stoll, G., Heiland, R., Barillot, E., Macklin, P., et al. (2018). Physiboss: A Multi-Scale Agent-Based Modelling Framework Integrating Physical Dimension and Cell Signalling. *Bioinformatics* 35, 1188–1196. doi:10.1093/bioinformatics/bty766

Liu, H., Sidiropoulos, P., Song, G., Pagliari, L. J., Birrer, M. J., Stein, B., et al. (2000). TNF-alpha Gene Expression in Macrophages: Regulation by NF-Kappa B Is Independent of C-Jun or C/EBP Beta. *J. Immunol.* 164, 4277–4285. doi:10.4049/jimmunol.164.8.4277

Müssel, C., Hopfensitz, M., and Kestler, H. A. (2010). Boolnet—an R Package for Generation, Reconstruction and Analysis of Boolean Networks. *Bioinformatics* 26, 1378–1380. doi:10.1093/bioinformatics/btq124

Naldi, A., Hernandez, C., Abou-Jaoude, W., Monteiro, P. T., Chaouiya, C., and Thieffry, D. (2018). Logical Modeling and Analysis of Cellular Regulatory Networks with Ginsim 3.0. *Front. Physiol.* 9, 646. doi:10.3389/fphys.2018.00646

Remy, E., Rebouissou, S., Chaouiya, C., Zinovyev, A., Radvanyi, F., and Calzone, L. (2015). A Modelling Approach to Explain Mutually Exclusive and Co-occurring Genetic Alterations in Bladder Tumorigenesis. *Cancer Res.* 75 (19), 4042–4052. doi:10.1158/0008-5472.CAN-15-0602

Shakhov, A. N., Collart, M. A., Vassalli, P., Nedospasov, S. A., and Jongeneel, C. V. (1990). Kappa B-Type Enhancers Are Involved in Lipopolysaccharide-Mediated Transcriptional Activation of the Tumor Necrosis Factor Alpha Gene in Primary Macrophages. *J. Exp. Med.* 171, 35–47. doi:10.1084/jem.171.1.35

Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W. (2002). Probabilistic Boolean Networks: A Rule-Based Uncertainty Model for Gene Regulatory Networks. *Bioinformatics* 18, 261–274. doi:10.1093/bioinformatics/18.2.261

Smyth, M. J., Cretney, E., Kershaw, M. H., and Hayakawa, Y. (2004). Cytokines in Cancer Immunity and Immunotherapy. *Immunol. Rev.* 202, 275–293. doi:10.1111/j.0105-2896.2004.00199.x

Stoll, G., Caron, B., Viara, E., Dugourd, A., Zinovyev, A., Naldi, A., et al. (2017). Maboss 2.0: An Environment for Stochastic Boolean Modeling. *Bioinformatics* 33, 2226–2228. doi:10.1093/bioinformatics/btx123

Stoll, G., Viara, E., Barillot, E., and Calzone, L. (2012). Continuous Time Boolean Modeling for Biological Signaling: Application of gillespie Algorithm. *BMC Syst. Biol.* 6, 116. doi:10.1186/1752-0509-6-116

Swat, M. H., Thomas, G. L., Belmonte, J. M., Shirinifard, A., Hmeljak, D., and Glazier, J. A. (2012). Multi-scale Modeling of Tissues Using Compucell3d. *Methods Cel. Biol.* 110, 325–366. doi:10.1016/b978-0-12-388403-9.00013-8

Udommethaporn, S., Tencomnao, T., McGowan, E. M., and Boonyaratanakornkit, V. (2016). Assessment of Anti-TNF-α Activities in Keratinocytes Expressing Inducible TNF- α: A Novel Tool for Anti-TNF-α Drug Screening. *PloS one* 11 (7), e0159151. doi:10.1371/journal.pone.0159151

Varela, P. L., Ramos, C. V., Monteiro, P. T., and Chaouiya, C. (2019). EpiLog: A Software for the Logical Modelling of Epithelial Dynamics. *F1000Research* 7, 1145. doi:10.12688/f1000research.15613.2

Viceconti, M., and Hunter, P. (2016). The Virtual Physiological Human: Ten Years after. *Annu. Rev. Biomed. Eng.* 18, 103–123. doi:10.1146/annurev-bioeng-110915-114742

Wang, Y., Xu, J., Zhang, X., Wang, C., Huang, Y., Dai, K., et al. (2017). Tnf-alpha-induced Lrg1 Promotes Angiogenesis and Mesenchymal Stem Cell Migration in the Subchondral Bone during Osteoarthritis. *Cell Death Dis* 8 (3), e2715. doi:10.1038/cddis.2017.129

Wolkenhauer, O., Auffray, C., Brass, O., Clairambault, J., Deutsch, A., Drasdo, D., et al. (2014). Enabling Multiscale Modeling in Systems Medicine. *Genome Med.* 6, 21. doi:10.1186/gm538

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Stoll, Naldi, Noël, Viara, Barillot, Kroemer, Thieffry and Calzone. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 08 March 2022
doi: 10.3389/fmolb.2022.806213

Applications of Coarse-Grained Models in Metabolic Engineering

Dieu Thi Doan[1]*, *Manh Dat Hoang*[2], *Anna-Lena Heins*[2] and *Andreas Kremling*[1]

[1]Systems Biotechnology, TUM School of Engineering and Design, Technische Universität München, Garching, Germany, [2]Biochemical Engineering, TUM School of Engineering and Design, Technische Universität München, Garching, Germany

Mathematical modeling is a promising tool for better understanding of cellular processes. In recent years, the development of coarse-grained models has gained attraction since these simple models are able to capture and describe a broad range of growth conditions. Coarse-grained models often comprise only two cellular components, a low molecular component as representative for central metabolism and energy generation and a macromolecular component, representing the entire proteome. A framework is presented that presents a strict mass conservative model for bacterial growth during a biotechnological production process. After providing interesting properties for the steady-state solution, applications are presented 1) for a production process of an amino acid and 2) production of a metabolite from central metabolism.

Keywords: mathematical modeling, resource allocation, kinetic model, metabolic engineering, systems biotechnology, metabolic and gene regulatory networks, microbial growth

OPEN ACCESS

Edited by:
Ernesto Perez-Rueda, Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Luis Diambra, National University of La Plata, Argentina
Nuno F Azevedo, University of Porto, Portugal

***Correspondence:**
Dieu Thi Doan
thi.doan@tum.de

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *31 October 2021*
Accepted: *19 January 2022*
Published: *08 March 2022*

Citation:
Doan DT, Hoang MD, Heins A-L and Kremling A (2022) Applications of Coarse-Grained Models in Metabolic Engineering. Front. Mol. Biosci. 9:806213. doi: 10.3389/fmolb.2022.806213

1 INTRODUCTION

To gain a full understanding of cellular processes, the usage of mathematical modeling and the analysis of those have become a standard in metabolic engineering, systems biology, and process engineering. Predictive models which can describe relevant cellular processes can be used as a basis for process observation and process design with the intention to optimize the properties and behavior of the cells. Even though bacteria are very diverse, the basic principles of their metabolism are quite similar. In general, every bacterial population has to cope with its environment, scavenge for nutrients, and then coordinate its central metabolism accordingly for growth and survival. The underlying regulatory networks are very densely intertwined, large, complex, and not fully known or incompletely understood, thus providing a big challenge to understand the processes in its entirety. Mathematical models trying to describe all these processes in detail are challenging and nearly impossible because of the huge number of variables and uncertain parameters.

Coarse-grained models have been used in the recent years and are now frequently used to get a better understanding on cellular control strategies, gene expression, and resource allocation Bollenbach et al. (2009), Scott et al. (2010). In this type of model, levels of cellular organization with similar functions are "lumped" together into a small number of modules Maitra and Dill (2015), Giordano et al. (2016), Pandey and Jain (2016), Sharma et al. (2018), Molenaar et al. (2009). In contrast to whole-cell models with hundreds of individual reactions and components, the number of state variables in coarse-grained models is very low, and kinetic parameters are obtained by either a rough estimation from literature data or by regression from experimental data. An important hallmark of coarse-grained models is allocation of cellular resources. This is expressed, for example, by linking biochemical reactions to the available fraction of the proteome for the respective module. In this way, a reaction can only take place if enough resources are available. The goal can be achieved by efficient proteome allocation in a way where no resources are wasted. Using this fundamental

assumption, many coarse-grained models have been proposed to analyze certain metabolic effects such as metabolic overflow Basan et al. (2015), production of heterologous protein Scott et al. (2010), or applications in synthetic biology Weiße et al. (2015).

Typically, coarse-grained models are written down as a set of differential equations for the components of the model, whose unit, for example, is the number of molecules per cell or mol/g dry cell weight. However, mass balance equations must fulfill the conservation of mass as dictated by the first fundamental theorem of thermodynamics, and often, a consistent transfer from mass balance equations to differential equations for the concentrations of the model components is faulty or inadequate. Therefore, we start by a brief recapitulation of the structure of the ordinary differential equations for coarse-grained models that are combined with models for the environment, for example, in a bioreactor process system. Here, a new relationship for the specific growth rate in dependence on the exchange reactions of the entire network is given. This equation is fundamental since it guarantees strict mass conservation for the complete system. Conventionally, the growth rate is an empirical function and, therefore, strict mass conservation is not ensured. In a second step, we analyze the model and show interesting properties of the steady-state behavior. We provide a general steady-state solution for biochemical networks and compare outcomes of a traditional flux balance analysis with our new approach. Finally, various applications and extensions for a broad spectrum of problem formulations in biotechnology are provided: 1) an L-phenylalanine production process and 2) the production of a metabolite from central metabolism. Hereby, problems of resource allocation as well as problems of parameter estimation are addressed.

2 MASS CONVERSATION IN MODELS FOR MICROBIAL SYSTEMS

From thermodynamic principles, mass conversion is crucial and plays the major role of determining the time course of selected quantities of interest (system volume, concentration of reaction partners, and temperature) which are called state variables. From a static view on the biochemical reaction equations alone, however, it is not possible to infer on the time course of the state variables. A mass balance equation that describes the change of a compound over time and sums up the material flow in and out of the system comes into play here. It is a differential equation. Since we are interested in the mass m_i of a component i, the mass balance reads as

$$\frac{d\,m_i}{dt} = J + P. \tag{1}$$

In this general equation, J describes the mass flow into the system while P describes conversion inside the system, for example, by biochemical reactions. Both summands depend on other state variables in the system. In the current form, the equation cannot be applied. The reason is as follows: for biochemical reaction networks, P describes mass conversion by reactions, and the reaction velocity strongly depends on the concentration of a compound given by $c_i = m_i/V$ of the reaction partners and not on the mass m_i of the reaction partners alone. For applications in systems biology, synthetic biology, and biotechnology, a different convention is used for the definition of the concentration of the cellular components (but not for environmental compounds). Since it is much easier to determine the entire biomass m_X than the cellular volume, the following definition is used, instead, for the concentration of an intracellular metabolite: $c_i = m_i/m_X$.

To avoid inconsistencies, it is recommended to always start from the mass balance and reformulate the mass balance into an equation for the concentration (the resulting equation is not a mass balance in the strict sense, but in literature, we often find this term). For a cellular network, the basic differential equation then reads[1]

$$\dot{\underline{c}} = N\ \underline{r} - \mu\ \underline{c}, \tag{2}$$

for the vector $\underline{c}$, for the concentration of all components, a reaction system given by the stoichiometric matrix N and a rate vector $\underline{r}(\underline{c})$ that is dependent on $\underline{c}$ as well. The specific growth rate μ is an integral parameter that—in a strict sense—is determined by the mass exchange of the population with its environment. Therefore, it is defined as

$$\mu = \dot{m}_X \Big/ m_X. \tag{3}$$

The specific growth rate is related to the doubling time τ of the population in the relationship $\tau = \ln 2/\mu$. For the biomass itself, also, a mass balance is set up that takes into account the changes of biomass concentration due to removal from the bioreactor and biomass formation due to growth. However, a different approach can be used as shown in the *Supporting Information* that describes the changes of biomass based on the changes of all compounds representing the biomass. Therefore, from a formal point of view, the specific growth rate μ only depends on the rate vector $\underline{r}$ that describes internal processes as well as mass exchange with the surrounding and is given with the vector of all molecular weights of the components $\underline{w}$

$$\mu = \underline{w}^T\ N\ \underline{r}. \tag{4}$$

Plugging **Eq. 4** into **Eq. 2**, the latter can be rewritten as

$$\dot{\underline{c}} = N\ \underline{r} - \underline{w}^T\ N\ \underline{r}\ \underline{c} = \left(\mathrm{Id} - \underline{c}\ \underline{w}^T\right) N\ \underline{r} = W\ N\ \underline{r}, \tag{5}$$

with W representing the mass matrix. Together with the equations for the concentration for a substrate S and biomass X in a bioreactor system with feeding rate q_{in}, feed concentration S_{in}, and stoichiometric vector $\underline{n}_S$

$$\dot{S} = \frac{q_{in}}{V}\ (S^{in} - S) - \underline{n}_S^T\ \underline{r}\, X, \tag{6}$$

[1]Please note that underlined symbols are used for vectors.

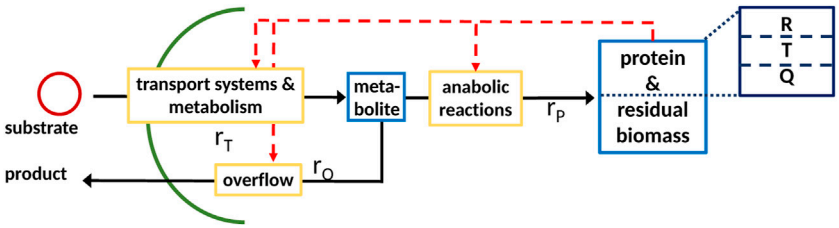


FIGURE 1 | General scheme of a coarse-grained model with partitioned proteome (ribosomal proteins *R*, proteins linked with the central metabolism *T*, and residual protein fraction *Q*) as the self-replicator system Scott et al. (2014); it consists of two components, indicated as blue boxes; (metabolite, low molecular weight), protein; and residual biomass (high molecular weight; protein is assumed to be 50% of total biomass). The pools are connected by a minimal set of reactions, indicated by yellow boxes, for substrate uptake, overflow metabolism, and protein synthesis.

$$\dot{X} = \left(\mu - \frac{q_{in}}{V}\right) X, \tag{7}$$

the system is completely described. The first term of **Eq. 6** accounts for the feeding substrate, the second the dilution due to the feed, and the last term the substrate uptake of the biomass. The second term of **Eq. 7** represents the growth and a dilution due to the feed.

3 STEADY-STATE ANALYSIS

3.1 Flux Analysis

In classic flux balance analysis, the equation for a cellular network only consists of the stoichiometric matrix *N*. Solutions for rate vector $\underline{r}$ are investigated by determining the kernel of *N*, providing possible fluxes through the cellular network Orth et al. (2010). With the proposed approach, in addition, properties of the mass matrix *W* has to be taken into account as well, and the steady-state solution for the intracellular network is obtained from the relationship

$$0 = W\,N\,\underline{r}, \tag{8}$$

which is obtained from **Eq. 5** by setting the left side to zero. The solution of this equation is determined by not only the kernel of *N* as in the classic flux balance analysis but also by the kernel of matrix *W*, which is determined by the molecular composition of the cell. The determinant of *W* is given by

$$\det(W) = 1 - \sum_i w_i\, c_i. \tag{9}$$

The addends of the second term are the mass fractions of cellular component *i*, and given the strict mass conservation, the sum over all mass fractions equals one. Therefore, the determinant of *W* is always zero, and *W* is nonsingular. Additionally, one can show that the kernel of *W* is in fact one-dimensional (the proof can be found in *Supplementary Information*). The complete solution $\underline{r}_0$ of the relationship given in **Eq. 8** comprises two terms: the kernel of *N*, denoted by $\underline{r}_{n,0}$, and second, the product of the Moore–Penrose inverse of *N*, denoted by N^+, and the kernel of *W*, denoted by $\underline{c}_{w,0}$

$$\underline{r}_0 = \underline{r}_{n,0} + N^+ \underline{c}_{w,0}. \tag{10}$$

The two summands in the solution are not given in a unique way and can be written with scalar factor *s* and an arbitrary vector $\underline{a}$ with the same dimension as the rate vector $\underline{r}_0$ as follows:

$$\underline{c}_{w,0} = \underline{c}\; s, \qquad s \in \mathbb{R} \tag{11}$$

$$\underline{r}_{n,0} = (\mathrm{Id} - N^+ N)\; \underline{a}\,. \tag{12}$$

This principle holds true for all types of cellular networks independently of its size and form, which can range from whole-cell models to, in this case, coarse-grained models as shown in **Figure 1**.

A typical minimal reaction system as shown in **Figure 1** is considered with a cellular network that represents the entire biomass (the sum of all components in the network weighted with their molecular weight) and only one anabolic reaction. The scheme is given as follows[2]:

$$\begin{aligned} (S) &\rightarrow \alpha\, M \\ \gamma\, M &\rightarrow \beta\, P \\ M &\rightarrow (\text{by} - \text{product}). \end{aligned} \tag{13}$$

It is to be noted that an extension to two or more anabolic reactions can be performed easily since, in general, the mass fractions of the macromolecules are well-known. The reaction systems, therefore, consist of a pool of metabolites *M*, proteins *P*, and reactions r_i that connect the pools with each other and the environment. Rate r_T describes the transport of the substrate into the cell, while rate r_O describes overflow metabolism. Proteins *P* are synthesized with rate r_P Maitra and Dill (2015), Giordano et al. (2016), Pandey and Jain (2016)) The vector of components reads $\underline{c} = (M, P)$, and the stoichiometric matrix for this system is as follows:

$$N = \begin{pmatrix} \alpha & -\gamma & -1 \\ 0 & \beta & 0 \end{pmatrix}. \tag{14}$$

Thus, the intracellular network for this minimal model can be written as

$$\begin{pmatrix} \dot{M} \\ P \end{pmatrix} = N \begin{pmatrix} r_T \\ r_P \\ r_O. \end{pmatrix} - \mu \begin{pmatrix} M \\ P \end{pmatrix}. \tag{15}$$

[2]Stoichiometric coefficients are given with Greek symbols.

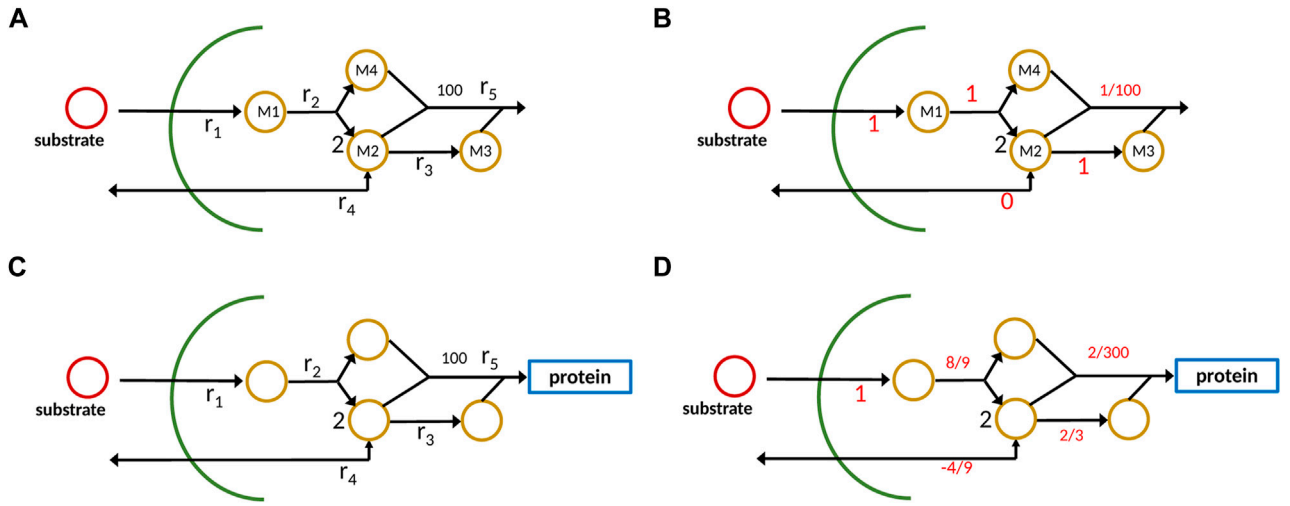


FIGURE 2 | Example network with metabolites M_i and reactions r_i; conventional representation **(A)**, new approach with one component representing protein/ biomass sector **(C)**. On the right side **(B,D)**, output data from the calculations are shown; it is to be noted that for the new approach, the dilution term must be taken into account which is not shown in the plot.

For the basic structure, with **Eq. 4**, the specific growth rate is given by

$$\mu = (\alpha\, r_T - r_O)\, w_M + r_P\, (\beta\, w_P - \gamma\, w_M), \tag{16}$$

where w_M and w_P are the molecular weights of metabolites M and proteins P, respectively. With the observation from **Eq. 10**, the solution of the rate vector reads

$$\underline{r} = \begin{pmatrix} \frac{1}{\alpha^2+1} \\ 0 \\ \frac{\alpha}{\alpha^2+1} \end{pmatrix} a + N^+ \underline{c}\; s \tag{17}$$

with the first term representing the solution from the stoichiometry of the system and the second term the solution determined by the molecular composition of the cell. With a closer look at solution **Eq. 17**, the rate connecting only intracellular components of the system, which is the protein synthesis rate r_P, is only defined by the second term. Thus, for the assumption of the known specific growth rate μ and molecular composition of the cell, this rate is fixed, while the remaining rates, meaning the substrate uptake rate r_T and the overflow metabolism rate r_O, are coupled through one degree of freedom a. If one of these rates is known, the degree of freedom a can be determined and, therefore, the last remaining rate.

3.2 Flux Analysis in Comparison to a Constraint-Based Method

To illustrate the different outcomes when applying the new approach with strict mass conservation as seen in **Eq. 5** in comparison to a standard analysis with a constraint-based method, a small network with four metabolites and five reactions is considered (**Figure 2A**).

Substrate S is taken up and four metabolites are generated which in the conventional approach are consumed in reaction r_5 to produce biomass. To make it realistic, we assume that the stoichiometric coefficient for biomass is 100, that is, 100 small molecules are used to get 1 mol protein/ biomass. The reaction system is given as follows:

$$\begin{aligned} (S) &\rightarrow M_1 \\ M_1 &\rightarrow 2\, M_2 + M_4 \\ M_2 &\rightarrow M_3 \\ M_3 &\rightarrow \\ 100\, M_2 + 100\, M_3 + 100\, M_4 &\rightarrow (\text{protein/biomass}). \end{aligned} \tag{18}$$

In the case of constraint-based models that are used, for example, for flux balance analysis or by Bollenbach et al. (2009), the stoichiometric matrix noted by N_1 for the internal network has four rows and five columns:

$$N_1 = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 2 & -1 & 0 & -100 \\ 0 & 0 & 1 & -1 & -100 \\ 0 & 1 & 0 & 0 & -100 \end{pmatrix}. \tag{19}$$

Hence, the null space of N_1 is one-dimensional, and the only possible solution, when providing 1 unit flux of substrate uptake, results in 0.01 unit of biomass (it is to be noted that in reaction 2, 2 mol of M_2 are produced). The situation becomes different when strict mass conservation is taken into account. Here, we consider the formation of protein/biomass. Therefore, the overall biomass composition dictates the flux distribution in this case. From the depicted scheme, we infer the molecular weight for the components as follows (it is to be noted that these numbers are not unique, but used here for demonstration purposes)

$\underline{w} = (w_1, w_1/4, w_1/4, w_1/2, 100\,w_1)$. Given the stoichiometric matrix N_2 (extension of N_1 by one component for the protein, that is, one additional row) and the vector of molecular weights, the specific growth μ can be calculated as mentioned before with **Eq. 4** with a possible flux vector $\underline{r}$ to be

$$\mu = \begin{pmatrix} w_1 & 0 & 0 & -w_1/4 & 0 \end{pmatrix} \underline{r}. \tag{20}$$

It is to be noted that here only reactions that exchange with the environment (here r_1 and r_4) appear. In the case at hand, the stoichiometric matrix N_2, taking into account the protein/biomass fraction as additional component, has five rows and five columns and has full rank. However, the null space of $W \cdot N_2$ with W given in **Eq. 5** is one-dimensional and represents the only possible flux distribution. Matrix W strongly depends on the cellular composition; the composition itself is the steady-state solution, if the system is given in the standard form (**Eq. 2**), and all reaction kinetics are known and well-parameterized. However, for the example, we choose a different way and start with a possible composition for biomass and back-calculate the fluxes for this case. Taking the following composition vector (mass fraction) $\underline{f} = (0.1, 0.1, 0.1, 0.1, 0.6)$ as an example, the resulting flux vector (scaled to 1 for the uptake rate) is $\underline{r} = (1, 8/9, 2/3, -4/9, 2/300)$. In this case, to fulfill all steady-state equations, an additional input flux is necessary (r_4 is negative, that is, a second substrate is needed). This flux vector is very different from the solution not considering strict mass conservation.

3.3 Differential Algebra System

In the case of regulated systems, that are also named self-replicator systems, a superimposed control structure (shown in red in **Figure 1**) determines the allocation of protein resources in the individual reactions. The division of the entire proteome in fractions results in additional algebraic equations representing conservation conditions. This is shown exemplarily in **Figure 1** with three fractions; fraction R represents ribosomes, fraction T represents transport and catabolism, and fraction Q represents the remaining proteins. The dependency of the rate for protein synthesis r_p on the ribosomal fraction R of the proteome is common to many approaches Scott et al. (2014). To derive a consistent system that can be used for numerical simulation, the resource allocation problem must be formulated in mathematical terms; here, we will describe two different approaches that result in a differential algebra system or in an optimization program.

First, the rate vector $\underline{r}$ of the minimal coarse-grained model is fixed with kinetic rate laws. For the rates involved in central metabolism, a dependency to the T fraction is applied, while rate r_P will depend on the R fraction. Furthermore, the drain from central metabolism will depend on metabolite M, while the transport reaction will depend on the main substrate S. The following rates are taken as examples for the case study:

$$\begin{aligned} r_T &= k_T \frac{S}{S + K_T} T \\ r_P &= k_P \frac{M}{M + K_P} R \\ r_O &= k_O\, M\, T. \end{aligned} \tag{21}$$

The dependencies of T and R from the entire proteome P are exploited from a data set that was published by Schmidt et al. (2016). The fraction of the T and R fractions is given in dependence on the specific growth rate μ. From the data, a linear relationship can be deduced. However, a direct implementation of functions $T, R = f(\mu)$ is not possible since μ itself depends on R and T. Therefore, we proceed as follows: for the case at hand, an algebraic system could be set up for the steady-state solution of the differential equation system. The system reads

$$\begin{aligned} \underline{0} &= N \begin{pmatrix} r_T \\ r_P \\ r_O \end{pmatrix} - \mu \begin{pmatrix} M \\ P \end{pmatrix} \\ 0 &= \mu - \underline{w}^T N\, \underline{r} \\ 0 &= T - f_1(\mu) \\ 0 &= R - f_2(\mu). \end{aligned} \tag{22}$$

Thereby, the last two equations are determined by experimental data. From the solution for a broad range of the input variable, in our case, substrate concentration S, relationships of the form $T = g_1(M, P)$, and $R = g_2(M, P)$ are determined. In this way, also the dynamical system could be simulated. In addition, in this format, the system for the intracellular network consists only of two independent variables M and P. **Figure 3** shows the dependencies of sectors T and R from the specific growth rate as measured experimentally (**Figure 3A**) and the kinetics for T and R as a function of metabolite M (**Figure 3B**).

3.4 Optimization Program

The first approach with fixed reaction kinetics is compared with an optimization program to check if the given experimental data for fractions T and R are optimal for the given specific growth rates. For this, we omit the determined dependencies given in the last subsection, and the following program is formulated:

$$\begin{aligned} &\max_R \Phi, \qquad \Phi = \mu \\ &s.t. \\ &\quad \underline{0} = N\, \underline{r} - \mu\, \underline{c} \\ &\quad \mu = \underline{w}^T N\, \underline{r} \\ &\quad 0 = P_0 - T - R. \end{aligned} \tag{23}$$

The program is simplified to only one design variable R and the constraint that the sum of R and T is fixed to a constant value P_0. The kinetic rate laws are taken as given above. **Figure 4** compares the steady-state output of both approaches for a given range of the substrate concentration S (kinetic expressions and kinetic parameters are the same in both cases).

A comparison of the growth rate indicates that only a slightly higher growth rate could be achieved in the optimal case. This is based on the observation that the protein fraction that is allocated to the T fraction is always higher than that for the R fraction which is not the case in **Figure 3B**.

4 L-PHENYLALANINE PRODUCTION WITH *ESCHERICHIA COLI*

The proposed coarse-grained model approach can be used to model a biotechnological production process. In the scope of this

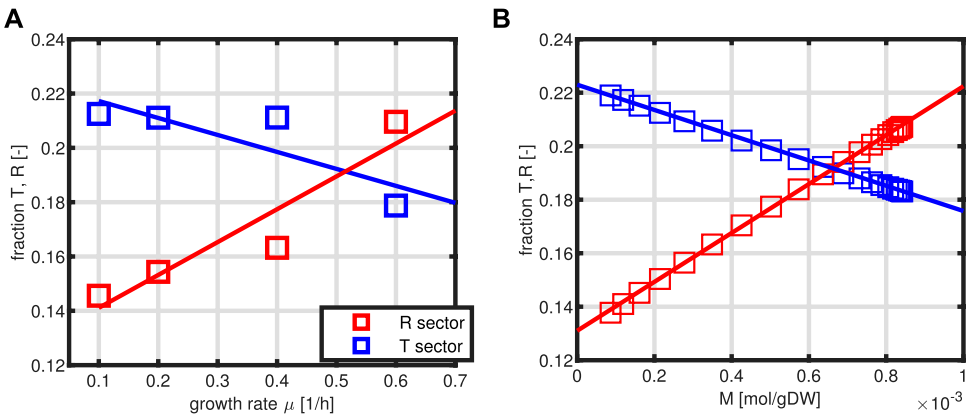


FIGURE 3 | Relationship between the sectors *T* and *R* as a function of the specific growth rate μ. Experimental data are taken from the study by Schmidt et al., (2016) with protein representing 50% of the biomass **(A)**. Estimated kinetics of first order for *T/P* and *R/P* as a function of variable *M* **(B)**.

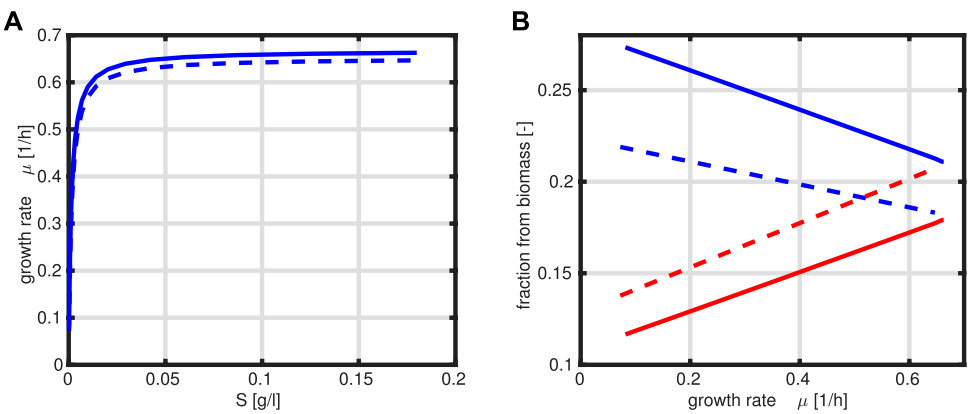


FIGURE 4 | Specific growth rate μ as a function of substrate *S*, where the solid line represents the optimal case **(A)**. Fractions *R* (red) and *T* (blue) as a function of the growth rate μ, where the solid line represents the optimal case, and dashed lines are from presented data in **Figure 3B**.

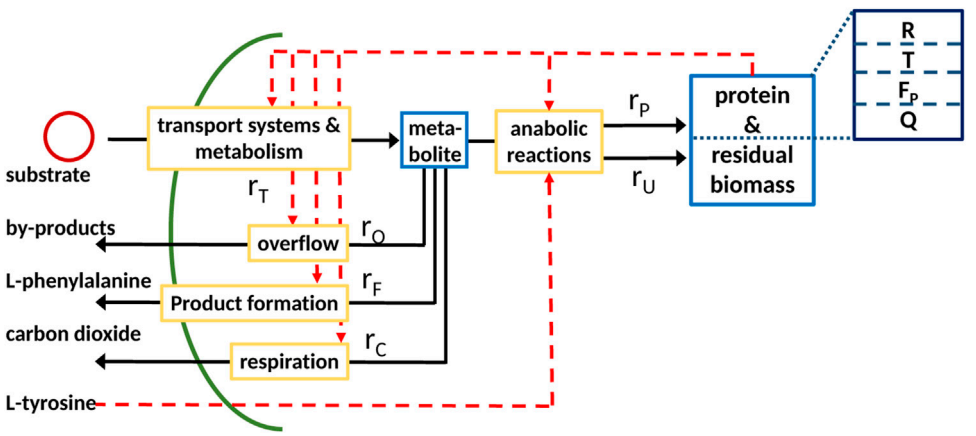


FIGURE 5 | Scheme of the coarse-grained model expanded to include the formation of L-phenylalanine, respiration r_C, and residual biomass *U*.

research, we consider an L-phenylalanine producing *Escherichia coli* strain with glycerol as the substrate and decoupled biomass and product formation due to L-tyrosine auxotrophism, meaning biomass is only formed if L-tyrosine is available (Sprenger (2007), Weiner et al. (2014)). A more detailed description of the strain used can be found in the *Material and Methods* section. The L-phenylalanine production process is considered here as an example for a bioreactor production process. Due to the nature of coarse-grained models, the resulting model can easily be adapted to depict other production processes. The basic model is extended to include an additional rate r_F describing L-phenylalanine formation and a corresponding protein sector F as seen in **Figure 5**. Furthermore, we consider respiration implemented as rate r_C and the residual biomass fraction U.

The system (**Eqs 2–4**) is represented as follows:

$$N = \begin{pmatrix} \alpha & -\gamma & -1 & -\delta & -2 & -1 \\ 0 & \beta & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \epsilon & 0 & 0 \end{pmatrix} \tag{24}$$

with the differential equations

$$\begin{pmatrix} \dot{M} \\ P \\ U \end{pmatrix} = N \begin{pmatrix} r_T \\ r_P \\ r_O \\ r_U \\ r_F \\ r_C \end{pmatrix} - \mu \begin{pmatrix} M \\ P \\ U \end{pmatrix}. \tag{25}$$

The next step is to determine the reaction rates. The rates concerning the central metabolism and overflow remain the same as in the basic model (**Eq. 21**). The L-phenylalanine production rate is dependent on the F fraction and the respiration rate r_C on the T fraction as it is part of the central metabolism. In order to incoporate the L-tyrosine auxotrophism, the protein synthesis rate r_P and synthesis rate of residual biomass r_U are modified to be multiplied with a function which is 1 if L-tyrosine is available and otherwise set to a low value of 0.1, corresponding to a low biomass formation during this phase, since it cannot be practically ensured that no L-tyrosine is available during this phase. This leads to the following set of reaction rates

$$\begin{aligned} r_T &= k_T \frac{S}{S + K_T} T \\ r_P &= k_P \frac{M}{M + K_P} R\, \tau(A) \\ r_O &= k_O M T \\ r_U &= k_U \frac{M}{M + K_U} R\, \tau(A) \\ r_F &= k_F M F_P \\ r_C &= k_C M T, \end{aligned} \tag{26}$$

where A is the available L-tyrosine and

$$\tau(A) = \begin{cases} 1, & A > 0 \\ 0.1, & A \leq 0. \end{cases} \tag{27}$$

As the reaction rates are determined by the composition of the proteome, we can take advantage of the observations from the previous section. The allocation of T and R in the proteome is given by the estimated linear function of M as seen in **Figure 4**. A part of the proteome is allocated to fraction F after induction, which is accomplished through the shift of biomass production to product formation due to lack of L-tyrosine in the feed at time t_{ind} with delay as follows:

$$F_P = \phi(t)\, F_{\max}\, P, \tag{28}$$

where

$$\phi(t) = \frac{t - t_{ind}}{(t - t_{ind}) + t_\phi}. \tag{29}$$

The remaining protein fraction Q is not of further interest.

Now that we have formulated a system for the intracellular components forming the total biomass X, equations can be set up to model a complete bioprocess consisting of two process phases: a biomass production phase, followed by a batch phase, followed by two fed-batch phases with two different feeding solutions, and an L-phenylalanine production process phase which was initiated with induction of the cells with IPTG. For the process, we assume ideal mixing conditions in a bioreactor of the volume V with feeding rate q_{in} as seen in **Figure 6**

$$\dot{V} = q_{in} \tag{30}$$

and only one feeding substrate (glycerol) S in [g/l] with feed substrate concentration S_{in}

$$\dot{S} = \frac{q_{in}}{V} (S_{in} - S) - r_T\, X\, w_S, \tag{31}$$

where the first term represents the ingoing substrate and dilution due to volume change and the second term the substrate uptake by the cells. The equation for the biomass X is given by

$$\dot{X} = \mu\, X - \frac{q_{in}}{V} X, \tag{32}$$

and the product equations are

$$\dot{F} = r_F\, X\, w_F - \frac{q_{in}}{V} F, \tag{33}$$

$$\dot{O} = r_O\, X\, w_O - \frac{q_{in}}{V} O, \tag{34}$$

with acetate O as the exemplary byproduct. Feeding profiles (**Figure 6**) determining the variables q_{in} and S_{in} and L-tyrosine concentrations during the process and initial values for the differential **Eqs 30–34** are obtained from experimental data (see Supplementary Material). The bioprocess has been run in a stirred-tank bioreactor with a starting volume of $V_0 = 1$ l; thus, we can that assume the environment in the bioreactor is well-mixed, and the description of the biomass by an average cell, as in the model presented, is sufficient. Experimental data suggest a stop in both biomass and product formation and a high accumulation of by-products after process time point $t = 71$ h (experimental data for the full process can be found in Supplementary Material). As reasons for this behavior have not been investigated at this point, the mechanistics to depict this are not incorporated in the model. Using **Eqs 25–34**, a numerical simulation up to process time $t = 71$ h was performed using MATLAB R2020a with ode15s as ordinary differential

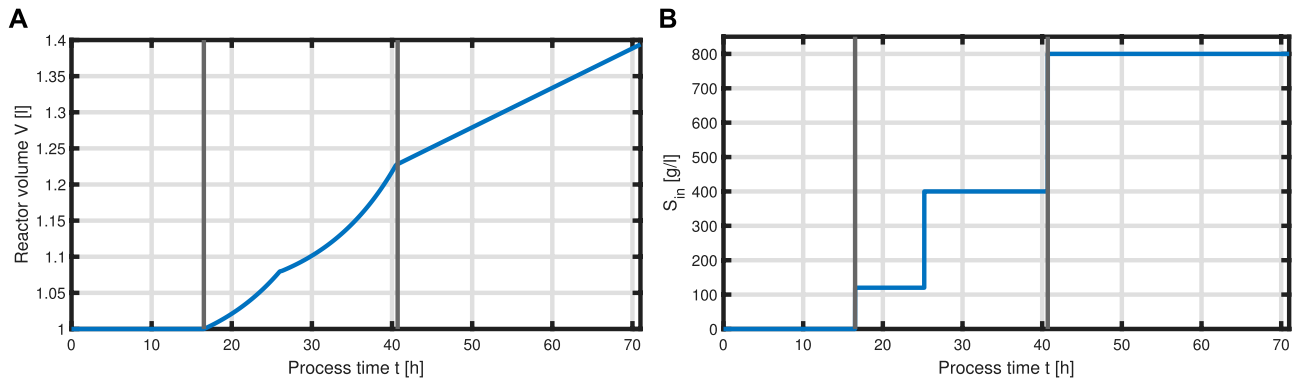


FIGURE 6 | Feeding profile of the process. Bioeactor volume V **(A)** and substrate concentration (glycerol) of the feed S_{in} **(B)** over the time course of the process, where vertical lines indicate the three process phases (batch phase, fed-batch phase, and production phase with constant feeding).

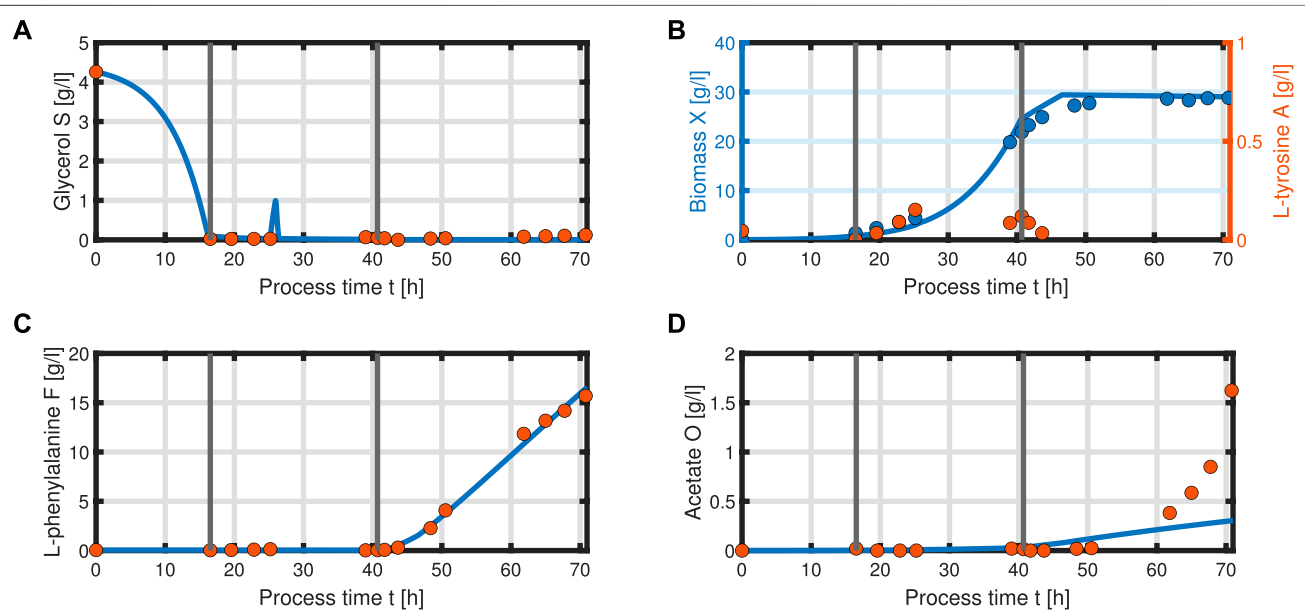


FIGURE 7 | Comparison of the simulated quantities (solid blue line) against the experimental data (points) of the L-phenylalanine production process up to $t = 71$ h. Time course of the following concentrations: glycerol S **(A)**, biomass X, L-tyrosine A **(B)**, L-phenylalanine F **(C),** and acetate O as representative of the by-products of the process **(D)**. The data points of L-tyrosine, which are negative due to insufficient measurement sensitivity, are set to zero.

equation solver and was compared to experimental results (**Figure 7**). Analogous to the experiments, the simulation is divided into different process phases, and slightly different parameter sets were used for the biomass and L-phenylalanine production phase. The parameter set for the product formation phase contains higher reaction constants for byproduct formation and respiration. After each process phase, the solution of the end point was used as the initial value for the differential equations for the next process phase. Parameter values used for this simulation can be found in *Supplementary Materials*. **Figure 7** shows good agreement between the simulated concentrations of substrate S, biomass X, L-phenylalanine F, and the experimental data. The peak of substrate concentration S during the fed-batch phase can be explained by the change of substrate concentration in the feed. The simulation of acetate concentration O shows the right trend, although not the exact behavior, of the by-products. Nevertheless, for the given parameter sets, the model can reproduce the overall dynamics of the L-phenylalanine production process. Besides the measurable quantities, the model can provide the intracellular concentrations as seen in **Figure 8B**, where the macromolecules P and U make up most of the biomass with both occupying nearly half of the biomass, and the mass fraction of the metabolites is negligible compared to that of the macromolecules, especially in the product formation phase. With a closer look at the allocation of the proteome, the fraction R follows the behavior of the metabolites M, decreases over the course of the process, and

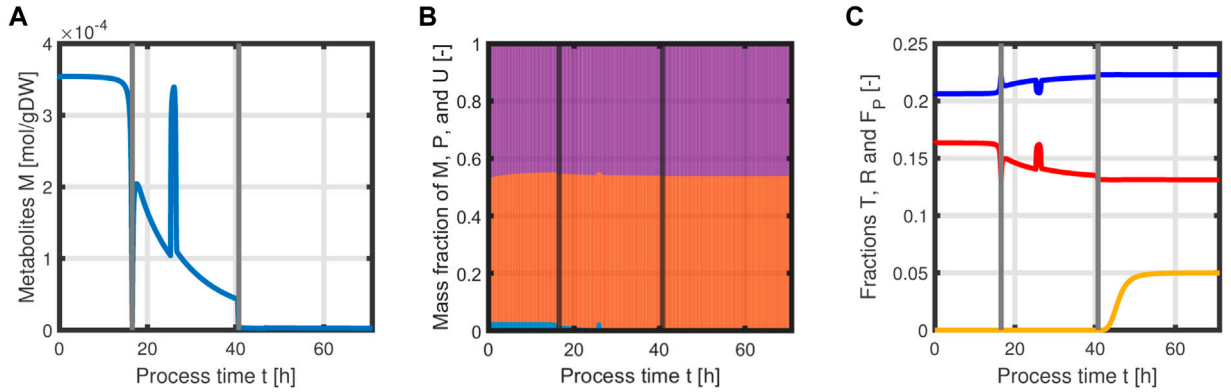


FIGURE 8 | Time course of intracellular concentrations of the L-phenylalanine process: metabolites *M* **(A)**, mass fractions of proteins *P* (violet), and residual biomass *U* (orange) and metabolites *M* (blue), where *M* becomes negligible in the last process phase **(B)**. Mass fractions of *T* (blue), *R* (red), and *F* (yellow) over the course of the process **(C)**.

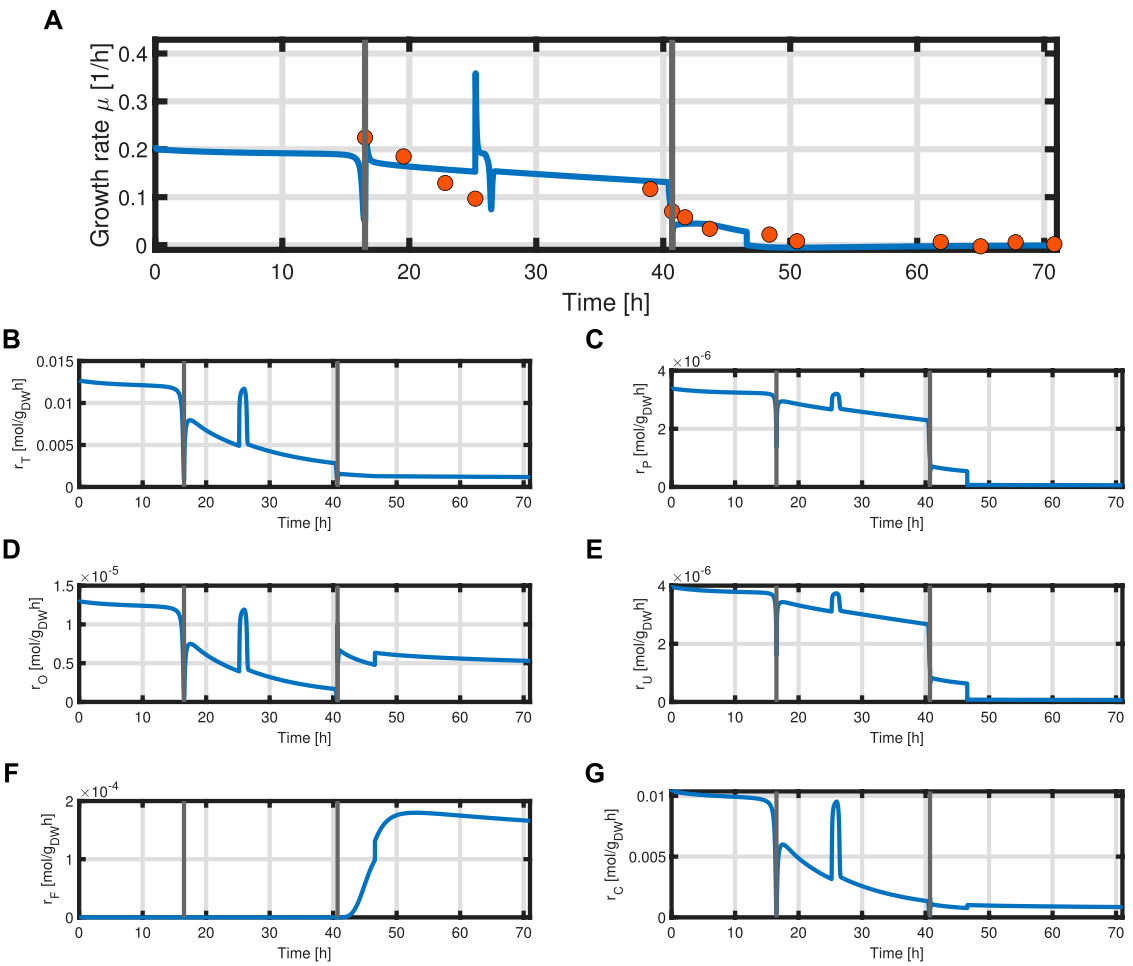


FIGURE 9 | Time course of the simulated specific growth rate μ (blue) and the point-wise calculated specific growth rate obtained from experimental data indicated as orange dots **(A)**, substrate transport rate r_T **(B)**, protein synthesis rate r_P **(C)**, overflow metabolism rate r_O **(D)**, residual biomass synthesis rate r_U **(E)**, product formation rate r_F **(F),** and respiration rate r_C **(G)**.

remains at a constant level during the product formation phase with fraction T forming the counterpart. The fraction for product formation F follows the description of **Eqs 28**, **29**.

In addition, the specific growth rate μ and the different reaction rates can be obtained from the model (**Figure 9**). The simulated growth rate μ roughly follows the trend of the growth rate pointwise calculated from the experimental data of biomass as seen in **Figure 9A**. The calculated growth rate has to be taken with caution as each point is calculated from two consecutive points with a large time difference and can heavily deviate from the actual growth rate. One can observe that due to the dependency of all rates on the metabolites M and the constant protein fraction T, all reaction rates follow the course of the metabolites during the biomass production phase (**Figures 9B–D**). In the product formation phase, the rates for the synthesis of all macromolecules deviate from the course of the metabolites as it is determined by the L-tyrosine auxotrophism (**Figures 9C,D**).

5 OPTIMAL BY-PRODUCT SECRETION

The second example considers the optimal production of a metabolite (in this case M) that is excreted into the medium *via* reaction r_3 (see **Figure 1**). Since M represents a metabolite from central metabolic pathways, it could stand for ethanol, acetate, or succinate which are all interesting biotechnological products. The stoichiometry and parameters are the same as in **Eqs. 16–19**. For the simulation study, and for a fair comparison of the outcoming results, the following conditions are fixed:

- A fed-batch process in a bioreactor is considered with a flexible input profile for the incoming substrate feed rate $q_{in}(t)$ as a function of time and a fixed-end time $t_{end} = 20$ h. With the feed, the substrate concentration can be adjusted in such a way that the metabolite is excreted at best. In contrast to a batch process, the substrate is fed into the medium and, therefore, high sugar concentrations in the beginning (as for the batch process) are avoided. Since a continuous process requires much more time, a steady state is reached normally first after five times the respective time constant (in our case approx. 20 h); this type of process design is also not considered here.
- The initial conditions are set fix for all model state variables.
- The bioreactor has a maximum working volume of 5 l, while in the beginning, the experiment starts with 1 l.
- The objective function is the amount of product expressed in mole at t_{end}: $M_{ex}\ V$.

For the study, three different profiles are investigated. 1) A standard procedure, often applied in bioprocess engineering tries to feed the substrate in an exponential way to keep the specific growth rate μ constant. This requires that the substrate concentration in the bioreactor is nearly constant. The differential equation for the substrate S with function q_{in} and feed concentration S_{in} reads as follows:

$$\dot{S} = \frac{q_{in}}{V}\ (S_{in} - S) - r_1\ X, \tag{35}$$

and after setting this equation to zero, a function for q_{in} can be obtained:

$$q_{in} = \frac{r_1\ X\ V}{S^{in} - S}; \tag{36}$$

with $m_X = X\ V$ and a constant growth rate μ_0 for this condition, we get

$$q_{in} = \frac{r_1\ m_{X0}\ e^{\mu_0\ t}}{S^{in} - S}. \tag{37}$$

Typically, r_1 is estimated given the biomass yield coefficient Y during the batch phase and since the current substrate concentration is low (due to a small half-saturation value for substrate uptake), we finally obtain

$$q_{in} = \frac{\mu_0\ m_{X0}\ e^{\mu_0\ t}}{Y\ S^{in}}. \tag{38}$$

The feeding profile is applied after the end of the batch phase.

2) The second profile uses a polynomial function of time for the feeding rate:

$$q_{in} = \sum_{i=1}^{4} a_i\ t^i, \tag{39}$$

with four parameters a_i to optimize.

3) The last profile is a piecewise linear profile with six fixed switching points t_k:

$$q_{in} = q_{in}^k = const. \quad \text{for } t_{k-1} \le t \le t_k. \tag{40}$$

Here, the values q_{in}^k are the parameters that have to be optimized.

First, a typical outcome for the standard case is shown in **Figure 10**. After 3.5 h, the substrate runs out and the feeding starts. The batch phase is characterized by very low productivity while the growth rate is at its maximum. In addition, during this phase, the R fraction is high (as already shown above) and, therefore, due to the coupling to the T fraction, the rate of byproduct formation is low. After starting the feeding, the intracellular system switches to high values for the T fraction Next, a comparison of the feeding profiles and the final value of the objective function is shown in **Figure 11**. Although the profiles are different, the final values of the objective function are comparable.

The simulation studies are based on a fixed set of kinetic parameters and, so far, did not consider any uncertainties with respect to the quality of these. Typically, kinetic parameters are obtained by parameter identification and subsequent parameter estimation and analysis. For our example at hand, in the next step, the uncertainty of the kinetic parameters is taken into account during the optimization procedure. Important parameters of the model are the maximal reaction velocities $r_{\max}$ for all reactions. To consider these uncertainties, an ensemble of 20 models is generated during each iteration step of the optimization. This results also in an ensemble for the values of the

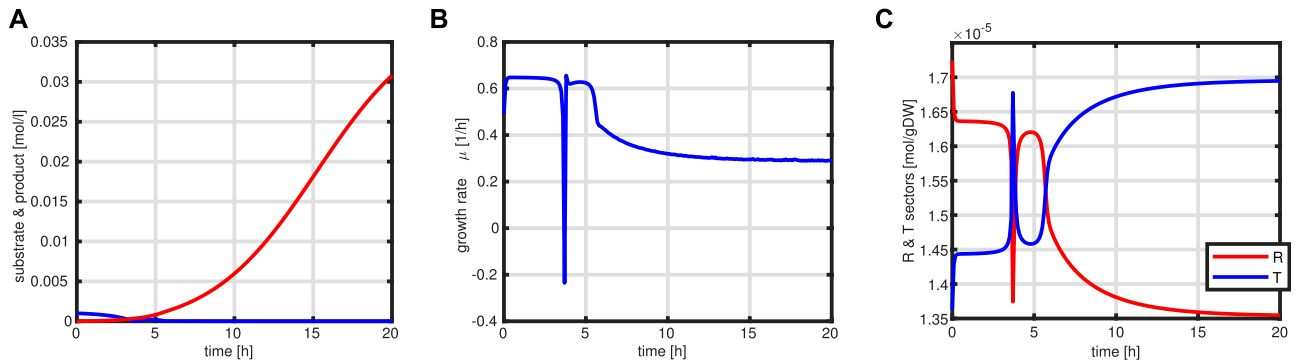


FIGURE 10 | Simulation outcome for the standard fed-batch process. Time course of substrate (blue) and product (red) in the medium **(A)**. Growth rate μ **(B)**. Time course of the protein sectors R (red) and T (blue) **(C)**.

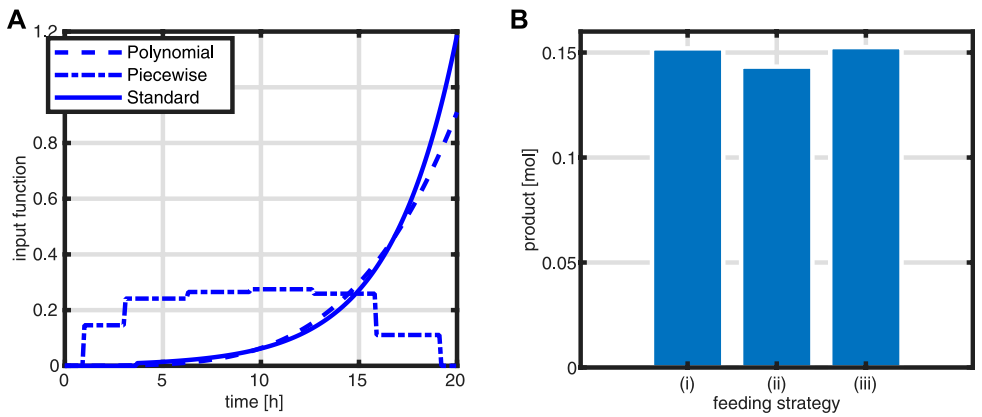


FIGURE 11 | Comparison of the outcome of the three strategies. Time course for the feeding q_{in} as a function of time **(A)**; value of the objective function (from left to right: standard feeding, polynomial function, and piece-wise function) **(B)**.

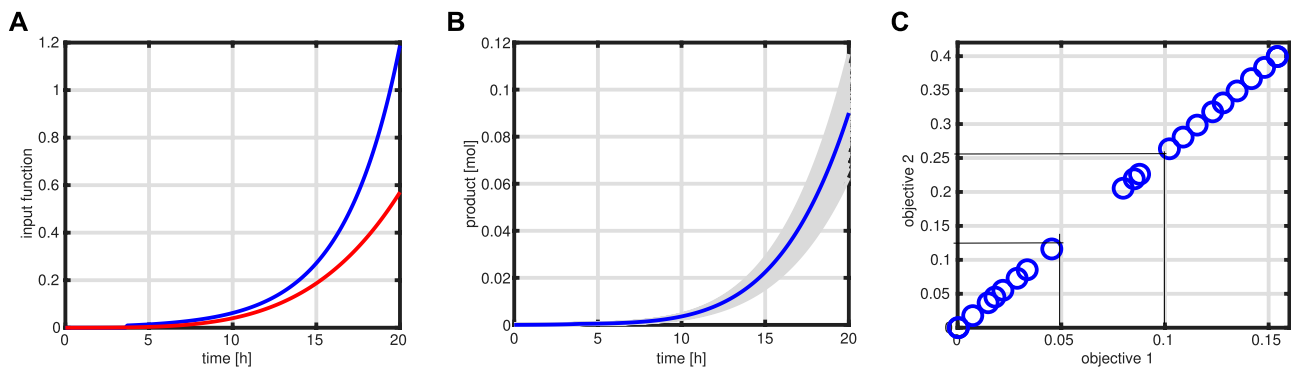


FIGURE 12 | Robust optimization. Input profile (red) for the robust case in comparison to the standard case **(A)**. Product course of time for 100 simulations (in gray) with the variation in the maximal rate of the enzymatic processes and simulation with standard parameter (blue). **(B)**. Pareto front for two objective functions: product amount at time t_{end} and total amount of the substrate needed to produce the product. Thin lines indicate the cost change for an increase in product amount as described in the text **(C)**.

objective function Φ, following the approach proposed by Nimmegeers et al., (2016), and the objective function in this case is given by

$$\max \ \mathrm{E}[\Phi] - \alpha \ \mathrm{Var}[\Phi], \tag{41}$$

with a weighting factor α, expectation E, and variance Var. With this formulation, stronger variations in the values of the objective function, expressed in the variance of Φ, are penalized. Since the feeding profile is geared to the growth rate, a much more conservative output is expected; if the substrate uptake, for example, would be higher than expected, more biomass would be produced, and few byproducts will be released. As can be seen in left of **Figure 12**, the input function (red curve) (and with this also the bioreactor volume) is lower than that in case of the standard procedure (blue curve). On the right side, different outcomes (in gray) for the variable product Pr are shown for 100 simulation runs, together with a simulation of the nominal values (blue curve).

Besides the optimization of the product at t_{end}, also the cost of substrate is of interest. In a fed-batch process, the total amount of substrate S_{tot} fed, considering the substrate concentration at the beginning S_0, can be calculated by

$$S_{tot} = \int_{t_1}^{t_{end}} q_{in}(t) \ S_{in} \ dt + S_0. \tag{42}$$

The outcome of having two objective functions, maximization of product at t_{end}, and minimization of the substrate cost for the entire process can be presented with a Pareto front that is shown in **Figure 12**. As can be seen, a nearly perfect linear relationship is detected (it is to be noted that both objective functions have unit mole); increasing the product, for example, for 0.05 mol units results in cost for the substrate of 0.13 mol. In this way, an economic assessment of the process is possible.

6 DISCUSSION

In bioprocess engineering, the design of experiments often is based on a mathematical description. While simple growth models taking into account only biomass, substrate, and product often are insufficient to describe the observed dynamics, whole-cell models are cumbersome and are difficult to calibrate. A good model comprises coarse-grained models because they are simple in the model structure but take into account the most important cellular processes. In this study, we propose an approach to use coarse-grained models based on a strict mass conservation to model bacterial growth as a basis for metabolic engineering applications. In this way, classical flux analysis could be extended to take into account fluxes into macromolecules such as the proteome. Additional solutions are provided by the null space of the mass matrix W that requires information on the mass fraction of the components of the model. With a simple example, we could show that depending on the mass composition of the cell, larger differences in the flux distribution in comparison to the standard approach could appear. With the condition of strict mass conservation, we are also able to provide a general solution for a cellular network independently of its internal structure.

As the focus of this research is coarse-grained models, we provided a formulation for a minimal model whose structure is in accordance with that in previous studies Sharma et al. (2018), Bertaux et al. (2020). Steady-state solutions are determined based on measured data (Schmidt et al., 2016) for the molecular composition of the cell. Especially, we consider two main protein fractions, an R fraction representing the transcription and translation apparatus and a T fraction, taking into account metabolic and transport enzymes. The outcome of these simulation studies is compared with that of a model where the R fraction is an adjustable quantity in an optimization program (the sum of the R fraction and T fraction is taken as constant). The results of the optimization program show that an optimal allocation of proteins led to a slightly higher growth rate with a comparable course of the R fraction as a function of the specific growth rate, and we conclude that the measured data are in good agreement with the expectation of an efficient and optimal acting organism.

Many studies have dealt with the derivation of growth laws under various conditions Klumpp et al. (2013), Bosdriesz et al. (2015), Hui et al. (2015). Based on the structure of the minimal model, we have expanded the model to include the dynamic environment in a bioreactor system that allows us to realize also different process design strategies such as feeding or continuous culture. Experimental data from an L-phenylalanine production process are taken as an example for parameter identification and estimation, and a good agreement between simulation and experimental data is obtained. A different design problem was addressed by finding optimal input profiles if the production of a metabolite from central pathways is of interest. Here, also, parametric uncertainties can be taken into account that leads to a much more conservative input profile. To summarize, coarse-grained models are a sound basis for the development of bioprocesses due to their simple structure with only a minor number of parameters and the flexibility to simulate and optimize different biotechnological process designs.

MATERIAL AND METHODS FOR EXPERIMENTAL CULTIVATION OF TRIPLE REPORTER STRAIN

Strain

For the L-phenylalanine production process, in a stirred-tank bioreactor of 3.6 l working volume, a recombinant *Escherichia coli* FUS4 ($pF81_{kan}$) strain was used as described by Gottlieb et al. (2014). This is a genetically modified strain with auxotrophies for L-phenylalanine and L-tyrosine by deletion of the chromosomal genes *aroF*, *pheA*, and *tyrA* (decoding for a DAHP synthase, bifunctional chorismate mutase / prephenate dehydratase, and a t-protein, respectively) along the aromatic biosynthesis pathway. Simultaneously, it harbors the $pF81_{kan}$ plasmid decoding for the genes *aroF*, *pheA*, *aroB* (3-dehydroquinate synthase), and *aroL* (shikimate kinase 2) under the control of an inducable *tac* promoter. Furthermore, kanamycin resistance is integrated as the selection marker Gottlieb et al. (2014), Weiner et al. (2014).

Cultivation Media

The cells were cultivated in a defined minimal medium with glycerol as the sole carbon source. All the components with their corresponding concentrations as well as its production protocol were adapted from the study by Weiner et al. (2014).

Preculture Strategy

Provision of cell biomass for the inoculation for the cultivation in a stirred-tank bioreactor was realized by a two-step preliminary cultivation in shake flasks. First, a single colony of cells grown on minimal medium agar plates (> 66 h at 37°C) was picked for inoculation of a single 100-ml shake flask with 10 ml minimal medium and cultivated at 37 °C and 150 rpm in an orbital shaker (Multitron, Infors HT, Switzerland) for 24 h. Afterward, the cells were transferred for further cultivation in two 500-ml shake flasks with 100 ml minimal media each and a starting optical density at 600 nm (OD_{600}) of 0.01. After incubation at 37°C and 250 rpm for at least 24 h, the cells were centrifuged (4,500 rpm, 10 min) and resuspended in fresh minimal medium. These cell suspensions were used for inoculation of cultivations in the stirred-tank bioreactor with a starting OD_{600} of 0.1.

Bioreactor Cultivation

For laboratory-scale cultivation of recombinant *E. coli* FUS4 ($pF81_{kan}$) for L-phenylalanine production, a 3.6 glass stirred-tank bioreactor was used (Infors HT, Switzerland). The bioreactor was equipped with two six-bladed flat-blade turbines and three equidistant baffles. The minimal medium for cultivation was prepared *ex situ* and pumped into the bioreactor under sterile conditions to a starting volume of 1 l. The temperature was kept at 37 °C. 42% phosphoric acid and 25% ammonia were used as titration solutions to keep the pH at 7 ± 0.1. Dissolved oxygen levels above 30% were provided by step-wise increase of either stirrer speed or aeration rate up to 1,500 rpm and 5 l/min, respectively. Furthermore, an antifoam probe was used for controlled titration of antifoam solution, if necessary (AF204, 1:10 diluted, Sigma Aldrich, United States). The cultivation started with an initial batch phase. After depletion of glycerol, an exponential feeding was set for the biomass production phase with a defined growth rate of μ_{set} = 0.1 1/h. Two fed-batch media were used with either 120 g/l glycerol, 2.5 g/l L-phenylalanine, 3.6 g/l L-tyrosine, 60 g/l ammonium sulfate, and 0.1 g/l kanamycin (fed-batch medium 1) or 400 g/l glycerol, 1.11 g/l L-phenylalanine, 3.8 g/l L-tyrosine, 25 g/l ammonium sulfate, and 0.1 g/l kanamycin (fed-batch medium 2). The former and the latter were titrated with 25% ammonia or 5 M potassium hydroxide to allow complete dissolution of L-tyrosine. After a sufficiently high biomass concentration of at least 20 g/l was provided, the cells were induced with 0.3 mM IPTG. Fed-batch medium 3 with the components 800 g/l, 8 g/l ammonium sulfate, 8 g/l ammonium phosphate, and 0.1 g/l kanamycin was then constantly provided with a rate of 0.18 $g_{glycerol}/g_{Biomass}h$. At the start of each fed-batch media supply, the concentrated media components calcium chloride dihydroxide (15 g/l) and iron(II) sulfate heptahydrate with sodium citrate dihydrate (22.5 g/l and 200 g/l), magnesium sulfate heptahydrate (300 g/l), and thiamine hydrochloride (7.5 g/l) were mixed in a 1:5:1:1 ratio and injected to the fermentation broth *via* a septum. For the start of fed-batch phase 1, 2, or 3, a mixture of 4.8, 9.6, and 8.8 ml were injected, respectively Weiner et al. (2014).

Analytics

Cell dry weights were measured gravimetrically. Pre-weighted dried 2-ml microcentrifuge tubes (80 °C for at least 24 h) were used for centrifugation of 2 ml of the cell suspension (21130 × g, 20 min, 4°C). The supernatant was further used for sample preparation for high-performance liquid chromatographies (HPLCs) or discarded. The cell pellet was dried again at 80°C for at least 24 h. The biomass concentration was calculated by the difference of weight between the microcentrifuge tube with dried cell pellets and the empty microcentrifuge tube. For the quantification of the amino acid concentrations of L-phenylalanine and L-tyrosine as well as for the organic compounds glycerol, acetate, lactate, succinate, pyruvate, malate, and ethanol, two different HPLCs were used. The samples for HPLC analysis were prepared by filtration of the supernatant of each sample through a 0.2-μm filter and were stored at 4°C upon measurement. The quantification method for both amino acids is already described by Weiner et al. (2014) and was adapted from there. The organic compounds were quantified using a HPLC (Prominence-i LC-2030C, Shimadzu, Japan) with an ion-exchange column (Aminex HPX-87H 300 mm × 7.8 mm, Bio-Rad, CA, United States) and a refractive index detector (RID-20A, Shimadzu, Kyoto, Japan). 10 μl of samples was injected to an isocratic flow of 0.6 ml/min and 5 mM sulfuric acid as mobile phase with a constant temperature of 60°C. The quantification of each component was realized by measurement of external standards.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

AK and AH designed and supervised the study. DD performed null space calculations, simulation studies, and parameter estimation. AK developed the concept and performed simulation studies. MH performed the experimental studies for the L-phenylalanine process.

FUNDING

Funding for DD and MH was provided by the DFG priority program SPP 2170 "Novel production processes and multiscale analysis, modeling, and design of cell–cell and cell–bioreactor interactions (InterZell)."

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.806213/full#supplementary-material And the used MATLAB files in a repository under the following link: https://sourceforge.net/projects/coarse-grained-modeling/

REFERENCES

Basan, M., Hui, S., Okano, H., Zhang, Z., Shen, Y., Williamson, J. R., et al. (2015). Overflow Metabolism in escherichia Coli Results from Efficient Proteome Allocation. *Nature* 528, 99–104. doi:10.1038/nature15765

Bertaux, F., von Kügelgen, J., Marguerat, S., and Shahrezaei, V. (2020). A Bacterial Size Law Revealed by a Coarse-Grained Model of Cell Physiology. *Plos Comput. Biol.* 16, e1008245. doi:10.1371/journal.pcbi.1008245

Bollenbach, T., Quan, S., Chait, R., and Kishony, R. (2009). Nonoptimal Microbial Response to Antibiotics Underlies Suppressive Drug Interactions. *Cell* 139, 707–718. doi:10.1016/j.cell.2009.10.025

Bosdriesz, E., Molenaar, D., Teusink, B., and Bruggeman, F. J. (2015). How Fast-growing Bacteria Robustly Tune Their Ribosome Concentration to Approximate Growth-rate Maximization. *Febs J.* 282, 2029–2044. doi:10.1111/febs.13258

Giordano, N., Mairet, F., Gouzé, J.-L., Geiselmann, J., and de Jong, H. (2016). Dynamical Allocation of Cellular Resources as an Optimal Control Problem: Novel Insights into Microbial Growth Strategies. *Plos Comput. Biol.* 12, e1004802. doi:10.1371/journal.pcbi.1004802

Gottlieb, K., Albermann, C., and Sprenger, G. A. (2014). Improvement of L-Phenylalanine Production from Glycerol by Recombinant escherichia Coli Strains: the Role of Extra Copies of Glpk, Glpx, and Tkta Genes. *Microb. Cel Fact* 13, 96. doi:10.1186/s12934-014-0096-1

Hui, S., Silverman, J. M., Chen, S. S., Erickson, D. W., Basan, M., Wang, J., et al. (2015). Quantitative Proteomic Analysis Reveals a Simple Strategy of Global Resource Allocation in Bacteria. *Mol. Syst. Biol.* 11, 784. doi:10.15252/msb.20145697

Klumpp, S., Scott, M., Pedersen, S., and Hwa, T. (2013). Molecular Crowding Limits Translation and Cell Growth. *Proc. Natl. Acad. Sci.* 110, 16754–16759. doi:10.1073/pnas.1310377110

Maitra, A., and Dill, K. A. (2015). Bacterial Growth Laws Reflect the Evolutionary Importance of Energy Efficiency. *Proc. Natl. Acad. Sci. USA* 112, 406–411. doi:10.1073/pnas.1421138111

Molenaar, D., van Berlo, R., de Ridder, D., and Teusink, B. (2009). Shifts in Growth Strategies Reflect Tradeoffs in Cellular Economics. *Mol. Syst. Biol.* 5, 323. doi:10.1038/msb.2009.82

Nimmegeers, P., Telen, D., Logist, F., and Impe, J. V. (2016). Dynamic Optimization of Biological Networks under Parametric Uncertainty. *BMC Syst. Biol.* 10, 86. doi:10.1186/s12918-016-0328-6

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What Is Flux Balance Analysis. *Nat. Biotechnol.* 28, 245–248. doi:10.1038/nbt.1614

Pandey, P. P., and Jain, S. (2016). Analytic Derivation of Bacterial Growth Laws from a Simple Model of Intracellular Chemical Dynamics. *Theor. Biosci.* 135, 121–130. doi:10.1007/s12064-016-0227-9

Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrné, E., Volkmer, B., Callipo, L., et al. (2016). The Quantitative and Condition-dependent *Escherichia coli* Proteome. *Nat. Biotechnol.* 34, 104–110. doi:10.1038/nbt.3418

Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z., and Hwa, T. (2010). Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science* 330, 1099–1102. doi:10.1126/science.1192588

Scott, M., Klumpp, S., Mateescu, E. M., and Hwa, T. (2014). Emergence of Robust Growth Laws from Optimal Regulation of Ribosome Synthesis. *Mol. Syst. Biol.* 10, 747. doi:10.15252/msb.20145379

Sharma, P., Pandey, P. P., and Jain, S. (2018). Modeling the Cost and Benefit of Proteome Regulation in a Growing Bacterial Cell. *Phys. Biol.* 15, 046005. doi:10.1088/1478-3975/aabe43

Sprenger, G. A. (2007). From Scratch to Value: Engineering escherichia Coli Wild Type Cells to the Production of L-Phenylalanine and Other fine Chemicals Derived from Chorismate. *Appl. Microbiol. Biotechnol.* 75, 739–749. doi:10.1007/s00253-007-0931-y

Weiner, M., Tröndle, J., Albermann, C., Sprenger, G. A., and Weuster-Botz, D. (2014). Carbon Storage in recombinantEscherichia Coliduring Growth on Glycerol and Lactic Acid. *Biotechnol. Bioeng.* 111, 2508–2519. doi:10.1002/bit.25306

Weiße, A. Y., Oyarzún, D. A., Danos, V., and Swain, P. S. (2015). Mechanistic Links between Cellular Trade-Offs, Gene Expression, and Growth. *Proc. Natl. Acad. Sci. USA* 112, E1038–E1047. doi:10.1073/pnas.1416533112

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Doan, Hoang, Heins and Kremling. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 14 March 2022
doi: 10.3389/fmolb.2022.801011

GAL Regulon in the Yeast *S. cerevisiae* is Highly Evolvable *via* Acquisition in the Coding Regions of the Regulatory Elements of the Network

Rajeshkannan[†], *Anjali Mahilkar*[†] *and Supreet Saini* *

Department of Chemical Engineering, Indian Institute of Technology Bombay, Mumbai, India

GAL network in the yeast *S. cerevisiae* is one of the most well-characterized regulatory network. Expression of GAL genes is contingent on exposure to galactose, and an appropriate combination of the alleles of the regulatory genes GAL3, GAL1, GAL80, and GAL4. The presence of multiple regulators in the GAL network makes it unique, as compared to the many sugar utilization networks studied in bacteria. For example, utilization of lactose is controlled by a single regulator LacI, in *E. coli*'s lac operon. Moreover, recent work has demonstrated that multiple alleles of these regulatory proteins are present in yeast isolated from ecological niches. In this work, we develop a mathematical model, and demonstrate via deterministic and stochastic runs of the model, that behavior/gene expression patterns of the cells (at a population level, and at a single-cell resolution) can be modulated by altering the binding affinities between the regulatory proteins. This adaptability is likely the key to explaining the multiple GAL regulatory alleles discovered in ecological isolates in recent years.

Keywords: *S. cerevisiae*, galactose, GAL regulon, gene expression kinetics, evolvability

OPEN ACCESS

Edited by:
Alberto Jesus Martin, Universidad Mayor, Chile

Reviewed by:
Mikael Molin, Chalmers University of Technology, Sweden
Sandeep Choubey, Institute of Mathematical Sciences, Chennai, India

***Correspondence:**
Supreet Saini
saini@che.iitb.ac.in

[†]*These authors have contributed equally to this work*

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *24 October 2021*
Accepted: *22 February 2022*
Published: *14 March 2022*

Citation:
Rajeshkannan, Mahilkar A and Saini S (2022) GAL Regulon in the Yeast S. cerevisiae is Highly Evolvable via Acquisition in the Coding Regions of the Regulatory Elements of the Network. Front. Mol. Biosci. 9:801011. doi: 10.3389/fmolb.2022.801011

INTRODUCTION

Genetic circuits are evolvable. Depending on the precise environmental niches, acquisition of a mutation could alter gene expression dynamics more suited for survival and growth. The changes in the network dynamics could be facilitated by two types of mutations. First, mutations which change processes like transcription and translation, and hence, shape regulatory networks (Wray, 2007; Kim and Przytycka, 2012; Hill et al., 2021). These mutations change the timing and levels of transcription and translation. On the other hand, mutations could also change protein activity, and as a result the affinity of a protein with DNA or another protein; resulting in downstream changes in gene expression (Golding and Dean, 1998). While several examples of the first kind are known, relatively fewer examples of changes in expression patterns by protein modifications are known (Lewontin, 2002; Rodriguez-Trelles et al., 2003).

The GAL regulon in the yeast *Saccharomyces cerevisiae* (*S. cerevisiae*), which enables the organism to utilize and grow on galactose, is one of the most well-studied regulatory networks in yeast (Bhat and Murthy, 2001). The regulatory network, is briefly explained below (**Figure 1**). All genes involved in utilization of galactose are under the control of a transcriptional regulator, GAL4p. Gal4p binds to and drive expression from several promoters, which control expression of the GAL regulon (Chasman and Kornberg, 1990). In the absence of galactose, Gal80p binds Gal4p, leading to formation of a Gal80p-Gal4p protein complex. This sequestration of Gal4p thus switches OFF expression from promoters of the

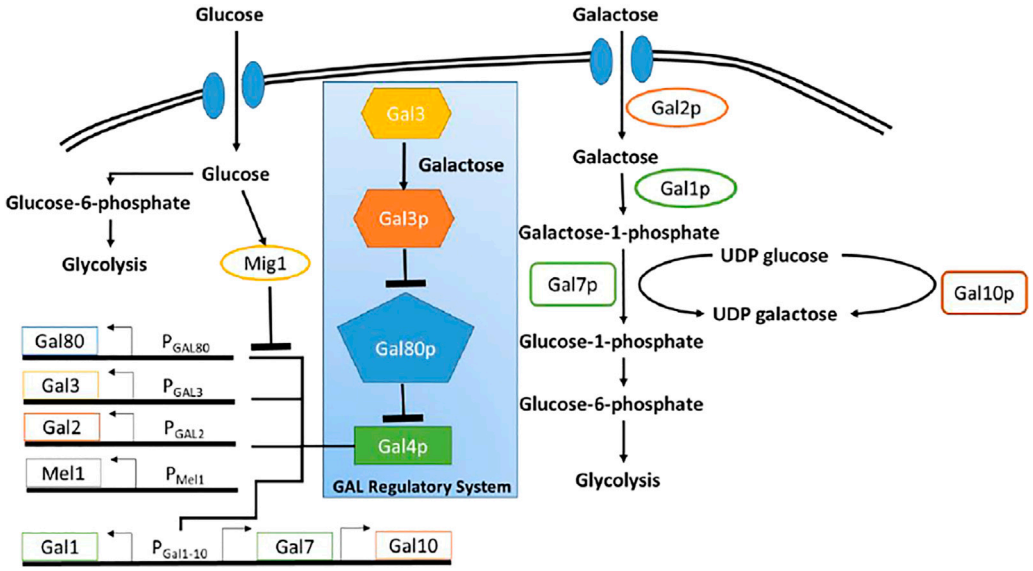


FIGURE 1 | Schematic of galactose and glucose metabolic pathway in Yeast. Blue box highlighted the regulatory network pathway of GAL genes. Glucose as a primary carbon resource, suppress the GAL genes through Mig1 protein. In the presence of galactose and in the absence of glucose, GAL4 protein activates the GAL genes.

GAL regulon. In the presence of galactose, however, the signal transducer Gal3p, binds galactose and in its activated form (Gal3p*) binds Gal80p, forming the complex Gal3p*-Gal80p. The Gal3p*-dependent sequestration of Gal80p, thus frees, Gal4p to activate gene expression from the GAL regulon promoters (Bram et al., 1986; Johnston, 1987). Galactose is brought into the cell via the galactose-specific transporter, Gal2p (Chiang et al., 1996; Hawkins and Smolke, 2006). Upon entry, galactose is first activated upon by the galactokinase, Gal1p (Zenke et al., 1996).

Laboratory work in the 1970s led to isolation of several alleles of GAL regulatory proteins (Douglas and Hawthorne, 1966; Nogi and Fukasawa, 1984; Salmeron et al., 1990). More recently, analysis of yeast isolated from ecological niches has revealed that several alleles of the regulators in the GAL system are present in nature (Wang et al., 2015; Lee et al., 2017). Analysis of several thousands of these isolates was used to reveal that distribution of these alleles is non-random; and certain combinations of alleles confer a greater fitness than others (Boocock et al., 2021).

In this study, we develop a quantitative model to study gene expression dynamics in the GAL system. Several existing models to study this system exist. However, we use the model to ask the following question: how does changing the allelic combinations make to the dynamics of gene expression of the GAL network? Are there specific regulatory elements in the GAL network which control the system's sensitivity, evolvability than others? We particularly focus on the Gal4p-Gal80p and Gal3*-Gal80p interaction. Towards this, we first develop a quantitative model, and study the system's behavior at a population as well as at a single-cell resolution. We then validate our model through analysis of the wild type laboratory strain *S. cerevisiae*, and two distinct regulatory mutants. Thereafter, we analyze our model for studying the regulatory evolvability of the network. Our analysis reveals that in the regulatory structure of the GAL network, the gene expression behavior is highly evolvable via acquisition of mutations among the several regulatory elements in the network, thus, suggesting that the GAL network is highly evolvable, and suggesting an explanation regarding the high allelic diversity among its regulatory elements in yeast isolated from ecological isolates.

MATERIALS AND METHODS

Mathematical Modeling of the GAL Network

In *S. cerevisiae,* enzymes encoded in the GAL regulon control metabolism of galactose. Seven proteins play a role in the metabolism of galactose. Among them, there are three enzymes involved in the catalytic pathway to convert galactose to glucose-6-phosphate and there are four other enzymes involved in transportation of galactose and activation of GAL genes (Sellick et al., 2008). **Figure 1** shows the schematic diagram of the utilization of glucose and galactose to produce biomass through glycolysis process.

Both glucose and galactose have specific transporter through which those sugars are imported into the cell. Glucose is transported by HXT proteins and galactose is imported through Gal2p. Glucose represses the metabolism of other sugar through global carbon catabolic repression. Therefore, once glucose entered into the cell, the galactose gene network shuts off because of the repression by glucose molecules. Once glucose is completely utilized, the repression is relieved and the galactose gene regulatory network resumes the production of GAL genes. The GAL genes include GAL1, GAL2, GAL3, GAL80, GAL7, and GAL10.

Gal1p, Gal7p, Gal10p constitute the enzymatic pathway (Leloir pathway) involved in the process of converting galactose into

glucose-1-phosphate (Sellick et al., 2008). The Gal1p and Gal3p proteins become active in their signal transduction role only when they bind galactose. While Gal3p is exclusively a signal transducer, Gal1p is a galactokinase but also retains some signal transduction activity (Lohr et al., 1995; Platt et al., 2000).

Modeling of GAL Regulatory System

The modeling of GAL regulatory network based on the deterministic approach adopted in (Venturelli et al., 2012). The rate equations for all the GAL proteins are given below for a single cell. We assume that the environment is a chemostat, and that the extracellular concentration of sugars is constant with time. The rate equation for GAL1 protein is given as follow,

$$\frac{d[G1]}{dt} = \alpha_{0G1} + \alpha_{G1}\left(\frac{[G4]^{n1}}{[G4]^{n1} + K_{G1}^{n1}}\right)\left(\frac{K_{R1}^{nG1}}{[R]^{nG1} + K_{R1}^{nG1}}\right) - k_{f1}[G1][B_i] + k_{r1}[G1*] - \gamma_{G1}[G1] \quad (1)$$

The above equation contains the production rate of Gal1p through basal expression and induced expression through transcription factor Gal4p, which has multiple binding sites and the binding mechanism is modeled as Hill function with coefficient, *n1* equal to 2. The glucose repression is also modeled using the Hill equation for repressor with Hill coefficient, *nG1* equal to 1. The third and fourth terms describe the activation of Gal1p protein when binding with galactose. The last term in the equation describes the degradation of Gal1p.

The rate equation of Gal2p is as follows,

$$\frac{d[G2]}{dt} = \alpha_{0G2} + \alpha_{G2}\left(\frac{[G4]^{n2}}{[G4]^{n2} + K_{G2}^{n2}}\right) - \gamma_{G2}[G2] \quad (2)$$

The rate equation of Gal3p consists of production, transformation, and degradation term. The production terms depend on the basal rate and induced expression rate with glucose repression. The active form of Gal3p, Gal3p* (represented as *G3** in the equation) forms when Gal3p binds with galactose. The induction and the repression rates are modeled using the Hill equation with coefficients *n3* equal to 2, and *nG3* equal to 1.

$$\frac{d[G3]}{dt} = \alpha_{0G3} + \alpha_{G3}\left(\frac{[G4]^{n3}}{[G4]^{n3} + K_{G3}^{n3}}\right)\left(\frac{K_{R3}^{nG3}}{[R]^{nG3} + K_{R3}^{nG3}}\right) - k_{f3}[G3][B_i] + k_{r3}[G3*] - \gamma_{G3}[G3] \quad (3)$$

The rate of formation of Gal4p has basal expression and glucose repression term that modeled as Hill equation with Hill coefficient, *nG4* equal to 2.

$$\frac{d[G4]}{dt} = \alpha_{G4}\left(\frac{K_{R1}^{nG4}}{[R]^{nG4} + K_{R4}^{nG4}}\right) - k_{f84}[G4][G80] + k_{r84}[C84] - \gamma_{G4}[G4] \quad (4)$$

The rate equation for GAL80 is given as,

$$\frac{d[G80]}{dt} = \alpha_{0G80} + \alpha_{G80}\left(\frac{[G4]^{n80}}{[G4]^{n80} + K_{G80}^{n80}}\right)\left(\frac{K_{R80}^{nG80}}{[R]^{nG80} + K_{R80}^{nG80}}\right) - k_{f81}[G1*][G80] + k_{r81}[C81] - k_{f83}[G3*][G80] + k_{r83}[C83] - k_{f84}[G4][G80] + k_{r84}[C84] - \gamma_{G80}[G80] \quad (5)$$

The above equation includes the production rate Gal80p with basal and induced expression, interaction with Gal1p, Gal4p, Gal3p, and the degradation of Gal80p. The active form of Gal1p (GAL1p*) has weak affinity towards Gal80p. The following equation gives the dynamics of Gal1p.

$$\frac{d[G1*]}{dt} = k_{f1}[G1][B_i] - k_{r1}[G1*] - k_{f81}[G1*][G80] + k_{r81}[C81] - \gamma_{G1p}[G1*] \quad (6)$$

The active form of Gal3p (Gal3p*) is formed as the result of galactose and Gal3p binding. Gal3p* interacts with Gal80p and form Gal80p-Gal3p* complex that prevents Gal80p Gal4p binding. The binding affinity between Gal3p* and Gal80p is stronger than Gal1p* and GAL80.

$$\frac{d[G3*]}{dt} = k_{f3}[G3][B_i] - k_{r3}[G3*] - k_{f83}[G3*][G80] + k_{r83}[C83] - \gamma_{G3p}[G3*] \quad (7)$$

The following equation describes the formation of intermediate complex molecules formed between Galp1*-Gal80p, Gal3p*-Gal80p, and Gal4p-Gal80p,

$$\begin{aligned}\frac{d[C81]}{dt} &= k_{f81}[G1*][G80] - k_{r81}[C81] - \gamma_{C81}[C81]\\ \frac{d[C83]}{dt} &= k_{f83}[G3*][G80] - k_{r83}[C83] - \gamma_{C83}[C83]\\ \frac{d[C84]}{dt} &= k_{f84}[G4][G80] - k_{r84}[C84] - \gamma_{C84}[C84]\end{aligned} \quad (8)$$

Transport Processes of Glucose and Galactose

The following equations describe the transport processes of glucose and galactose across the cell membrane. The rate change of the internal concentration of galactose molecules depends on the formation of biomass from galactose, the concentration of galactose transporter Gal2p and the formation of active protein Gal1p* and Gal3p*.

$$\frac{dB_i}{dt} = \frac{\lambda_{gal}}{G2_{max}}\left(\frac{[B_e]}{[B_e] + K_{\lambda B}}\right)[G2] - \left(\frac{\mu_{B,max}}{G1_{max}}\right)\left(\frac{[B_i]}{B_i + K_B}\right)[G1][X] - k_{f3}[G3][B_i] + k_{r3}[G3_*] - k_{f1}[G1][B_i] + k_{r1}[G1_*] \quad (9)$$

The rate of intake of galactose molecules also depends on the concentration of Gal2 transporter protein.

$$\begin{aligned}\frac{dB_e}{dt} &= -\frac{\lambda_{gal}}{G2_{max}}\left(\frac{[B_e]}{[B_e] + K_{\lambda B}}\right)[G2][X]\\ \frac{dA_e}{dt} &= -\lambda_{glu}\left(\frac{[A_e]}{[A_e] + K_{\lambda A}}\right)[X]\end{aligned} \quad (10)$$

TABLE 1 | Parameters value used in both deterministic and stochastic model to simulate GAL network. Parameter values taken from (Venturelli et al., 2012), unless otherwise noted with a * in column 1.

Parameters	Value	Unit
α_{10} - GAL1 basal expression rate*	0.001	$nM.min^{-1}$
α_{80} - GAL80 basal expression rate	0.6	$nM.min^{-1}$
α_{30} - GAL3 basal expression rate*	0.001	$nM.min^{-1}$
α_4 - GAL4 basal expression rate	0.5	$nM.min^{-1}$
α_1 - GAL1 induced expression rate	15	$nM.min^{-1}$
K_{G1} - GAL1 transcriptional feedback threshold	50	*nM*
n_1 - GAL1 Hill coefficient	2	*no unit*
α_8 - GAL80 induced expression rate	0.9	$nM.min^{-1}$
K_{G80} - GAL80 transcriptional feedback threshold*	20	*nM*
n_{80} - GAL80 Hill coefficient	2	*no unit*
α_3 - GAL3 induced expression rate	2	$nM.min^{-1}$
K_{G3} - GAL3 transcriptional feedback threshold	50	*nM*
n_3 - GAL3 Hill coefficient	2	*no unit*
k_{f1} - GAL1-galactose forward binding rate*	0.001	$(nM.\mathrm{min})^{-1}$
k_{f3} - GAL3-galactose forward binding rate*	1	$(nM.\mathrm{min})^{-1}$
k_{r1} - GAL1-galactose unbinding rate*	3	min^{-1}
k_{r3} - GAL3-galactose unbinding rate*	10	min^{-1}
k_{f81} - GAL1-GAL80 forward binding rate*	1	$(nM.\mathrm{min})^{-1}$
k_{f83} - GAL3-GAL80 forward binding rate*	1	$(nM.\mathrm{min})^{-1}$
k_{f84} - GAL4-GAL80 forward binding rate*	0.8	$(nM.\mathrm{min})^{-1}$
k_{r81} - GAL1-GAL80 unbinding rate*	0.1	min^{-1}
k_{r83} - GAL3-GAL80 unbinding rate*	0.1	min^{-1}
k_{r84} - GAL4-GAL80 unbinding rate*	1	min^{-1}
γ_{G1} - GAL1 decay rate	0.004	min^{-1}
γ_{G80} - GAL80 decay rate	0.004	min^{-1}
γ_{G3} - GAL3 decay rate	0.004	min^{-1}
γ_{G4} - GAL4 decay rate	0.004	min^{-1}
γ_{G1p} - GAL1p decay rate	0.004	min^{-1}
γ_{G3p} - GAL3p decay rate	0.004	min^{-1}
γ_{C81} - Complex C81 decay rate	0.004	min^{-1}
γ_{C83} - Complex C83 decay rate	0.004	min^{-1}
γ_{C84} - Complex C84 decay rate	0.004	min^{-1}

Here, A and B represent the glucose and galactose concentration respectively. The subscript *e* refers to extracellular concentration.

Repression of Gal Network Through Mig1 Protein

The following equations are the rate equations for the Mig1p concentration (R). Mig1 protein becomes activated in presence of glucose.

$$\begin{aligned}\frac{d[R]}{dt} &= \alpha_R - k_{fR}[R][A_i] + k_{rR}[Rp] - \gamma_R[R] \\ \frac{d[Rp]}{dt} &= k_{fR}[R][A_i] - k_{rR}[R_p] - \gamma_{Rp}[Rp]\end{aligned} \tag{11}$$

Modelling of Mutant Strain

The rate equation of Gal3p for the *gal3*Δ mutant is modified as follow,

$$\begin{aligned}\frac{d[G3]}{dt} &= 0 \\ \frac{d[G3p]}{dt} &= 0\end{aligned} \tag{12}$$

The epistatically altered strain was modelled by changing the appropriate parameter values in the model. The Gal3*-Gal80p interaction is weakened by a factor of 4, and the Gal80p-Gal4p interaction is weakened by a factor of 5, compared to the magnitude of these interactions in the ancestral strain.

The above equations are simulated simultaneously using Matlab ODE functions. **Table 1** lists the values of the parameters used in the simulations.

Stochastic Modelling of GAL Network

We use the Gillespie algorithm to implement the stochastic model for the GAL network (Gillespie, 1976). The dynamic reactions with their corresponding stochastic rates are listed in the **Table 2**. Galactose units are represented as A.U. in the model. These numbers closely match with the system's response to galactose, when A.U. is replaced by nM. However, we note that different *S. cerevisiae* strains exhibit quantitatively different response to exposure to galactose (Lee et al., 2017). Hence, in this work, we focus on capturing the qualitative response of the system, to exposure to galactose.

Experimental Methods

The ancestor, epistatically-altered strain, and the *gal3*Δ strains used in this study are as described previously (Johnston and Hopper, 1982; Das Adhikari et al., 2014). To measure GAL1 promoter activity in these three strains GAL1-lacZ fusions were integrated at the *ura3* locus, as described previously (Das Adhikari et al., 2014).

Growth Kinetics

Glycerol-lactate pre-grown strains were plated onto SCM agar plates (containing 2% glucose). The plates were thereafter incubated at 30°C for 2–3 days. Colonies were randomly selected from the agar plates and subjected to two rounds of serial passage in appropriate media [1% glucose, or 1% galactose, or glycerol/lactate (gly/lac)]. The resulting cultures were then washed with SCM and then growth curves were initiated with an initial optical density of 0.1 in SCM containing the appropriate carbon source (glucose, galactose, or gly/lac at the concentrations mentioned above). Three replicates of culture were transferred to a 96-well plate and OD was measured periodically until the cultures reach stationary phase. The plates were overlaid with a *Breathe Easy* membranes (Sigma) to prevent evaporation.

GAL1 Expression Levels

GAL1-lacZ levels were determined in the three strains, ancestor, the epistatically altered strain, and the *gal3*Δ as described previously (Das Adhikari et al., 2014).

Cell-Cell Heterogeneity (2-Deoxygalactose Experiments)

2DG-induced toxicity has been used previously to observe the metabolic state of a population (Platt, 1984; Das Adhikari et al., 2014). Cells were inoculated into gly/lac medium and incubated at 30 deg C for 48–72 h with shaking. These gly/lac pregrown cultures were inoculated into 5 ml of CSM containing 1%

TABLE 2 | Reaction events and reaction rate used to simulate stochastic model.

Reaction ID	Reaction	Rate
R1: GAL1 basal expression	$\varphi \xrightarrow{\alpha_{10}} Gal1$	α_{10}
R2: GAL80 basal expression	$\varphi \xrightarrow{\alpha_{80}} Gal80$	α_{80}
R3: GAL3 basal expression	$\varphi \xrightarrow{\alpha_{30}} Gal3$	α_{30}
R4: GAL4 basal expression	$\varphi \xrightarrow{\alpha_4} Gal4$	α_4
R5: GAL1 expression by GAL4 induction	$\varphi \xrightarrow{\alpha_1} Gal1$	$\alpha_1 (\frac{G4^{n1}}{G4^{n1}+K_{G1}^{n1}})$
R6: GAL80 expression by GAL4 induction	$\varphi \xrightarrow{\alpha_8} Gal80$	$\alpha_8 (\frac{G4^{n80}}{G4^{n80}+K_{G80}^{n80}})$
R7: GAL3 expression by GAL4 induction	$\varphi \xrightarrow{\alpha_3} Gal3$	$\alpha_3 (\frac{G4^{n3}}{G4^{n3}+K_{G3}^{n3}})$
R8: GAI1 binding with galactose	$Gal1 + gal \xrightarrow{k_{f1}} Gal1_p$	$k_{f1}(G1)(gal)$
R9: GAL3 binding with galactose	$Gal3 + gal \xrightarrow{k_{f3}} Gal3_p$	$k_{f3}(G3)(gal)$
R10: Unbinding of GAL1-galactose	$Gal1_p \xrightarrow{k_{r1}} Gal1 + gal$	$k_{r1}(G1_p)$
R11: Unbinding of GAL3-galactose	$Gal3_p \xrightarrow{k_{r3}} Gal3 + gal$	$k_{r3}(G3_p)$
R12: Binding of GAL1p and GAL80	$Gal1_p + Gal80 \xrightarrow{k_{f81}} Gal81$	$k_{f81}(G1_p)(G80)$
R13: Binding of GAL3p and GAL80	$Gal3_p + Gal80 \xrightarrow{k_{f83}} Gal83$	$k_{f83}(G3_p)(G80)$
R14: Unbinding of GAL4-GAL80	$Gal4 + Gal80 \xrightarrow{k_{f84}} Gal84$	$k_{f84}(G4)(G80)$
R15: Unbinding of GAL1p-GAL80	$Gal81 \xrightarrow{k_{r81}} Gal1_p + Gal80$	$k_{r81}(C81)$
R16: Unbinding of GAL3p-GAL80	$Gal83 \xrightarrow{k_{r83}} Gal3_p + Gal80$	$k_{r83}(C83)$
R17: Unbinding of GAL4-GAL80	$Gal84 \xrightarrow{k_{r84}} Gal4 + Gal80$	$k_{r84}(C84)$
R18: decaying of GAL1	$Gal1 \xrightarrow{\gamma_{G1}} \theta$	$\gamma_{G1}(G1)$
R19: decaying of GAL80	$Gal80 \xrightarrow{\gamma_{G80}} \theta$	$\gamma_{G80}(G80)$
R20: decaying of GAL3	$Gal3 \xrightarrow{\gamma_{G3}} \theta$	$\gamma_{G3}(G3)$
R21: decaying of GAL4	$Gal4 \xrightarrow{\gamma_{G4}} \theta$	$\gamma_{G4}(G4)$
R22: decaying of GAL1p	$Gal1_p \xrightarrow{\gamma_{G1p}} \theta$	$\gamma_{G1_p}(G1_p)$
R23: decaying of GAL3p	$Gal3_p \xrightarrow{\gamma_{G3p}} \theta$	$\gamma_{G3_p}(G3_p)$
R24: decaying of Complex C81	$Gal81 \xrightarrow{\gamma_{C81}} \theta$	$\gamma_{C81}(C81)$
R25: decaying of Complex C83	$Gal83 \xrightarrow{\gamma_{C83}} \theta$	$\gamma_{C83}(C83)$
R26: decaying of Complex C84	$Gal84 \xrightarrow{\gamma_{C84}} \theta$	$\gamma_{C84}(C84)$

galactose, such that the starting OD of the culture was equal to 0.01. Cells were harvested from this growing culture, serially diluted with PBS and thereafter, plated onto gly/lac solid media containing 2-Dexoygalactose (2DG) (0.3 μM). As a control, cells were also plated on media with gly/lac and no 2DG. The two sets of plates were incubated at 30 deg C for 3–4 days. The number of colonies that grew on gly/lac plates and those containing 2DG were counted and the percentage of Gal-positive cells calculated. All experiments were repeated three times. A minimum of 500 colonies on gly/lac plates were counted in each experiment.

RESULTS

Experimental Characterization of the GAL Regulon/Network in Wild-Type and Mutant *S. cerevisiae*

We first start by experimentally characterizing the growth kinetics of cells in a galactose medium. In this work, we focus on three different strains. First, the ancestral wild type; second, an epistatically-altered strain in which the Gal3*-Gal80p interaction is weakened by a factor of four, and the Gal80p-Gal4p interaction is weakened by a factor of five (Das Adhikari et al., 2014). These changes in the interaction strengths are due to the strain having mutant alleles of Gal80p and Gal4p.

We first study the growth kinetics of the wild type strain, as the cells are brought to a galactose environment from the same/different carbon source environment. As shown in the Figure, the populations exhibits a lag phase duration which is a function of the environment from which the cells are brought into the galactose media. At a single-cell resolution, the transition from a GAL OFF to a GAL ON state is heterogeneous; and the wild type strain exhibits a characteristic gene expression levels, when exposed to a certain galactose concentration. (**Figure 2**; **Supplementary Figure S1**). As shown in the Figure, as cells are brought to a media containing galactose, the initial lag phase duration is a function of the environment the cells are brought from. The lag phase is the longest when cells are brought from repressing conditions (glucose), medium when brought from non-inducing non-repressing (NINR) conditions (glycerol/lactate), and shortest when brought from inducing conditions (galactose). In the time region when the cells transition from lag to log phase, the single-cell behavior of the gene expression levels was characterized. For this purpose, we withdrew cells at regular intervals, and plated an equal volume on gly/lac plates, and on

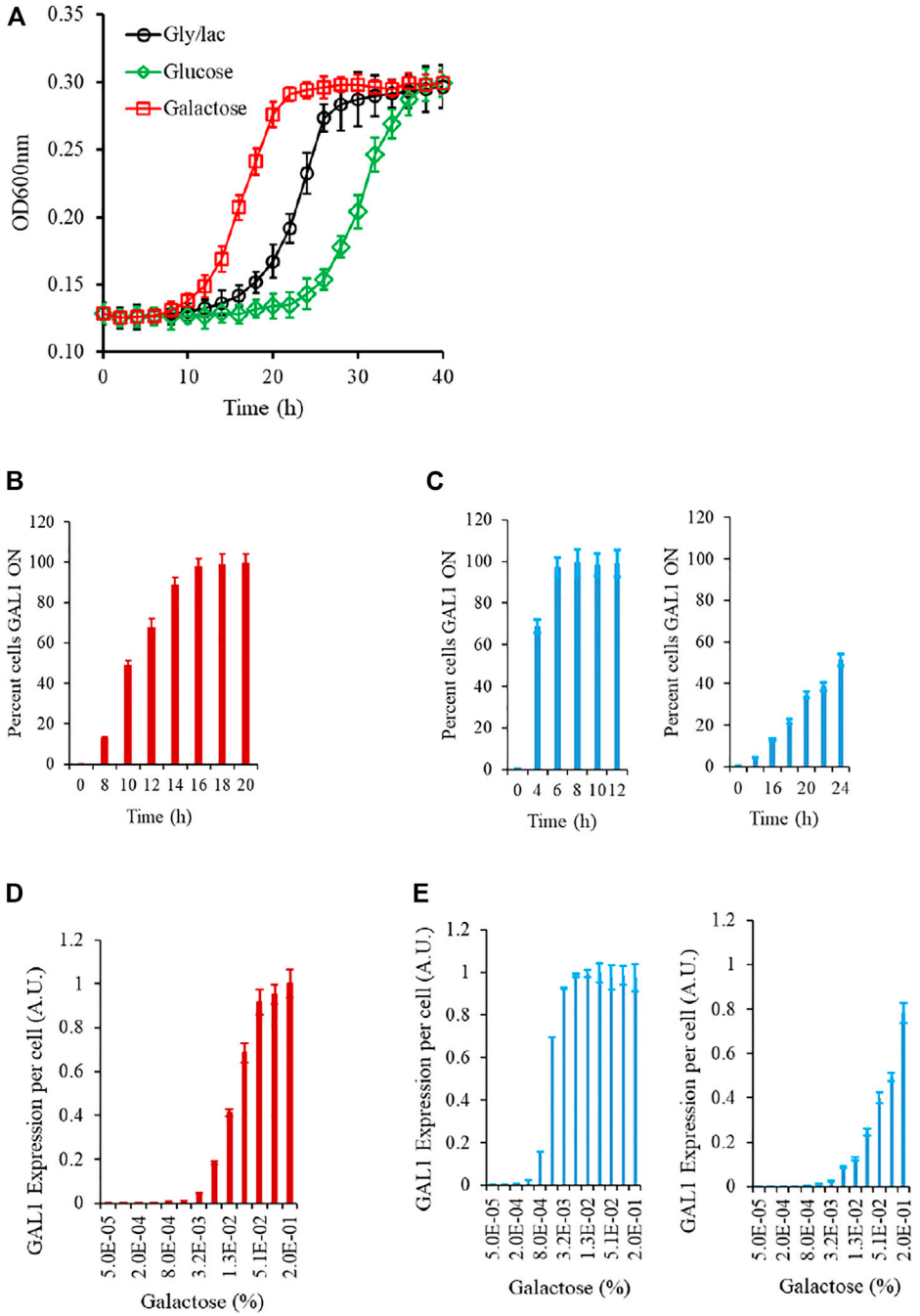


FIGURE 2 | Experimental characteristics of the GAL network in *S. cerevisiae.* **(A)** Growth kinetics in galactose is contingent on the conditions in which the cells are introduced from. The lag is the longest in the cells introduced from glucose, and shortest in those introduced from galactose. **(B)** During transition from gly/lac to galactose, the percent cells which are expressing GAL1 is heterogeneous in the initial phase of the growth. **(C)** Same as **(B)**. The transition from lag to log phase in the epistatically-altered strain (left) and the *gal3*Δ strain (right) is qualitatively different. **(D)** Steady state GAL1 expression in ancestral strain, when cells are grown in different galactose concentrations. **(E)** Same as **(C)** but for the epistatically-altered strain (left) and the *gal3*Δ strain (right). The nature of galactose gene expression in the three strains is qualitatively different from each other. All experiments were performed in triplicate. The average and standard deviations are reported.

gly/lac plates containing 2DG. All cells expressing Gal1p only grow on the gly/lac plates. This assay thus allows us to quantify, at a single-cell resolution, how cells transition from a GAL OFF state to a GAL ON state. As shown in the Figure, the wild-type strain exhibits a heterogeneous induction kinetics in the window of transition from lag to log phase. On the other hand, the epistatically-altered mutant exhibits a kinetics where the induction is homogeneous and faster, as compared to the wild

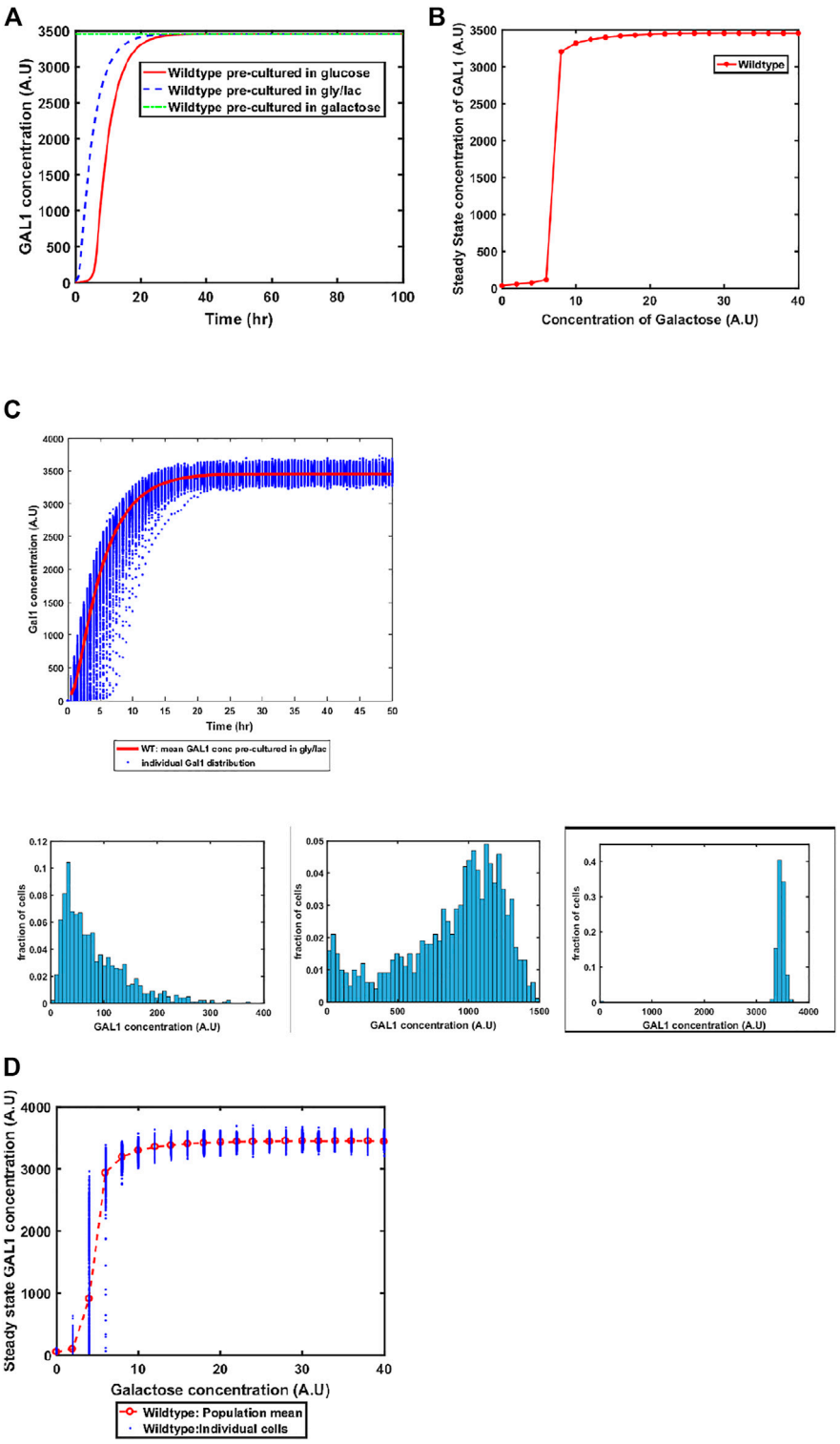


FIGURE 3 | (A) Activation of GAL network in galactose [concentration, 30 (A.U)]. GAL1 profile of wildtype pre-cultured in gly/lac (dashed blue line), glucose (solid red line), and galactose (dashed green line). **(B)** Steady state concentration of GAL1 protein at different concentrations of galactose in wild type *S. cerevisiae*. Note that the system exhibits a threshold-like activation. **(C)** Dynamic GAL1 response at galactose concentration of 30 (A.U) for 1000 wild-type cells pre-grown in gly/lac. Histograms of GAL1 expression at 30 min, 5 h, and 50 h. **(D)** Steady state concentration of GAL1 under different concentration of galactose for wild type.

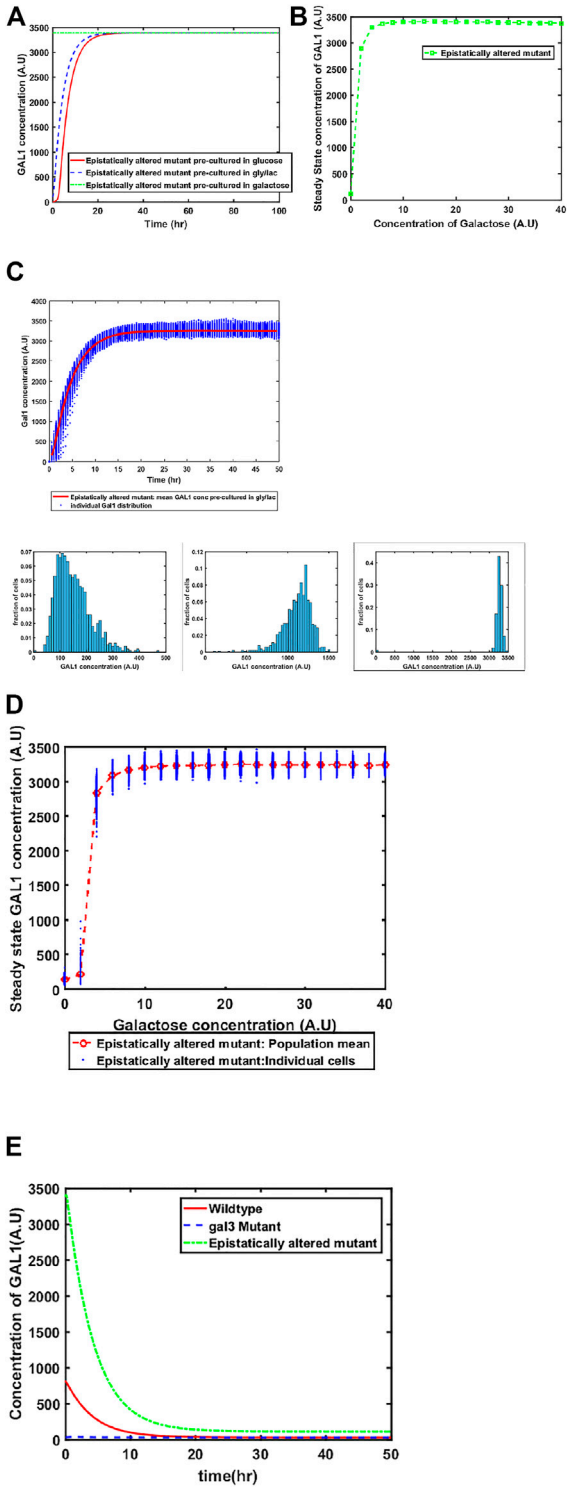


FIGURE 4 | Dynamics of gene expression for the epistatically-altered strain. **(A)** Activation of GAL network in galactose [concentration, 30 (A.U)]. GAL1 profile when cells are pre-cultured in gly/lac (dashed blue line), glucose (solid red line), and galactose (dashed green line). **(B)** Steady state concentration of GAL1 protein at different concentrations of galactose. Note that the system exhibits a behaviour which is qualitatively different than the wild type behavior. The threshold of galactose concentration is no longer

(*Continued*)

type. On the other hand, in the *gal3*Δ, the induction kinetics are considerably slower, and heterogeneous, even at long durations of time. Note that in the *gal3*Δ, the signal transduction activity is carried out by the kinase, Gal1p, which is a bifunctional protein is capable of weak signal transduction activity also (Meyer et al., 1991).

The steady state Gal1p levels, when studied via a proxy of a promoter fusion reporter indicate that GAL genes are turned ON, when the concentration of galactose in the media exceeds a threshold. These thresholds and the nature of induction are qualitatively different in the three strains being analyzed in this study. Thus, we establish that the nature of response to a galactose environment in *S. cerevisiae* is strongly influenced by the protein-protein interactions in the regulatory network.

GAL Regulon Model Captures the Kinetics of Gene Expression

Kinetics of GAL System Induction are Strongly Dependent on the Initial Conditions

The GAL system in *S. cerevisiae* is induced in the presence of galactose, and is actively repressed in the presence of glucose. On the other hand, the system is in a Non-Induced Non-Repressed conditions (NINR), when cells are grown in gly/lac media. This aspect of the system is captured in the model (**Figure 3A**). When cells are transferred to a media containing galactose, the duration of the lag phase is contingent on the pre-existing state of the system. The lag phase is the longest in the cells which are introduced into galactose-containing media from glucose, and is shortest in the cells which come from galactose.

Switch-Like Induction of GAL Genes

Next, we simulate the steady state concentration of GAL1 levels when cells are grown in different galactose concentrations. As shown in the **Figure 3B**, GAL gene induction follows a step-like behavior. For small concentrations of galactose in the media, the increase in GAL gene induction is small and linear in nature. Beyond a threshold, however, the GAL1 gene expression levels increases in a step function manner to their maximal levels. Thereafter, the steady state concentrations of GAL1 do not change with further increase in galactose.

Positive Feedback, and Cell-Cell Heterogeneity in the GAL System

Sugar utilization systems are a combination of positive and negative feedback loops. The GAL4p-encoded positive feedback leads to rapid induction of the system. As cells transition from the GAL OFF to a GAL ON state, there is

FIGURE 4 | present. **(C)** Dynamic GAL1 response at galactose concentration of 30 (A.U) for 1000 cells pre-grown in gly/lac. Histograms of GAL1 expression at 30 min, 5 h, and 50 h. **(D)** Steady state concentration of GAL1 under different concentration of galactose. **(E)** The dynamics of transition from ON to OFF state in the ancestor and the epistatically altered strain. Note that even after long times, the epistatically altered strain has a non-zero residual GAL1 expression. The GAL1 profile for the *gal3*Δ is shown as control.

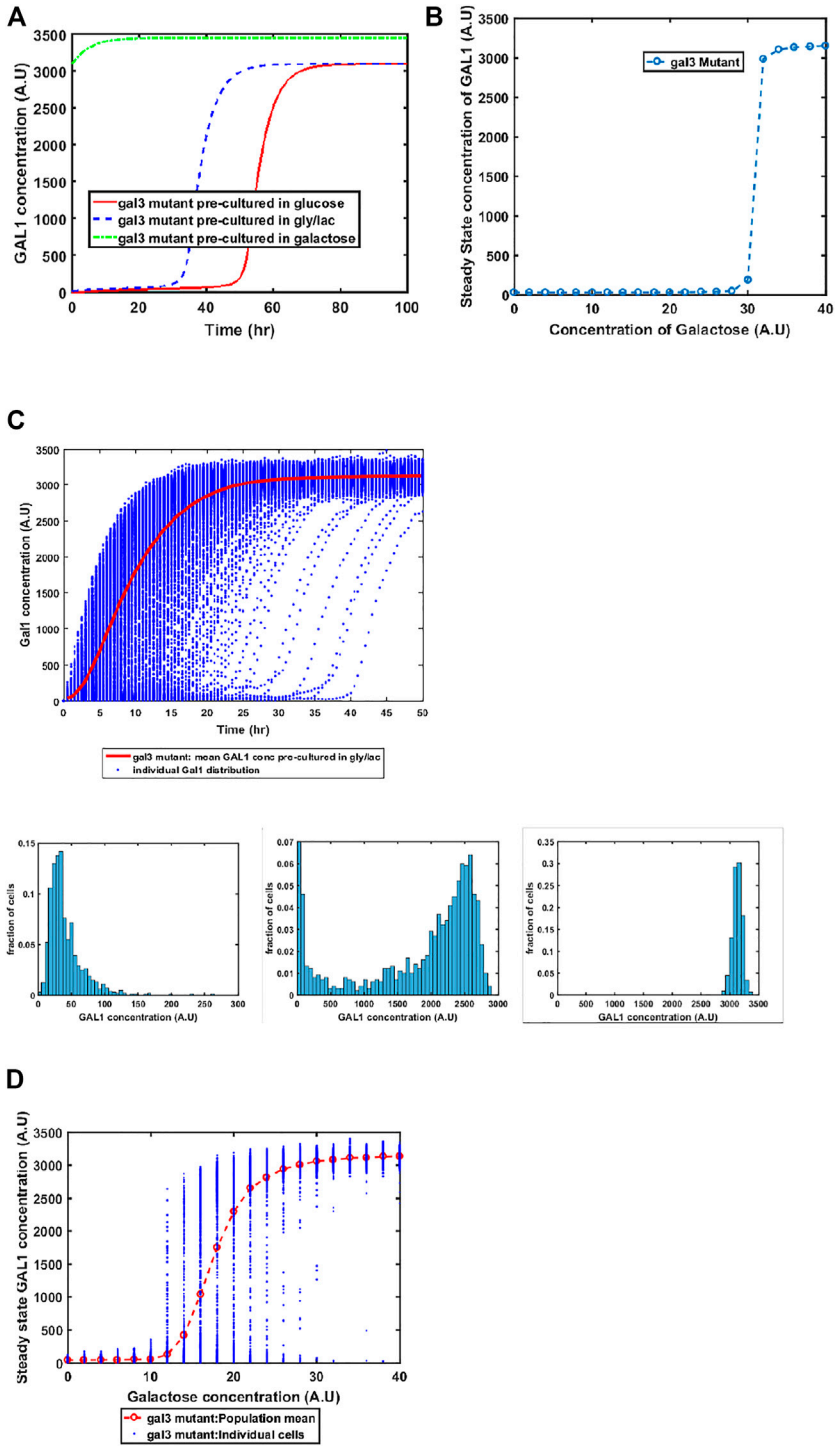


FIGURE 5 | Dynamics of gene expression for *gal3*Δ strain. **(A)** Activation of GAL network in galactose [concentration, 30 (A.U)]. GAL1 profile when cells are pre-cultured in gly/lac (dashed blue line), glucose (solid red line), and galactose (dashed green line). **(B)** Steady state concentration of GAL1 protein at different concentrations of galactose. Note that the system exhibits a behaviour which is qualitatively different than the wild type behavior. The threshold of galactose concentration is no longer present. **(C)** Dynamic GAL1 response at galactose concentration of 30 (A.U) for 1000 cells pre-grown in gly/lac. Histograms of GAL1 expression at 30 min, 5 h, and 50 h. **(D)** Steady state concentration of GAL1 under different concentration of galactose.

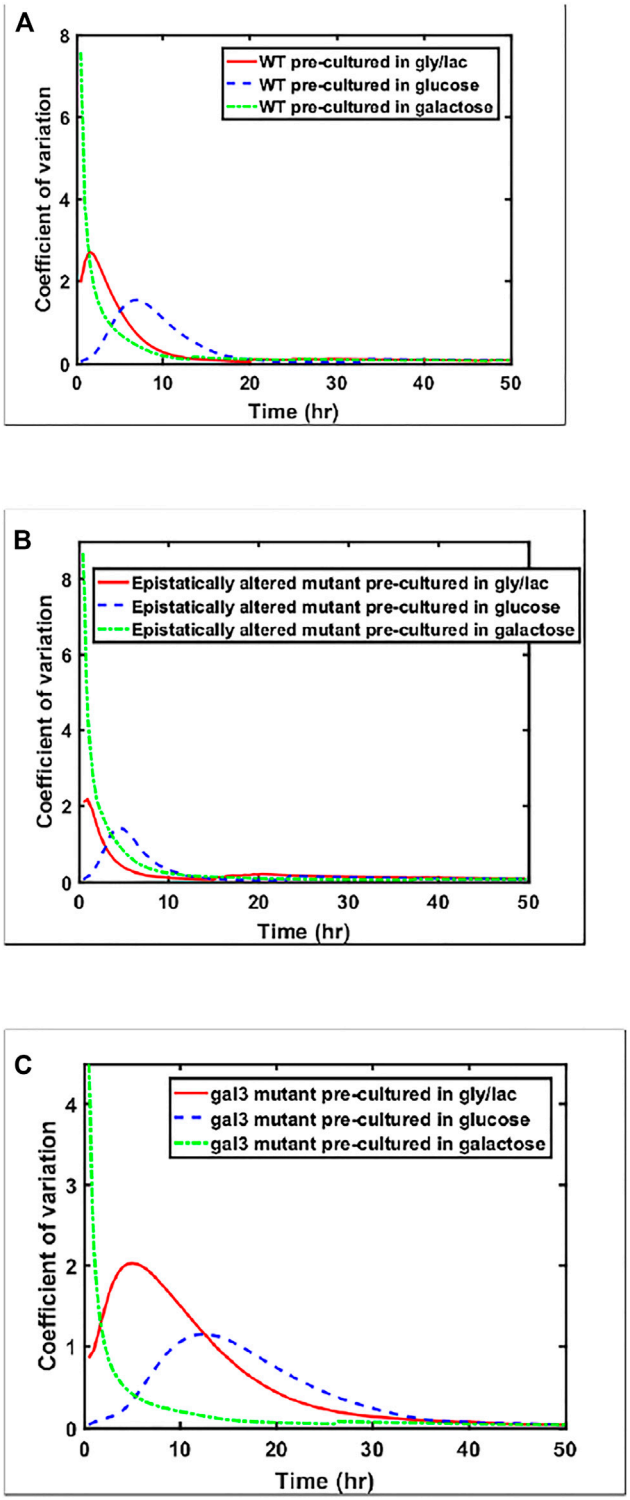


FIGURE 6 | Coefficient of variation in GAL1 expression levels, shown for **(A)** wild type, **(B)** Epistatically-altered strain, and **(C)** *gal3*Δ strain when grown in different environmental conditions.

cell-cell heterogeneity in the system (**Figure 3C**). At intermediate times, a fraction of the cells are in the ON state, and the rest of the population is in the OFF state. This heterogeneous induction kinetics are characteristic of sugar utilization systems in bacteria. At a steady state level, the GAL network exhibits cell-cell heterogeneity at steady state in low concentrations of galactose (**Figure 3D**).

Altering the GAL Network Architecture

Two aspects of the GAL network stand out. The first, GAL system in *S. cerevisiae* comprises of GAL1 and GAL3, two genes which resulted from a whole genome duplication event (Marcet-Houben and Gabaldon, 2015; Wolfe, 2015). Gal3p is a signal transducer, which dictates the kinetics of the system; and Gal1p is a galactokinase, which also has limited signal transduction activity. Therefore, a *gal1*Δ strain cannot grow on galactose, whereas a *gal3*Δ strain can exhibit growth on galactose. Second, ecological isolates of *S. cerevisiae* exhibit a wide range of growth kinetics, when grown on a mixture of glucose and galactose. The glucose-dependent repression of the GAL system is strain specific, and as a result, the transition from glucose to galactose varies from strain to strain. Interestingly, this behavior was found to be largely dictated by the allelic variation at the GAL3 locus (Lee et al., 2017). In a recent study, evolution on melibiose was studied, and interestingly, alternate GAL3 alleles were reported discovered (Anjali et al., 2021). Additionally, alternate GAL80 alleles were isolated which exhibit different kinetics of induction, when introduced to a galactose environment (Nogi and Fukasawa, 1984; Salmeron et al., 1990; Douglas and Hawthorne, 1966). This change in kinetics is due to the altered GAL80 allele exhibiting altered binding behavior with GAL4 and GAL3* (GAL3* is GAL3 bound to galactose). Alternate alleles have been isolated in short term evolution experiments. We next use the model to study these two aspects of the network. We mimic the *gal3*Δ mutant by putting GAL3 amounts in the cell equal to zero at all times. We mimic alternate GAL3 alleles by changing the strength of GAL3 interaction with GAL80.

Epistatically-Altered Strain of the GAL Network

Mutant of GAL4 which exhibited constitutive GAL network induction was isolated a long time back. This mutant allele ($GAL4^{c}$) does not interact with GAL80, and thus, cells containing this altered allele have the GAL network in the ON state, independent of the environmental conditions. To restore galactose-dependent induction of the GAL network, Mutants of GAL80 were isolated (e.g., $GAL80^{s-1}$). The $GAL80^{s-1}$ allele interacts with $GAL4^{c}$, and thus restores galactose-dependent induction. However, the mutant alleles ($GAL4^{c}$ and $GAL80^{s-1}$) exhibit altered binding behavior between Gal3*-Gal80p, Gal1*-Gal80p, and Gal80p-Gal4p. The biochemical interactions between $Gal4^{c}$-$Gal80^{s-1}$ and that between $Gal80^{s-1}$-Gal3* have been previously characterized (Das Adhikari et al., 2014). When we make these changes in the model to reflect altered binding kinetics, we predict that the behavior of the system at a 1) population level, and 2) single-cell resolution changes from that of the wild type. As shown in the Figure, our model predicts that the epistatically-altered strain exhibits 1) faster induction of the GAL system, 2) higher steady state levels of the GAL proteins, 3) a qualitatively different induction kinetics at

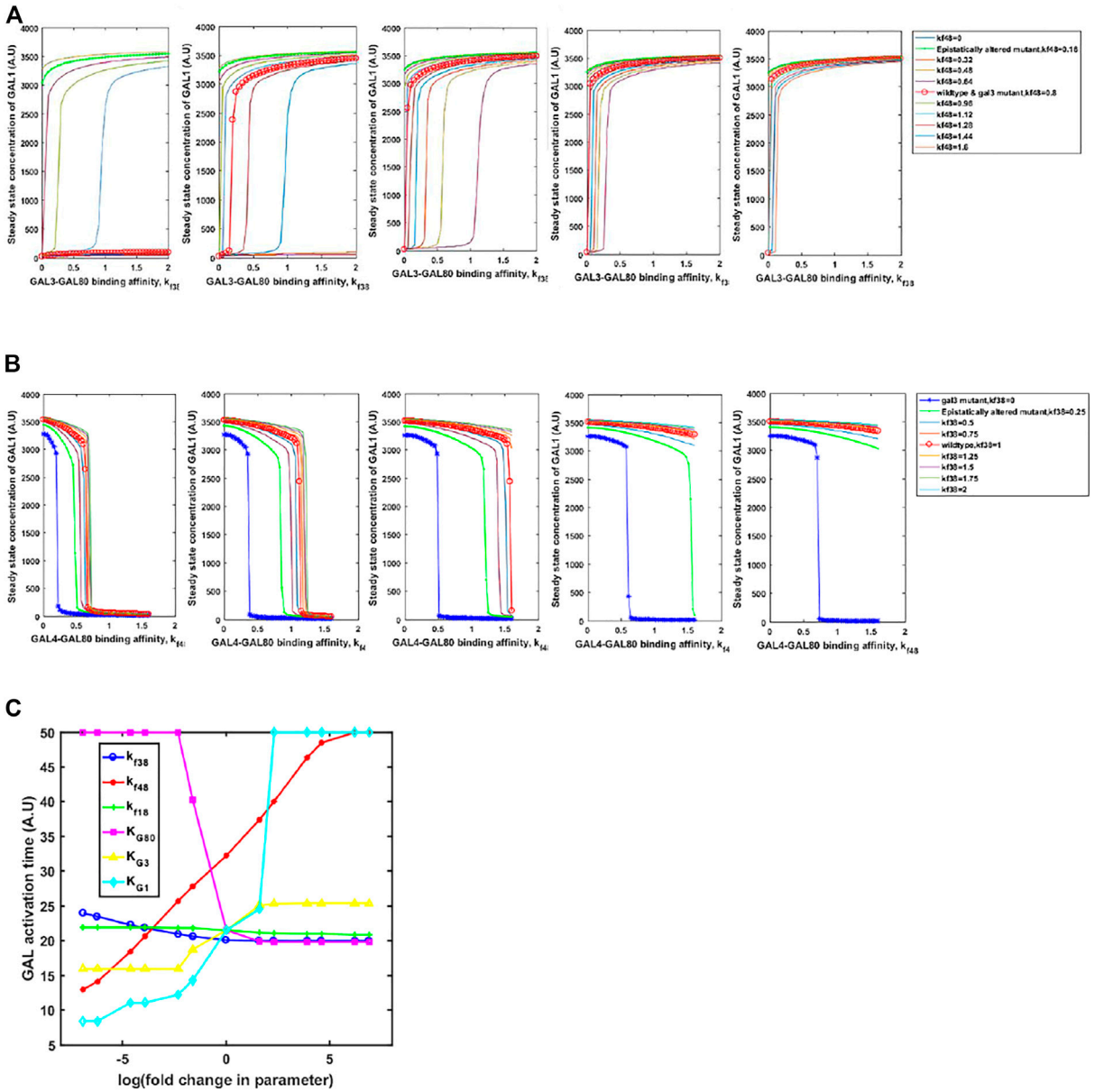


FIGURE 7 | Steady state GAL1 concentration plotted against the binding affinity between GAL3-GAL80 for different GAL4-GAL80. In **(A)** the change in GAL3-GAL80 interaction is represented on the *x*-axis. Different curves represent different strengths of GAL4-GAL80 interaction, as indicated in the legend. In **(B)**, the change in GAL4-GAL80 interaction is represented on the *x*-axis. Different curves represent different strengths of the GAL3-GAL80 interaction, as indicated in the legend. In both **(A)** and **(B)**, the five panels represent a galactose concentration of 5, 10, 15, 20, and 25 A.U. (from left to right). **(C)** Time taken to reach maximal GAL1p production rate, as cells transition from glucose to galactose.

a single-cell resolution, and 4) a slower transition from the ON to the OFF state, when cells are moved from a galactose- to a glucose-containing media (**Figure 4**).

GAL3 Mutant Exhibits Altered Gene Expression Kinetics, Including Bistability at Intermediate Galactose Concentrations

GAL1 and GAL3 are homologous genes, resulting from a whole genome duplication event about a 100 million years ago (Marcet-Houben and Gabaldon, 2015; Wolfe, 2015). The ancestral gene sequence presumably had both kinase and the signal transduction activity. Since duplication and divergence, Gal3p has lost the kinase activity in *S. cerevisiae* and has evolved to become a better signal transducer than the ancestor. On the other hand, Gal1p has evolved to become a specialist galactokinase, while still retaining small signal transduction activity. This adaptive divergence is thought to have resolved an adaptive conflict where one gene sequence was coding for two activities (Hittinger and Carroll, 2007). Since Gal3p does not have any kinase activity, a *gal1*Δ does not exhibit any growth in galactose. On the other hand, a GAL3Δ

exhibits growth, albeit with delayed kinetics, when grown in the presence of galactose. Our model successfully captures this facet of growth on galactose (**Figure 5**). Loss of GAL3 manifests in not only lower Gal1p activity, but also the switch-like induction of the GAL regulon takes place at a significantly higher concentration. At a single-cell resolution, a *gal3*Δ strain exhibits exaggerated heterogeneity as the population transitions from a GAL OFF to a GAL ON state. In addition, steady state levels of Gal1p also exhibit bistability at intermediate concentrations of galactose. Note that this bistability was not observed in the ancestral strain (**Figure 2**), or in the epistatically-altered strain (**Figure 3**). Thus, a *gal3*Δ strain exhibits significantly altered growth kinetics, both at a population level, and at a single-cell resolution.

While the steady state response of the three strains is within a factor of 1.2 for the three strains examined above (ancestor, epistatically-altered strain, and *gal3*Δ), the kinetics of gene expression vary widely. This difference can also be seen from analyzing the coefficient of variation between the expression levels of the members of the populations as cells are transitioned from a given culture conditions to galactose (**Figure 6**). Cell-cell heterogeneity among isogenic populations is now known to manifest itself in a number of ways—often providing adaptive value to the population. Hence, from the prospect of a population to move towards higher fitness levels, the cell-cell variability in a population is an evolutionarily important marker.

Evolution *via* Changes in the Coding Regions of the Regulatory Genes of the GAL Network

Adaptive changes are largely studied as a result of changes in gene expression, which are obtained via changes in one or more of the following: promoter activity, mRNA stability, translation rate. In all these mutations, adaptive benefit takes place as a result of changes in expression levels of the protein. However, as mentioned above, a large number of alleles of GAL regulatory genes, with altered biochemical interactions, have been identified in both, laboratory studies as well as analysis of ecological isolates (Peng et al., 2015; Roop et al., 2016; Lee et al., 2017). These alleles confer qualitatively different growth dynamics, depending on the environmental context. Thus, we hypothesize that the population level and the single-cell behavior of the GAL network can be tuned to a large extent *via* changes in the protein coding sequence which change the protein function rather than levels. This is indicated in **Figures 7A, B**. On changing the interaction strengths between the Gal3*-Gal80p and Gal4p-Gal80p binding propensities, the steady-state induction levels of Gal1p levels are highly tunable. This tuning of gene expression is achieved by simply changing one/two binding coefficients between the regulatory proteins. On the other hand, if the same change made in the *gal3*Δ, the corresponding change in Gal1p protein levels are not observed. Note that a *gal3*Δ strain is able to exhibit growth in galactose, because its regulatory function is compensated by the signal transduction activity of Gal1p.

On the other hand, kinetics of growth on glucose and galactose are also strongly influenced by the parameters. We examine the change in the diauxy lag when cells growing on glucose are introduced to a media containing galactose. As shown in **Figure 7C**, the duration of lag is highly tuned by changing the affinity of protein-protein interactions in the regulatory network, or by changing the half-maximum concentration of the proteins. Interestingly, our simulations show that changing the GAL3p-GAL80p (GAL1p-GAL80p) interaction strength has the least effect on the duration of the lag phase. This is in contrast to changing the interaction strength of GAL4p-GAL80p, which controls the lag phase duration much more strongly. The duration of the lag phase as yeast transition from glucose to galactose has been attributed to several genetic loci (Peng et al., 2015; Roop et al., 2016; Lee et al., 2017), thus indicating a distributed control strategy. Among these interactions, our results demonstrate that GAL4p-GAL80p interaction is the key major determinant of the lag-phase duration, as cells transition from glucose to galactose; and the GAL3p-GAL80p is a fine-tuning interaction of the lag phase duration during this transition.

DISCUSSION AND CONCLUSION

The GAL system in the yeast *S. cerevisiae* is, along with the lac system in *E. coli*, perhaps the most well studied and well characterized gene regulation and sugar utilization systems. Interestingly, while the lac system has been studied in thousands of reports, our understanding of allelic variation resulting in changes in kinetics or levels of gene expression are limited. On the other hand, several alleles of regulators are known. The reasons for this difference are not clear.

S. cerevisiae exists primarily as a diploid in its ecological niches. In fact, several laboratory experiments have conclusively demonstrated that propagation of a haploid for a few hundred generations leads to self-diploidization (Gerstein et al., 2006; Venkataram et al., 2016; Harari et al., 2018). This process is accelerated in conditions of stress (Gerstein et al., 2006). Presumably, doubling the genome size allows for a faster adaptation to the prevailing stressful conditions. In the context of the GAL network, *S. cerevisiae* has several mutational targets to change the kinetics of gene expression. These include the signal transducer GAL3, the repressor GAL80, the transcriptional activator GAL4, and the galactokinase (which has small regulatory activity) GAL1. This diversity in the mutational targets, presumably allows the organism to tune gene expression and physiology which is most appropriate for the surrounding environmental conditions. This is also suggested by a recent analysis of the GAL network, where specific combinations of regulatory alleles were found to confer high fitness; and others found to lead to unfit individuals (Boocock et al., 2021). In fact, analysis of the GAL network from several experimental studies has indicated that the tuning of the GAL gene expression is primarily controlled by modifications in the regulatory proteins, and not by changes in the promoter regions driving expression of the GAL enzymes or regulatory proteins (Wang et al., 2015; Lee et al., 2017).

Towards this end, recent evidence has indicated that mutations in regulatory proteins are likely significant contributors of gene expression evolution (Brem et al., 2002; Yvert et al., 2003; Bustamante et al., 2005). Interestingly, among

these mutations, gene duplication is an important driver of changing regulatory behavior in a gene network (Voordeckers et al., 2012; Baker et al., 2013; Perez et al., 2014; Pougach et al., 2014), resulting in GAL1 and GAL3. Gene duplication and the consequent divergence has been suggested to be involved in ecological adaptation in yeast, in another context (Thomson et al., 2005).

Comparing the GAL network with well-characterized sugar utilization systems in bacteria, we note that the GAL network in yeast is distinct in its structure. While most sugar utilization systems in *E. coli* employ only one regulaor (AraC for arabinose; XylR for xylose; LacI for lactose) (Aidelberg et al., 2014) [however, one exception, in the form of rhamnose utilization system does exist (Kolin et al., 2008)], the GAL system employs a cascade of three. We speculate that the additional number of regulators in the GAL network help individual strains to adapt to their precise ecological niche. Evolution experiments to test the relationship between complexity of regulatory structures and adaptive response of populations can shed more light on this aspect of cellular physiology and adaptation.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

AUTHOR CONTRIBUTIONS

ER: Performed simulations, analyzed the data; AM: Performed experiments, analyzed the data, and wrote the manuscript; SS: Conceived the study, wrote the manuscript.

FUNDING

This work is funded by DBT/Wellcome Trust (India Alliance), Grant number IA/S/19/2/504632 to SS. AM is supported by the Council of Scientific and Industrial Research (CSIR), Government of India, as a Senior Research Fellow (09/ 087(0873)/2017-EMR-I).

ACKNOWLEDGMENTS

AM, SS thank ICTS Bangalore for support towards attending the IVth Population Genetics and Evolution School, held in 2020. The authors thank Paike Jayadeva Bhat for discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.801011/full#supplementary-material

REFERENCES

Aidelberg, G., Towbin, B. D., Rothschild, D., Dekel, E., Bren, A., and Alon, U. (2014). Hierarchy of Non-glucose Sugars in *Escherichia coli*. *BMC Syst. Biol.* 8, 133.doi:10.1186/s12918-014-0133-z

Anjali, M., Alugoju, P., Kavatalkar, V., Rajeshkannan, E., Bhat, J., and Saini, S. (2021). *Public-good Driven Release of Heterogeneous Resources Leads to Genotypic Diversification of an Isogenic Yeast Population in Melibiose*. bioRxiv.

Baker, C. R., Hanson-Smith, V., and Johnson, A. D. (2013). Following Gene Duplication, Paralog Interference Constrains Transcriptional Circuit Evolution. *Science* 342, 104–108.doi:10.1126/science.1240810

Bhat, P. J., and Murthy, T. V. S. (2001). Transcriptional Control of the GAL/MEL Regulon of Yeast *Saccharomyces cerevisiae*: Mechanism of Galactose-Mediated Signal Transduction. *Mol. Microbiol.* 40, 1059–1066.doi:10.1046/j.1365-2958.2001.02421.x

Boocock, J., Sadhu, M. J., Durvasula, A., Bloom, J. S., and Kruglyak, L. (2021). Ancient Balancing Selection Maintains Incompatible Versions of the Galactose Pathway in Yeast. *Science* 371, 415–419.doi:10.1126/science.aba0542

Bram, R. J., Lue, N. F., and Kornberg, R. D. (1986). A GAL Family of Upstream Activating Sequences in Yeast: Roles in Both Induction and Repression of Transcription. *EMBO J.* 5, 603–608.doi:10.1002/j.1460-2075.1986.tb04253.x

Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic Dissection of Transcriptional Regulation in Budding Yeast. *Science* 296, 752–755.doi:10.1126/science.1069516

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Todd Hubisz, M., Glanowski, S., et al. (2005). Natural Selection on Protein-Coding Genes in the Human Genome. *Nature* 437, 1153–1157.doi:10.1038/nature04240

Chasman, D. I., and Kornberg, R. D. (1990). GAL4 Protein: Purification, Association with GAL80 Protein, and Conserved Domain Structure. *Mol. Cel Biol* 10, 2916–2923.doi:10.1128/mcb.10.6.2916-2923.1990

Chiang, H.-L., Schekman, R., and Hamamoto, S. (1996). Selective Uptake of Cytosolic, Peroxisomal, and Plasma Membrane Proteins into the Yeast Lysosome for Degradation. *J. Biol. Chem.* 271, 9934–9941.doi:10.1074/jbc.271.17.9934

Das Adhikari, A. K., Qureshi, M. T., Kar, R. K., and Bhat, P. J. (2014). Perturbation of the Interaction between Gal4p and Gal80p of theSaccharomyces Cerevisiae GAL Switch Results in Altered Responses to Galactose and Glucose. *Mol. Microbiol.* 94, 202–217.doi:10.1111/mmi.12757

Douglas, H. C., and Hawthorne, D. C. (1966). Regulation of Genes Controlling Synthesis of the Galactose Pathway Enzymes in Yeast. *Genetics* 54, 911–916.doi:10.1093/genetics/54.3.911

Gerstein, A. C., Chun, H.-J. E., Grant, A., and Otto, S. P. (2006). Genomic Convergence toward Diploidy in *Saccharomyces cerevisiae*. *Plos Genet.* 2, e145.doi:10.1371/journal.pgen.0020145

Gillespie, D. T. (1976). A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. *J. Comput. Phys.* 22, 403–434. doi:10.1016/0021-9991(76)90041-3

Golding, G. B., and Dean, A. M. (1998). The Structural Basis of Molecular Adaptation. *Mol. Biol. Evol.* 15, 355–369.doi:10.1093/oxfordjournals.molbev.a025932

Harari, Y., Ram, Y., Rappoport, N., Hadany, L., and Kupiec, M. (2018). Spontaneous Changes in Ploidy Are Common in Yeast. *Curr. Biol.* 28, 825–835.doi:10.1016/j.cub.2018.01.062

Hawkins, K. M., and Smolke, C. D. (2006). The Regulatory Roles of the Galactose Permease and Kinase in the Induction Response of the GAL Network in *Saccharomyces cerevisiae*. *J. Biol. Chem.* 281, 13485–13492.doi:10.1074/jbc.m512317200

Hill, M. S., Vande Zande, P., and Wittkopp, P. J. (2021). Molecular and Evolutionary Processes Generating Variation in Gene Expression. *Nat. Rev. Genet.* 22, 203–215.doi:10.1038/s41576-020-00304-w

Hittinger, C. T., and Carroll, S. B. (2007). Gene Duplication and the Adaptive Evolution of a Classic Genetic Switch. *Nature* 449, 677–681.doi:10.1038/nature06151

Johnston, M. (1987). A Model Fungal Gene Regulatory Mechanism: the GAL Genes of *Saccharomyces cerevisiae*. *Microbiol. Rev.* 51, 458–476.doi:10.1128/mr.51.4.458-476.1987

Johnston, S. A., and Hopper, J. E. (1982). Isolation of the Yeast Regulatory Gene GAL4 and Analysis of its Dosage Effects on the Galactose/melibiose Regulon. *Proc. Natl. Acad. Sci.* 79, 6971–6975.doi:10.1073/pnas.79.22.6971

Kim, Y. A., and Przytycka, T. M. (2012). Bridging the Gap between Genotype and Phenotype via Network Approaches. *Front. Genet.* 3, 227.doi:10.3389/fgene.2012.00227

Kolin, A., Balasubramaniam, V., Skredenske, J. M., Wickstrum, J. R., and Egan, S. M. (2008). Differences in the Mechanism of the Allosteric L-Rhamnose Responses of the AraC/XylS Family Transcription Activators RhaS and RhaR. *Mol. Microbiol.* 68, 448–461.doi:10.1111/j.1365-2958.2008.06164.x

Lee, K. B., Wang, J., Palme, J., Escalante-Chong, R., Hua, B., and Springer, M. (2017). Polymorphisms in the Yeast Galactose Sensor Underlie a Natural Continuum of Nutrient-Decision Phenotypes. *Plos Genet.* 13, e1006766.doi:10.1371/journal.pgen.1006766

Lewontin, R. C. (2002). Directions in Evolutionary Biology. *Annu. Rev. Genet.* 36, 1–18.doi:10.1146/annurev.genet.36.052902.102704

Lohr, D., Venkov, P., and Zlatanova, J. (1995). Transcriptional Regulation in the Yeast GAL Gene Family: a Complex Genetic Network. *FASEB j.* 9, 777–787.doi:10.1096/fasebj.9.9.7601342

Marcet-Houben, M., and Gabaldón, T. (2015). Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. *Plos Biol.* 13, e1002220.doi:10.1371/journal.pbio.1002220

Meyer, J., Walker-Jonah, A., and Hollenberg, C. P. (1991). Galactokinase Encoded by GAL1 Is a Bifunctional Protein Required for Induction of the GAL Genes in Kluyveromyces Lactis and Is Able to Suppress the Gal3 Phenotype in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* 11, 5454–5461.doi:10.1128/mcb.11.11.5454

Nogi, Y., and Fukasawa, T. (1984). Nucleotide Sequence of the Yeast Regulatory geneGAL80. *Nucl. Acids Res.* 12, 9287–9298.doi:10.1093/nar/12.24.9287

Peng, W., Liu, P., Xue, Y., and Acar, M. (2015). Evolution of Gene Network Activity by Tuning the Strength of Negative-Feedback Regulation. *Nat. Commun.* 6, 6226.doi:10.1038/ncomms7226

Pérez, J. C., Fordyce, P. M., Lohse, M. B., Hanson-Smith, V., Derisi, J. L., and Johnson, A. D. (2014). How Duplicated Transcription Regulators Can Diversify to Govern the Expression of Nonoverlapping Sets of Genes. *Genes Dev.* 28, 1272–1277.doi:10.1101/gad.242271.114

Platt, A., Ross, H. C., Hankin, S., and Reece, R. J. (2000). The Insertion of Two Amino Acids into a Transcriptional Inducer Converts it into a Galactokinase. *Proc. Natl. Acad. Sci.* 97, 3154–3159.doi:10.1073/pnas.97.7.3154

Platt, T. (1984). Toxicity of 2-deoxygalactose to *Saccharomyces cerevisiae* Cells Constitutively Synthesizing Galactose-Metabolizing Enzymes. *Mol. Cell. Biol.* 4, 994–996.doi:10.1128/mcb.4.5.994

Pougach, K., Voet, A., Kondrashov, F. A., Voordeckers, K., Christiaens, J. F., Baying, B., et al. (2014). Duplication of a Promiscuous Transcription Factor Drives the Emergence of a New Regulatory Network. *Nat. Commun.* 5, 4868.doi:10.1038/ncomms5868

Rodríguez-Trelles, F., Tarrío, R., and Ayala, F. J. (2003). Evolution of Cis-Regulatory Regions versus Codifying Regions. *Int. J. Dev. Biol.* 47, 665–673.

Roop, J. I., Chang, K. C., and Brem, R. B. (2016). Polygenic Evolution of a Sugar Specialization Trade-Off in Yeast. *Nature* 530, 336–339.doi:10.1038/nature16938

Salmeron, J. M., Jr., Leuther, K. K., and Johnston, S. A. (1990). GAL4 Mutations that Separate the Transcriptional Activation and GAL80-Interactive Functions of the Yeast GAL4 Protein. *Genetics* 125, 21–27.doi:10.1093/genetics/125.1.21

Sellick, C. A., Campbell, R. N., and Reece, R. J. (2008). Chapter 3 Galactose Metabolism in Yeast-Structure and Regulation of the Leloir Pathway Enzymes and the Genes Encoding Them. *Int. Rev. Cel Mol Biol* 269, 111–150.doi:10.1016/s1937-6448(08)01003-4

Thomson, J. M., Gaucher, E. A., Burgan, M. F., De Kee, D. W., Li, T., Aris, J. P., et al. (2005). Resurrecting Ancestral Alcohol Dehydrogenases from Yeast. *Nat. Genet.* 37, 630–635.doi:10.1038/ng1553

Venkataram, S., Dunn, B., Li, Y., Agarwala, A., Chang, J., Ebel, E. R., et al. (2016). Development of a Comprehensive Genotype-To-Fitness Map of Adaptation-Driving Mutations in Yeast. *Cell* 166, 1585–1596.doi:10.1016/j.cell.2016.08.002

Venturelli, O. S., El-Samad, H., and Murray, R. M. (2012). Synergistic Dual Positive Feedback Loops Established by Molecular Sequestration Generate Robust Bimodal Response. *Proc. Natl. Acad. Sci.* 109, E3324–E3333.doi:10.1073/pnas.1211902109

Voordeckers, K., Brown, C. A., Vanneste, K., Van Der Zande, E., Voet, A., Maere, S., et al. (2012). Reconstruction of Ancestral Metabolic Enzymes Reveals Molecular Mechanisms Underlying Evolutionary Innovation through Gene Duplication. *Plos Biol.* 10, e1001446.doi:10.1371/journal.pbio.1001446

Wang, J., Atolia, E., Hua, B., Savir, Y., Escalante-Chong, R., and Springer, M. (2015). Natural Variation in Preparation for Nutrient Depletion Reveals a Cost-Benefit Tradeoff. *Plos Biol.* 13, e1002041.doi:10.1371/journal.pbio.1002041

Wolfe, K. H. (2015). Origin of the Yeast Whole-Genome Duplication. *Plos Biol.* 13, e1002221.doi:10.1371/journal.pbio.1002221

Wray, G. A. (2007). The Evolutionary Significance of Cis-Regulatory Mutations. *Nat. Rev. Genet.* 8, 206–216.doi:10.1038/nrg2063

Yvert, G., Brem, R. B., Whittle, J., Akey, J. M., Foss, E., Smith, E. N., et al. (2003). Trans-acting Regulatory Variation in *Saccharomyces cerevisiae* and the Role of Transcription Factors. *Nat. Genet.* 35, 57–64.doi:10.1038/ng1222

Zenke, F. T., Engels, R., Vollenbroich, V., Meyer, J., Hollenberg, C. P., and Breunig, K. D. (1996). Activation of Gal4p by Galactose-dependent Interaction of Galactokinase and Gal80p. *Science* 272, 1662–1665.doi:10.1126/science.272.5268.1662

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Rajeshkannan, Mahilkar and Saini. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 29 March 2022
doi: 10.3389/fmolb.2022.801032

Modeling and Optimization of a Molecular Biocontroller for the Regulation of Complex Metabolic Pathways

*Yadira Boada, Fernando N. Santos-Navarro, Jesús Picó and Alejandro Vignoni**

Synthetic Biology and Biosystems Control Lab, Institut d'Automàtica i Informàtica Industrial, Universitat Politècnica de València, Valencia, Spain

Achieving optimal production in microbial cell factories, robustness against changing intracellular and environmental perturbations requires the dynamic feedback regulation of the pathway of interest. Here, we consider a merging metabolic pathway motif, which appears in a wide range of metabolic engineering applications, including the production of phenylpropanoids among others. We present an approach to use a realistic model that accounts for *in vivo* implementation and then propose a methodology based on multiobjective optimization for the optimal tuning of the gene circuit parts composing the biomolecular controller and biosensor devices for a dynamic regulation strategy. We show how this approach can deal with the trade-offs between the performance of the regulated pathway, robustness to perturbations, and stability of the feedback loop. Using realistic models, our results suggest that the strategies for fine-tuning the trade-offs among performance, robustness, and stability in dynamic pathway regulation are complex. It is not always possible to infer them by simple inspection. This renders the use of the multiobjective optimization methodology valuable and necessary.

Keywords: metabolic pathway, dynamic regulation, biomolecular antithetic controller, extended biosensor, tuning, gene circuit parts, multiobjective optimization, modeling biological systems

OPEN ACCESS

Edited by:
Alberto Jesus Martin, Universidad Mayor, Chile

Reviewed by:
Paola Lecca, Free University of Bozen-Bolzano, Italy
Frederic Cadet, INSERM Dynamics of Structures and Interactions of Biological Macromolecules (DSIMB), France

****Correspondence:***
Alejandro Vignoni
vignoni@isa.upv.es

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *24 October 2021*
Accepted: *22 February 2022*
Published: *29 March 2022*

Citation:
Boada Y, Santos-Navarro FN, Picó J and Vignoni A (2022) Modeling and Optimization of a Molecular Biocontroller for the Regulation of Complex Metabolic Pathways. Front. Mol. Biosci. 9:801032. doi: 10.3389/fmolb.2022.801032

1 INTRODUCTION

Microbial cell factory development using metabolic engineering seeks to obtain high levels of products of interest through genetic modification of microorganisms. Natural cells use complex regulatory networks to preserve robust growth and endure environmental changes by dynamically adapting cell metabolism (Liu et al., 2018). These regulation strategies are the long-term result of evolution. In most cases, they are not compatible with the addition of exogenous genes highly expressed to reach the production levels demanded by the industry. Constraint-based steady-state models of metabolism using only stoichiometric information and some basic information about the enzyme regulation have proved very valuable in providing predictions on maximum theoretical yields, optimal flux distribution to maximize flux towards some metabolite reaction bottlenecks and required ways of intervention on gene expression, leading to fluxes towards final products that achieve specified levels in productivity, titer and yield (Otero-Muras and Carbonell, 2021). This approach seeks the careful optimal selection of the constant expression levels of the exogenous genes in the pathway of interest and the endogenous ones with relevant interactions. Yet, as it is a static regulation approach, it fails to address the problem's dynamic and highly uncertain nature. Indeed,

the static strategy to regulate a metabolic pathway relies on an optimization process that is tailor-made for a particular situation, and therefore it is not able to respond to cell and environmental changes occurring during fermentation in a bioreactor (Wehrs et al., 2019).

Considering the metabolic network dynamics allows better analysis of the sensitivity of the metabolites or fluxes of interest to the optimal enzymatic intervention points under different environmental situations. Dynamic network models, from grey-box to black-box ones, of scales ranging from a subset of pathways to genome-scale, have been used to this end (Otero-Muras and Banga, 2017; Li et al., 2018; Yang et al., 2019; Lo-Thong et al., 2020). The optimal intervention points and intervention strategies (required up- or down-regulation) can be assessed using sensitivity analysis methods like metabolic control analysis (Lo-Thong et al., 2020) dynamic optimization (Otero-Muras and Banga, 2017; Li et al., 2018; Yang et al., 2019) and optimal control principles (Tsiantis and Banga, 2020). Thus, these methods address the fundamental problem of determining the structure of (optimal) control intervention points in complex metabolic networks. Yet, there are no generally applicable algorithms for designing metabolic dynamic feedback regulation systems to date. The regulation topology is generally pathway-specific, depending on both the potential presence of toxic pathway intermediates and the pathway topology (Hartline et al., 2020). Several typical metabolic topology motifs are usually considered: linear, branched, and merging (Blair et al., 2012). Most existing work has dealt with the dynamic regulation of linear pathways (Oyarzún and Stan, 2013; Liu and Zhang, 2018) or branched ones (Liu et al., 2018).

Once the optimal signals to be feedback and the intervention points are obtained, the problem of designing and tuning the proper dynamic feedback regulation biomolecular controller remains. Achieving robust optimal production in microbial cell factories requires considering the dynamic regulation of the pathway of interest. Dynamic feedback regulation constitutes a very interesting strategy to construct pathways with the ability to self-tune upon changing environmental conditions and to overcome many of the ongoing challenges faced in metabolic engineering (Liu and Zhang, 2018; Hartline et al., 2020). For example, it is often challenging to find the proper enzyme levels that maximize production while avoiding pathway bottlenecks or the accumulation of toxic intermediates. Feedback control circuits can solve these problems by dynamically changing enzyme expression in response to metabolic inputs and continuously regulating the activity in the pathway in response to either intracellular or bioreactor perturbations. This enables the industry to attain higher process performance indices than static regulation (Stevens and Carothers, 2015).

Despite the growing number of reported successful cases, engineering dynamic feedback control strategies in biological applications remains a major challenge (Gao et al., 2019). Model-based design, which leverages control engineering principles, can provide a powerful formalism to design dynamic feedback regulation circuits. This, together with the tools of synthetic biology, can lead to robust and efficient microbial production at the industrial level (Liu et al., 2018; Segall-Shapiro et al., 2018).

Here, we consider the design and tuning of a biomolecular controller for the dynamic feedback regulation of a merging metabolic pathway. Since we restrict to a single metabolic pathway, determining the dynamic regulation topology, i.e., the feedback variable and the intervention point, could be made by simple inspection and previous knowledge of the system. In this metabolic motif, two substrates, the primary precursor and an essential metabolite, are converted to an intermediate product which is subsequently transformed into a target product. The secondary essential metabolite plays an additional role in cell metabolism in many practical situations. Therefore, it is subject to environmentally-induced fluctuations. Over-expressing the enzyme that synthesizes this secondary metabolite or redirecting the flux towards it is not feasible in cases where its accumulation is toxic for the cell, leading to growth inhibition. This is the situation encountered in applications like the production of phenylpropanoids of industrial interest, e.g., naringenin (Sheng et al., 2020).

In previous work, we considered the problem of designing a dynamic regulation topology for the production of naringenin while coping with fluctuations in malonyl-CoA, the secondary essential metabolite (Boada et al., 2020). This work considers a detailed model of the whole system, including the metabolic pathway, the extended biosensor, and the molecular biocontroller. We address the problem of optimal choice (tuning) of the biocontroller and the biosensor components in the dynamic regulation topology. In particular, we considered a realistic model for the antithetic controller together with an extended biosensor based on the QdoR Transcription Factor (TF) that accounts for a straightforward *in vivo* implementation of the system. This gives us more information than the simplistic models of the antithetic biocontroller used in the literature that do not consider fundamental aspects like:

- non-linearities in the promoters. In the simplest models, the expression of proteins and sigma factors is always proportional to the number of transcription factors, i.e., there is no saturation of the promoters.
- formation of the antithetic complex, and the unbinding reaction of the complex.
- dilution rate of all the species due to cell growth. It is known that the dilution destroys the perfect adaptation property of the antithetic biocontroller, introducing a steady-state error. As we comment later, this forces us to use more than one objective to optimize.

Multiobjective optimization has already been demonstrated to be an appropriate tool for characterization of gene circuit parts (Boada et al., 2019a; Boada et al., 2019b), and for the design of gene circuits with the desired behavior (Boada et al., 2016; Boada et al., 2017b; Boada et al., 2021). Here, we present an approach to use multiobjective optimization for the optimal tuning of the gene circuit parts composing the biocontroller and biosensor in a dynamic metabolic regulation feedback loop. We show how this approach can deal with the trade-offs between the

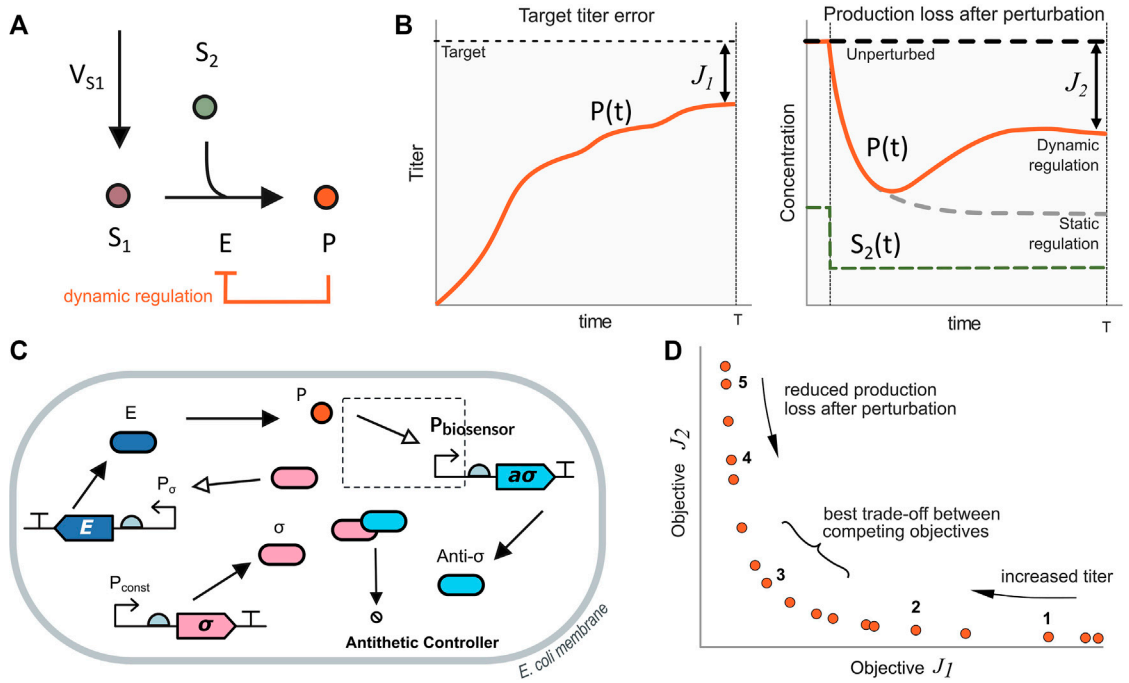


FIGURE 1 | Illustrative model system. **(A)** Metabolic pathway for the production of metabolite *P*. The main substrate S_1 and the secondary one S_2 are converted into the product *P* by the catalyst enzyme *E*. In the static regulation strategy (black lines), the expression level of the enzyme *E* remains constant in time. Conversely, in the dynamic regulation strategy (orange line), the expression of the enzyme *E* depends on the amount of product *P*. **(B)** Objective functions employed in this work for the maximization of the production up to a target value (J_1) together with the minimization of the production loss after perturbations (J_2) as defined in **Eq. 9**, **10 (C)** Biosensor and antithetic controller configuration for dynamic pathway regulation. The amount of free σ molecule determines the expression of the enzyme *E*. A TF-based biosensor detects the product levels and counteracts expressing the anti-σ molecule. When the amount of *P* decreases, the controller reduces the amount of expressed anti-σ, thus increasing the amount of free σ to up-regulate the enzyme *E*. **(D)** Pareto front of optimal solutions for the dynamical pathway regulation case. Solutions on the right side have large titer target error J_1 (i.e., lower titer) and a small production loss after perturbation J_2 (i.e., higher titer after the perturbation). Moving along the Pareto front towards the left, the titer target error decreases, and the production loss increases. Solutions in the middle of the Pareto front have the best trade-off between the competing objectives J_1 and J_2.

performance of the regulated pathway and robustness to fluctuations in the secondary metabolite. We also highlight that performance indices must include the standard steady-state industrial ones (e.g., titer) and indices related to the time-response transient (i.e., stability). As the complexity of the dynamic biocontrollers and biosensors integrated into the feedback loop regulation increases, the stability and transient performance issues that high order dynamics introduce must be taken into account. In this work, we consider, on the one hand, the case where having transcription factor (TF) based biosensors of the target product is not always possible. As an alternative, extended TF-based biosensors can be used, where an additional pathway is introduced from the target product to be regulated to a measurable metabolite (Boada et al., 2020). Yet, these extended biosensors present extra dynamics in the feedback loop. On the other hand, to regulate the amount of enzyme that catalyzes the conversion from the two precursor substrates into the product naringenin, we consider the use of the antithetic controller, a biomolecular integral feedback controller that achieves quasi-perfect adaptation (Briat et al., 2016; Aoki et al., 2019).

We first show our approach using a simple illustrative pathway that captures the essential topological features of merging metabolic pathways. We use a feedback regulation strategy encompassing a simple TF-based biosensor to obtain readouts of the product and a simplified model of the biomolecular antithetic controller. In this case, the final titer of the target product and the robustness to fluctuations in the secondary metabolite are evaluated. Then, we consider a detailed model of the metabolic merging pathway of naringenin, the biocontroller, and the extended biosensor of naringenin production that we previously introduced in (Boada et al., 2020). In this case, we use a more realistic model of the antithetic biocontroller. The extra dynamics introduced by both the extended biosensor and the biocontroller force us to consider the transient dynamics of the regulated feedback loop in the design process. A library of designs is obtained, each one corresponding to a different trade-off.

2 RESULTS

2.1 Tuning the Dynamic Regulation of a Merging Metabolic Pathway

To illustrate our approach's broad scope and usefulness, we first study a basic metabolic pathway that contains the main common features of a typical merging motif. As shown in **Figure 1A** (black

lines), we consider the production of a product metabolite P from a precursor substrate S_1 and a secondary substrate S_2. The reaction is catalyzed by the enzyme E. This metabolic pathway can be described using the following model:

$$\frac{dS_1}{dt} = V_{S_1} - V_{S_1,S_2} - \mu S_1 \tag{1}$$

$$\frac{dP}{dt} = V_{S_1,S_2} - \mu P \tag{2}$$

$$\frac{dX}{dt} = \mu X\left(1 - \frac{X}{X_{max}}\right) \tag{3}$$

Where S_i and P are the amount of substrates and product. X is the number of cells in the population. S_1 is the primary substrate, and S_2 is the secondary substrate. The first order dilution term represents the effect of cell growth on the amount of substrates and products, being μ the specific growth rate. X_{max} accounts for the maximum growing capacity of the population. The metabolic fluxes are given by the kinetic terms:

$$V_{S_1} = K_{S_1} \tag{4}$$

$$V_{S_1,S_2} = k_{cat} E \frac{S_1 S_2}{K_{mS_1} K_{mS_2} + K_{mS_2} S_1 + K_{mS_1} S_2 + S_1 S_2} \tag{5}$$

Where we assume that the uptake of the precursor S_1 has constant rate K_{S_1} (4), and the substrate S_2 is normally available at non-limiting amount. The flux V_{S_1,S_2} is described by means of the Michaelis-Menten kinetics in **Eq. 5**, where E is the amount of enzyme catalyzing the pathway, k_{cat} is the enzyme catalytic rate and K_{mS_i} are the Michaelis-Menten constants for the substrates.

In the case of static pathway regulation (**Figure 1A**, black line), the flux V_{S_1,S_2} has a constant maximum value determined by the amount of the constitutively expressed heterologous enzyme E. As its expression level is independent of any metabolite in the pathway, the production of P is affected in the presence of a sudden change in the availability of the secondary substrate S_2, as shown in the right plot of **Figure 1B** in dashed grey lines.

On the contrary, in the case of dynamic pathway regulation (**Figure 1A** in orange line), the level of expression of the enzyme E depends on the amount of the product metabolite. A biosensor provides product metabolite readouts, and a biomolecular controller changes the enzyme expression level as a function of the difference between the current amount of product and the target one encoded in the controller. Thus, when there is a change in the secondary substrate, the production of the metabolite P is affected but can recover (up to some extent) closer to its previous value (**Figure 1B**, right plot, solid orange line).

Different control architectures can be implemented with combinations of activation and repression feedback loops. Here, we focus on a class of biomolecular controller, the antithetic controller (Aoki et al., 2019), that allows for quasi-perfect adaptation.

To gain an initial understanding of the design trade-offs in the dynamic control of the merging metabolic pathway motif, we first consider a simplified version of the antithetic controller regulating the amount of enzyme E using a simple TF-based biosensor to obtain readouts of P (**Figure 1C**).

The control action is encoded in the amount of free σ molecules that activate the expression of the enzyme E through its promoter P_σ. We modeled the promoter using a generalized Hill function as in (Boada et al., 2020), including the effect of the plasmid copy number on the promoter activation function. The resulting dynamics of the amount of enzyme E is:

$$\frac{dE}{dt} = C_N a_0 + C_N a_1 \frac{\sigma^2}{k_{d20} C_N^2 + \sigma^2} - (d_E + \mu)E \tag{6}$$

The acting molecule σ is constitutively expressed (thus encoding for sort of a target set-point value) and binds the anti-σ molecule to form an inactive complex, effectively reducing the amount of free σ. The resulting dynamics of the amount of free σ molecules is:

$$\frac{d\sigma}{dt} = C_N k_\sigma - \gamma\sigma a\sigma - (d_\sigma + \mu)\sigma \tag{7}$$

Next, a TF-based biosensor detects the product P expressing the anti-σ molecule as a function of the product amount. A constitutively expressed Transcription Factor (TF) (equation omitted for brevity) binds to the product P inducing the expression of the anti-σ molecule. The dynamics of the amount of anti-σ molecules is:

$$\frac{da\sigma}{dt} = C_{N_{a\sigma}} k_{a\sigma} \frac{C_N^2\left(1 + \frac{P}{k_{dp}}\right)^2}{C_N^2\left(1 + \frac{P}{k_{dp}}\right)^2 + (TF)^2} - \gamma\sigma a\sigma - (d_{a\sigma} + \mu)a\sigma \tag{8}$$

When the amount of P decreases, so does the amount of expressed anti-σ, thus increasing the amount of free σ molecules and, this way, up-regulating the expression of the enzyme E. Next, we consider the optimal tuning of the gene circuit parts composing the antithetic biocontroller and the TF-based biosensor.

To characterize the trade-offs between reaching the desired titer target for P together with reducing the production loss after a perturbation on the level of the secondary metabolite S_2, as illustrated in **Figure 1B**, we considered two objective functions. For the first one, we looked for the difference between the titer of the product P in the bioreactor and the desired target value (J_1). For the second one, we focus on the production loss (amount of product expressed per cell) after a perturbation on S_2 (J_2). The corresponding expressions for both objectives to be jointly minimized are:

$$J_1 = \left|\text{Target} - KP_{\text{unperturbed}}(T)\right|, \quad \text{(target titer error)} \tag{9}$$

$$J_2 = \left|\frac{P_{\text{unperturbed}}(T) - P_{\text{perturbed}}(T)}{P_{\text{unperturbed}}(T)}\right|, \quad \text{(production loss after perturbation)}, \tag{10}$$

Where $P_{\text{unperturbed}}(T)$ is the product amount at the end of the experiment (time T), K is a conversion constant from the amount of product to titer, $P_{\text{perturbed}}(T)$ is the amount of product after a perturbation in the secondary metabolite. As J_1 describes the difference between the desired target titer and the actual one, lower values of J_1 correspond to larger titers. On the other hand, J_2 is related to the loss in production after a perturbation. Therefore,

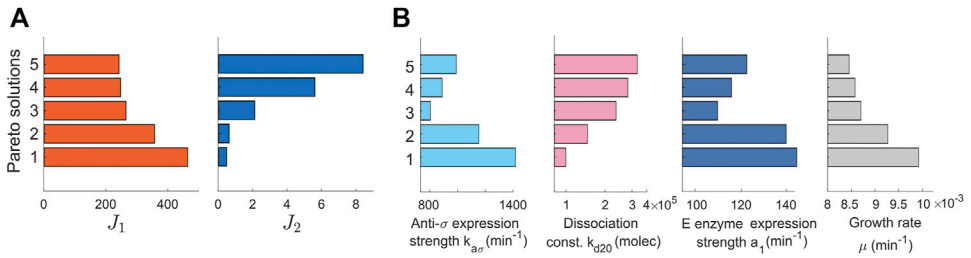


FIGURE 2 | Pareto solutions. Pareto front and Pareto set of selected solutions. **(A)** Pareto front showing the solutions of the multiobjetive problem. The values of the objectives J_1 and J_2 (x-axis) are represented for different solutions (y-axis). **(B)** The Pareto set represented with a plot for each tuned parameter. The tuned values of the parameters (x-axis) are shown for each selected Pareto solution (y-axis). The set of solutions constitutes the library of biocontrollers and biosensors obtained with the multiobjective optimization tuning process.

small values of J_2 correspond to low production loss after a perturbation. That is a better rejection of the perturbation on the secondary metabolite.

Next, we selected the biosensor and biocontroller set of parameters to be tuned. We took into account to what extent these parameters can be changed in the biological implementation of the system at the lab. Thus, we considered the set: expression strength for the enzyme E, a_1; the dissociation constant between σ and the enzyme promoter, k_{d20}; and the expression strength for anti-σ, $k_{a\sigma}$. Additionally, the specific growth rate, μ, was included as a decision parameter to account for the dependency of the results on cell growth.

The goal is to obtain a library of possible designs, each one corresponding to a different trade-off between the cost indices J_1, J_2. The resulting solutions are all equally optimal in the sense of Pareto (Boada et al., 2019b). When one of the objectives improves, the others necessarily deteriorate, so selecting the most appropriate solution depends on the designer.

We computed the values of the selected parameters as the solution of the multiobjective optimization problem min (J_1, J_2) subject to biologically plausible bounds on the values of the parameters (see **Supplementary Table S1** for a list of the solutions). Thus, for example, we set a growth rate corresponding to doubling times between 25 and 90 min, an upper bound for the dissociation constant of the promoter $k_{d20} <$ 3.5 μM and an upper bound on the maximum enzyme level of 180 μM. The optimization problem was solved using a multiobjective optimization genetic algorithm based on differential evolution. The detailed statement of the optimization problem is described in the Methods section, and the parameters used are in **Supplementary Table S2** (Supplementary Material).

The resulting Pareto front, **Figure 1D**, has three distinct regimes: 1) large titer target error and low production loss, 2) small titer target error and high production loss, and 3) the best trade-off regime between the two competing objectives. The convexity of the Pareto front indicates that the optimization problem is well-posed, in the sense that both objective functions oppose each other across the whole space of optimal solutions. We selected five solutions that represent the mentioned regimes. These solutions are highlighted in the Pareto front in **Figure 1D**. The achieved objective values of the selected solutions are shown in **Figure 2A** and the corresponding tuned optimal values for the controller and biosensor parameters and the growth rate are shown in **Figure 2B** (see **Supplementary Figure S1**, for details on the temporal responses of the selected solutions). Thus, the set of solutions of the optimization constitutes a library of optimally tuned controller-biosensor pairs.

A detailed inspection of the library of controller and biosensor pairs obtained (**Figure 2B**) reveals that the relations between parameters and objectives are not necessarily monotonous (**Supplementary Table S1**, Supplementary Material). For example, the dissociation constant k_{d20} must be chosen smaller to reduce the production loss after perturbation (J_2). Yet, there is no monotonous trend in neither the anti-σ expression strength $k_{a\sigma}$ nor in the E enzyme expression strength a_1.

Altogether these results suggest that strategies for fine-tuning the trade-off between target titer error and production loss in dynamic pathway regulation are complex and impossible to obtain by simple inspection even for a simplified case, rendering the use of the multiobjective optimization methodology not only helpful but necessary.

2.2 Model of the Dynamic Regulation of the Naringenin Metabolic Pathway

Naringenin is a flavonoid compound predominantly found in grapefruits and oranges. It has been reported to have many pharmacological properties, including anti-dyslipidaemic, anti-obesity and anti-diabetic (Liu et al., 2008; Zygmunt et al., 2010; Rahigude et al., 2012). Flavonoids are an essential subclass of phenylpropanoids, an important family of plant natural products with diverse uses as food supplements, antioxidants, flavoring and flavoring agents, pharmaceuticals, insecticides and colorants. Significant market opportunities clearly exist for flavonoids with enhanced bioavailability and bioactivity profiles that are used, among others, as flavorings and bioactive compounds for nutraceutical applications.

The naringenin pathway has four enzymatic steps from the L-tyrosine precursor (see **Figure 3**). The third step, catalysed by the naringenin chalcone synthase enzyme (CHS) requires

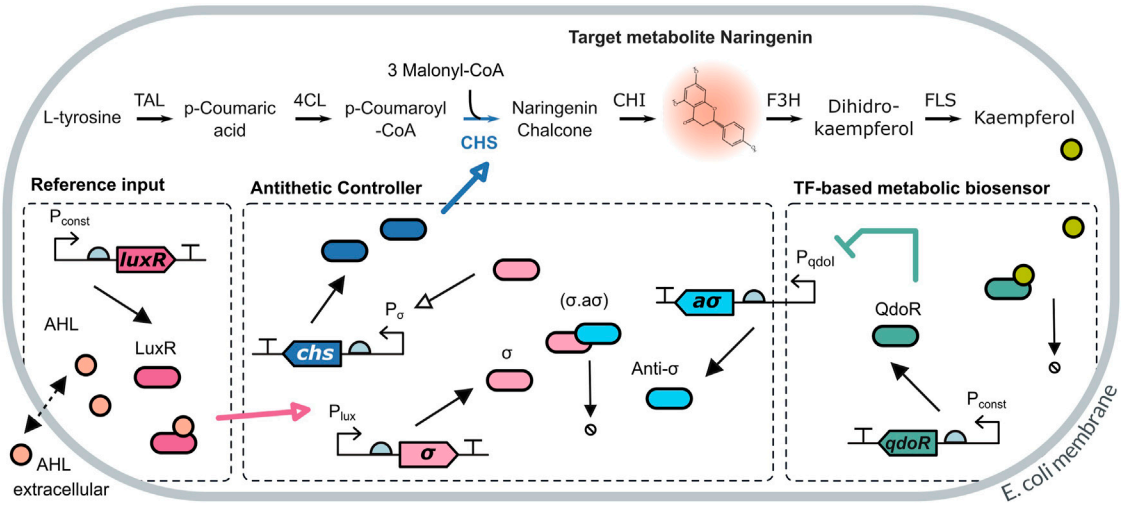


FIGURE 3 | Dynamic pathway regulation scheme for the naringenin pathway. The target metabolite naringenin is produced from L-tyrosine in four enzymatic steps, including a merging step catalysed by the enzyme naringenin chalcone synthase (CHS) which incorporates the secondary metabolite malonyl-CoA. The production level of naringenin is readout using a metabolic extended TF-based biosensor through the downstream metabolite kaempferol. This is sensed using the QdoR TF-based biosensor and feeds back to an antithetic biomolecular controller. The controller can be activated upstream by means of the external inducer AHL. Its actuating signal overdrives the basal constitutive expression of the CHS enzyme in the pathway in order to compensate for malonyl-CoA depletion.

the co-substrate malonyl-CoA, an essential metabolite that is used in fatty acid production and plays an important role in cell metabolism. Intracellular concentrations of malonyl-CoA are typically low (4–40 μM in *E. coli*) (Xu et al., 2014; Johnson et al., 2017). Moreover, its concentration is subject to fluctuations caused by cell environmental heterogeneity. Accumulation of malonyl-CoA is toxic for the cell, so that over-expressing it is not a feasible solution.

We considered a detailed model of the naringenin pathway by obtaining the mass balance equations of the enzyme-catalyzed reactions of the metabolic pathway from *L*-tyrosine to naringenin (see **Figure 3**). From the mass balance equations, we obtained the set of rate **Eq. 11**. The rate equations include the dilution effect of cell replication at a specific growth rate μ.

$$\begin{aligned}
\frac{dLt}{dt} &= V_0 - V_{Lt} - \mu Lt \\
\frac{dpC}{dt} &= V_{Lt} - V_{pC} - \mu pC \\
\frac{dpA}{dt} &= V_{pC} - V_{pA,Ma} - \mu pA \\
\frac{dNc}{dt} &= V_{pA,Ma} - V_{Nc} - \mu Nc \\
\frac{dN}{dt} &= V_{Nc} - V_N - \mu N
\end{aligned} \tag{11}$$

For each reaction, the corresponding flux is V_j (molecules min^{-1}). Lt is the number of molecules of *L*-tyrosine, pC is *p*-coumaric acid, pA is *p*-coumaroyl-CoA, Nc is naringenin chalcone, and N is the target metabolite naringenin. Next, we assumed that the fluxes V_j follow the Michaelis-Menten kinetics (Michaelis and Menten, 1913), and the flux V_0 from the L-tyrosine precursor is kept as a constant, obtaining the equations:

$$\begin{aligned}
V_0 &= K_{Lt} \\
V_{Lt} &= \mathrm{k_{catTAL}} TAL \frac{Lt}{\mathrm{K_{mLt}} + Lt} \\
V_{pC} &= \mathrm{k_{cat4CL}} 4CL \frac{pC}{\mathrm{K_{mpC}} + pC} \\
V_{pA} &= \mathrm{k_{catCHS}} CHS \frac{pAMa}{\mathrm{K_{mpA}}\,\mathrm{K_{mMa}} + \mathrm{K_{mMa}} pA + \mathrm{K_{mpA}} Ma + pAMa} \\
V_{Nc} &= \mathrm{k_{catCHI}} CHI \frac{Nc}{\mathrm{K_{mNc}} + Nc} \\
V_N &= \mathrm{k_{catF3H}} F3H \frac{N}{\mathrm{K_{mN}} + N}
\end{aligned} \tag{12}$$

Where Ma is the number of malonyl-CoA molecules naturally available inside the cell, $\mathrm{k_{catj}}$ is the catalytic rate of each enzyme (min^{-1}), and K_{mj} is the Michaelis-Menten constant for each substrate. The enzyme kinetic parameters, detailed in **Supplementary Table S4** (Supplementary Material) were obtained from Brenda (Schomburg et al., 2017) and optimized according to the requirements for the pathway implementation in the lab.

Malonyl-CoA is one of the major building blocks for cell metabolism. Its intracellular concentration is tightly regulated and maintained at small amounts (Yang et al., 2015). Therefore, our system, the exogenous naringenin pathway, will compete for this resource. Thus, from the point of view of our system, any variation in the Ma level caused by changes in the cell will act as a perturbation signal. We considered a basal value of Ma in the mid-range of values reported in the literature (Takamura and Nomura, 1988; Xu et al., 2014; Wu et al., 2015), and avoiding the accumulation of large amounts of intermediate metabolites that may lead to growth inhibition.

The amounts of the enzymes TAL, 4CL, CHI, CHS, and F3H involved in the naringenin pathway were previously optimized so that the flux of precursor L-tyrosine can yield the targeted 1 g/L of naringenin (see **Supplementary Table S4**, Supplementary Material). Compared to other models, we are explicitly modeling the amount of the enzymes of interest (CHS) as variables of our model to capture the interaction between the genetic control level and the metabolic pathway level.

2.2.1 Feedback Regulation *via* a Metabolic Biosensor and Biocontroller for the Naringenin Pathway

For the naringenin pathway, we implemented a dynamic regulation strategy including an extended biosensor to obtain the readout of naringenin and a biomolecular controller. On the one hand, the TF-based biosensor provides readouts of the amount of naringenin *via* a short metabolic pathway from naringenin to kaempferol (see **Figure 3**). Kaempferol is the effector flavonoid measured by the biosensor promoter region P_{qdoI} and the QdoR transcription factor (TF) (Siedler et al., 2014). When kaempferol captures QdoR, the TF is inactivated, and repression of the P_{qdoI} promoter becomes weaker while leading to an increase of anti-σ factor production. In contrast, lower concentrations of kaempferol allow higher amounts of QdoR transcription factor, which inhibits anti-σ expression.

For every *i*—cell in the population, the kinetics of the enzyme-catalyzed reactions involved in this extended pathway were modeled using the set of rate **Eq. 13**:

$$\begin{aligned}\frac{d\,Di}{dt} &= V_N - V_{Di} - \mu Di\\ \frac{dKa}{dt} &= V_{Di} - \mu Ka\\ \frac{d\,Q}{dt} &= \frac{\mathrm{p_Q C_N k_Q}}{\mathrm{dm_Q} + \mu} - (\mathrm{d_Q} + \mu)Q\end{aligned} \tag{13}$$

Where Di is the number of molecules of Dihydrokaempferol, Ka is kaempferol, the flux V_{Di} obeys the Michaelis-Menten kinetics $V_{Di} = \mathrm{k_{catFLS}} FLS \frac{Di}{\mathrm{K_{mDi}} + Di}$, and Q is the constitutively expressed QdoR protein. All the parameters are listed in **Supplementary Tables S4, S5** (Supplementary Material) and optimized according to the lab implementation and characterization in (Boada et al., 2019b; Dunstan et al., 2020).

The antithetic controller used for the dynamic regulation of naringenin production is depicted in **Figure 3**. The antithetic motif relies on the annihilation mechanism between both σ and anti $-\sigma$ factor proteins. The σ factor activity is controlled by the anti $-\sigma$ factor that binds to and keeps the σ factor sequestered. The anti $-\sigma$ is only released and de-repressed in response to the QdoR transcription factor. The dimer formed by the LuxR protein and AHL lactone activates the P_{LuX} promoter, inducing the synthesis of σ factor. The externally added concentration of AHL acts as the desired reference input for naringenin production. We do not assume the AHL concentration needed to set the desired value for naringenin must be equal to this one—implying an unnecessary metabolic burden—but simply proportional. Free σ factor binds to the P_{20} promoter to activate expression of the naringenin chalcone synthase CHS, which subsequently converts p-Coumaroyl-CoA and malonyl-CoA into the naringenin precursor. In other words, the CHS enzyme represents the controller output signal.

Considering the same assumptions as those to derive the TF-based biosensor model, the dynamics of the antithetic controller for every cell is given by the following set of equations:

$$\begin{aligned}\frac{d\sigma\cdot a\sigma}{dt} &= \frac{\mathrm{k_{-c}}}{\mathrm{k_{dc}}}\sigma\, a\sigma - \mathrm{k_{-c}}\sigma\cdot a\sigma - (\mathrm{d_c}+\mu)\sigma\cdot a\sigma\\ \frac{d\sigma}{dt} &= \frac{\mathrm{p_\sigma C_N k_\sigma}}{\mathrm{dm_\sigma}+\mu}\left(\alpha + \frac{(1-\alpha)A^2}{\mathrm{k_{dlux}}\left(\frac{\mathrm{k_{d2}C_N}}{R}\right)^2 + A^2}\right) - \frac{\mathrm{k_{-c}}}{\mathrm{k_{dc}}}\sigma\, a\sigma + \mathrm{k_{-c}}\sigma\cdot a\sigma - (\mathrm{d_\sigma}+\mu)\sigma\\ \frac{da\sigma}{dt} &= \frac{\mathrm{p_{a\sigma} C_{N_{a\sigma}} k_{a\sigma}}}{\mathrm{dm_{a\sigma}}+\mu}\left(\alpha + \frac{(1-\alpha)(\mathrm{k_{dq}C_N})^2(\mathrm{k_{dk}}+\mathrm{Ka})^2}{(\mathrm{k_{dq}C_N})^2(\mathrm{k_{dk}}+Ka)^2 + (\mathrm{k_{dk}Q})^2}\right) - \frac{\mathrm{k_{-c}}}{\mathrm{k_{dc}}}\sigma a\sigma\\ &\quad + k_{-c}\sigma\cdot a\sigma - (\mathrm{d_{a\sigma}}+\mu)a\sigma\\ \frac{dCHS}{dt} &= \beta\frac{\mathrm{p_{H_c} C_{Nh} k_H}}{\mathrm{dm_H}+\mu} + \frac{\mathrm{p_H C_{Nh} k_H}}{\mathrm{dm_H}+\mu}\left(\alpha + \frac{(1-\alpha)\sigma^2}{\mathrm{k_{d20}}(\mathrm{k_{d\sigma}C_{Nh}})^2 + \sigma^2}\right) - (\mathrm{d_H}+\mu)CHS\end{aligned} \tag{14}$$

Where $\sigma\cdot a\sigma$ is the amount of molecules from the generated complex after σ sequestration, σ and $a\sigma$ are the factor and its cofactor, respectively. All the parameters are listed in **Supplementary Table S5** (Supplementary Material).

As in (Boada et al., 2020), the desired naringenin set-point is regulated by the external addition of AHL and the constitutive expression of the LuxR protein. The passive diffusion of extracellular AHL inside the cell was modeled as a reversible pseudo-reaction using mass-action kinetics (Boada et al., 2017a). This resulted in the set of **Eq. 15**:

$$\begin{aligned}\frac{dR}{dt} &= \frac{\mathrm{p_R C_N k_R}}{\mathrm{dm_R}+\mu} - (\mathrm{d_R}+\mu)R\\ \frac{dA}{dt} &= \mathrm{D}(\mathrm{V_c}A_e - A) - (\mathrm{d_A}+\mu)A\\ \frac{dA_e}{dt} &= \mathrm{D}\left(-x\mathrm{V_c}A_e + \sum_{i=1}^{x} A\right) - \mathrm{d_{Ae}}A_e\\ \frac{dx}{dt} &= \mu\left(1 - \frac{x}{\mathrm{x_{max}}}\right)x\end{aligned} \tag{15}$$

Where R is the number of molecules of LuxR, A and A_e are the intra and extracellular AHL molecules, respectively, the term $\mathrm{V_c} = \frac{\mathrm{V_{cell}}}{\mathrm{V_{ext}}}$ is the ratio between the cellular and the culture volumes; and x is the number of cells. The parameters are also listed in **Supplementary Table S5** (Supplementary Material).

Using the set of preliminary parameters in **Supplementary Table S5**, we ran computational simulations to obtain the temporal response of the system to perturbation in Malonyl-CoA. Once the production of the target metabolite naringenin reached steady-state, we introduced a perturbation in the availability of Malonyl-CoA at 65 h of 60%. After that, the amount of σ factor increases leading to an increased expression of the enzyme CHS. This results in a slight increase in the naringenin production. However, as seen in **Figure 4**, there is room for improvement. This will be the goal obtained in the next section by means of the optimal tuning of the regulation strategy.

2.3 Optimal Tuning of the Dynamic Regulation for Naringenin Production

Having developed a detailed and realistic model of the pathway dynamic regulation, we optimally tuned the controller and

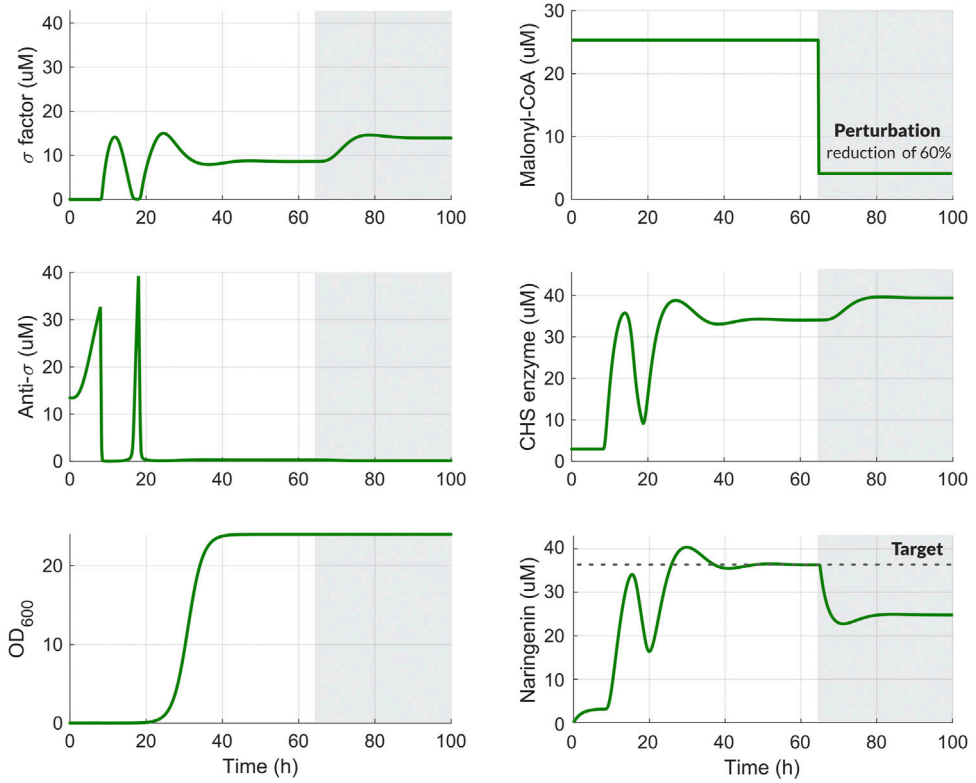


FIGURE 4 | Temporal response of the system including the dynamic pathway regulation scheme for the naringenin pathway. Time-course variation in the biocontroller species (σ and Anti-σ), malonyl-CoA secondary substrate, CHS enzyme, naringenin, and cellular growth (OD) before (white background) and after 60% reduction in malonyl-CoA availability (grey background). After the perturbation, the amount of naringenin begins to decline until it recovers steadily thank to the transient increase in the amount of the CHS enzyme. This increase is generated by the activation of factor when the perturbation occurs.

biosensor components of the dynamically regulated metabolic pathway that produces naringenin in *Escherichia coli* (*E. coli*).

We first established a baseline production pathway for naringenin, with the basal level of the CHS enzyme provided by a constitutive promoter. On top of this, the feedback control loop regulates the total level of expression of CHS to give a robust response to the fluctuations in the secondary co-substrate malonyl-CoA availability and drive the production of naringenin up to the target industrially relevant value of 1 g L^{-1}.

Next, we used our multiobjective optimization approach to find a library of optimal biocontroller and biosensor pairs for the dynamic regulation of the naringenin pathway. As in the previous example, we aim to determine controller and biosensor designs that allow reaching a target titer of naringenin production while minimizing the production loss after perturbations in the secondary metabolite. However, the extra dynamics introduced by the extended biosensor used in this case and the ones introduced by using a more realistic model of the biomolecular antithetic controller must be considered. These extra dynamics force us to consider the transient behavior of the regulated feedback loop in the design process to evaluate the overall stability properties of the designed system. Thus, we defined the following three objective functions to be minimized:

$$J_1 = \left|\text{Target} - KN_{\text{unp}}(T)\right|, \quad (\text{target titer error}) \tag{16}$$

$$J_2 = \left|\frac{N_{\text{unp}}(T) - N_{\text{pert}}(T)}{N_{\text{unp}}(T)}\right|, \quad (\text{production loss after perturbation}), \tag{17}$$

$$J_3 = (\text{Number of oscillations in } \sigma), \tag{18}$$

Where N_{unp} and N_{pert} are the amount of naringenin before and after the malonyl perturbation at the time T, respectively, and J_3 is an indirect measure of the frequency and damping factor of the transient in the antithetic biocontroller (see Methods for a detailed description).

A preliminary parameter sensitivity analysis of the biological parts from the biocontroller and the biosensor revealed six parameters to be tuned in the optimization process (Boada et al., 2020): the translation rates of both anti-σ and the CHS enzyme, $p_{a\sigma}$ and p_H; the plasmid copy numbers $C_{N_{a\sigma}}$ and C_{N_h}; the σ anti-σ complex dissociation rate, k_{-c}; and the dissociation constant between σ and the CHS enzyme promoter k_{d20}. The majority of these parameters are also easy to tune in the real implementation in the lab. Additionally, as in the simplified example before, we also considered the growth rate, μ, as a decision parameter. We computed them as the optimal

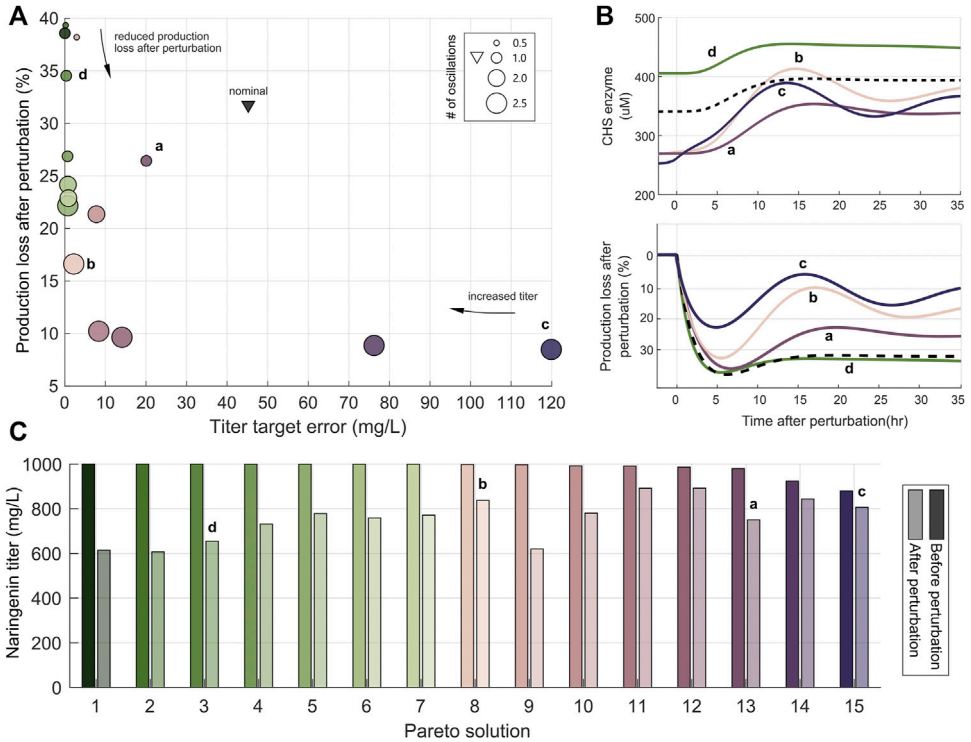


FIGURE 5 | Optimal controller and biosensor tuning in the naringenin dynamic pathway regulation. **(A)** Pareto front of the optimal solutions. The *x*-axis is the objective J_1 (titer target error), the *y*-axis is the objective J_2 (production loss after perturbation), and the size of the circles represent the objective J_3 (number of oscillations) that take into account the transient response of the controller to a perturbation. Solutions along the Pareto front are identified with color ranging from dark green to dark violet as the values of objective J_1 increase. Green solutions have smaller target error than pink/purple/violet ones. The black triangle represents the preliminary not optimized configuration. Solutions *a-d* are representative of the different zones along the Pareto front. **(B)** Time response of the selected solutions. Top plot: response of the CHS enzyme after a perturbation in the secondary metabolite Malonyl-CoA. Bottom plot: time response of the production loss after perturbation with respect to the level achieved before the perturbation for the four selected solutions. In both plots the black dashed line corresponds to the preliminary not optimized solution. **(C)** Naringenin titer before and after perturbation of each one of the Pareto solutions. The color codes are common to all the plots in the figure.

solutions of the multiobjective problem min (J_1, J_2, J_3). The details of the optimization problem can be found in the Methods section and details on the obtained solutions can be found in the Supplementary Material, including the list of the solutions in **Supplementary Table S3**, the representation of the Pareto front in **Supplementary Figure S2**, the Pareto set in **Supplementary Figure S3** and the temporal responses in **Supplementary Figure S4**.

The Pareto front resulting from the optimization is shown in **Figure 5A**. Several interesting aspects arise from it. First, all the solutions found are better than the preliminary configuration that was not optimized. They have either smaller titer target error or lower production loss after perturbation or both. Second, the relevance of considering the objective J_3, related to the transient characteristics, can be clearly seen. This objective is represented by the size of the circles that correspond to each of the solutions in the Pareto front. Notice that reducing the production loss can be achieved at the cost of increasing the titer error as also seen before in the previous example. But in this case, an increased capacity to reject perturbations also increases the number of oscillations in the response, and thus corresponds to a less marginally stable configuration of the controller.

However, unlike in the previous example, in which the oscillations were no taken into account in the optimization process, now it is possible to obtain a compromise design (see solution *a* in **Figure 5**) that has small titer error and low number of oscillations (a lightly under-damped response in **Figure 5B**) at a fairly low production loss cost. Notice that the apparent best trade-off without considering J_3 (with solution *b* as representative) has too an under-damped and lengthy transient, which could be unacceptable in some cases. Extreme solutions like (*c*) which has the lowest production loss or (*d*) with the smallest titer error may be of interest in particular cases when one objective has more practical importance than the others.

The trade-off is evident in the case of solutions *b*,*c* and *d* when only looking at the first two objectives. In this situation, solution *b* is the obvious best trade-off between titer error and production loss as it also clearly seen in **Figure 5C**. However, taking into account for the transient response of the biocontroller (J_3) shown in **Figure 5B**, solution *a* arises as a better compromise with less under-damped response. A detailed inspection of the library of controller and biosensor pairs obtained (**Figure 6**) again reveals the complex non monotonous relationship between parameters and objectives.

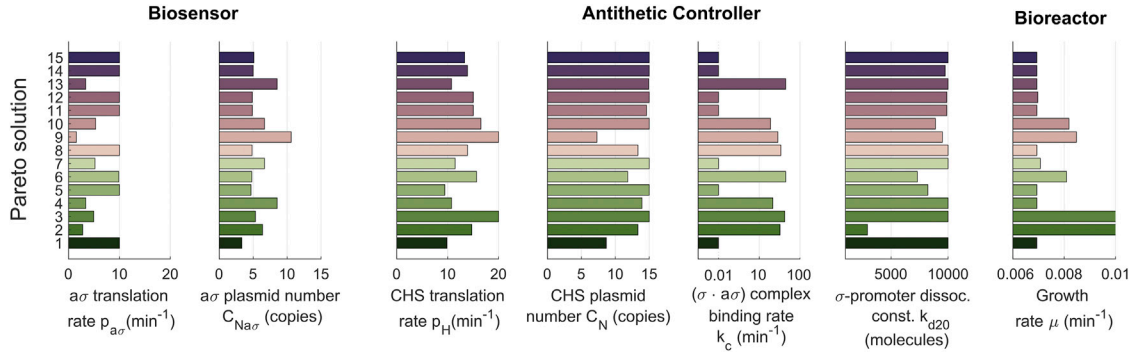


FIGURE 6 | Library of optimal controller-biosensor devices for the dynamic regulation of the naringenin pathway. Pareto set representing the optimal tuned values of the controller-biosensor parameters (decision variables). On the *x*-axis we show the optimal values of the parameters for each solution. Each one of the solutions (in the *y*-axis) constitutes an element of the controller-biosensor library. The color code is the same as the one used in **Figure 5**.

Taking a deeper look into the obtained library, we can also make an interesting observation: different combinations of parameters result in similar performances. This is nothing more than another evidence of the inherent robustness obtained with negative feedback control. For example, devices 5, 6, and 7 from the library (**Figure 6**, see **Supplementary Figures S2–S4** in the Supplementary Material for more details) have an approximately equivalent performance regarding the three objectives. Still, they significantly differ in their parameter values. Device 6 works in a faster-growing culture without losing titer, with a higher CHS translation rate than the other two devices, but needs a higher biding rate for the ($\sigma \cdot a\sigma$) complex. Depending on the available biological parts, one implementation can be more feasible than another, increasing the importance of having such a variety of elements in the library.

3 DISCUSSION

Dynamic regulation of metabolic pathways is a crucial strategy to achieve optimal production in microbial cell factories while coping with cell and environmental fluctuations. The appropriate dynamic regulation topology will be particular to the topology and characteristics of the metabolic pathway to be regulated. Yet, on the one hand, some basic metabolic motifs that often appear in practical industrial applications and their appropriate dynamic regulation topology can be identified. On the other hand, all dynamic regulation schemes share a set of common features that determine to a great extent the appropriate methodological tools required for the optimal selection of the gene circuit parts composing them. In particular, there is the common need to address a set of multiple goals related to the system's performance in terms of both the production of the targeted product and the rejection of perturbations affecting it. In addition, the stability issues that arise as a result of using feedback regulation strategies must be addressed. This is even more important as we use complex biomolecular controllers and metabolic extended biosensors that introduce extra dynamics that may compromise the regulated system's transient time response and stability.

In this work, we have shown the application of a general approach based on multiobjective model-based optimization for building libraries of gene circuit parts that achieve optimal performance of a dynamically regulated merging metabolic pathway. This metabolic motif appears in many situations of practical interest and, in particular, is a pervasive motif in producing phenylpropanoid-derived natural products.

The multiobjective optimization approach obtains devices within resulting libraries with different combinations of parameter values but similar performances. This is another sign of the inherent robustness obtained with negative feedback control. Interestingly, depending on the available biological parts, one implementation can be more feasible than another, increasing the importance of having such a variety of elements in the library.

We used detailed models of the metabolic kinetics and the biosensor and biocontroller dynamics constituting a sort of *in vivo* construction guidelines, as some of the model parameters can be directly related to biological parts or devices in the laboratory. Our results show that using this type of model with enough granularity also forces us to consider the transient and stability issues that are often disregarded.

The need for enough detailed models arguably includes the use of host-aware models. Indeed, the library of designs we obtained might suffer some modifications in case we considered the interactions between the regulated metabolic pathway and the host cell caused by competition for cell resources (Santos-Navarro et al., 2021). Our goal in this work was to present the general multiobjective optimization approach, emphasizing the tuning of the biomolecular controller and biosensor. In any case, the use of host-aware models will not change the general framework; it will only change the obtained solutions.

Altogether our results suggest that strategies for fine-tuning the trade-offs between target performance, robustness, and stability in complex dynamic pathway regulation topologies are intricate and not possible to obtain by simple inspection, rendering the use of the multiobjective optimization methodology not only helpful but necessary. As a consequence, it will not be generally possible to obtain widely

applicable optimal simple rules for the design. Instead, the expected outcome of the tuning process should be libraries of gene circuit components that achieve specific trade-offs and specific nominal environmental situations.

4 MATERIALS AND METHODS

4.1 Multiobjective Optimization

Generally, a multiobjective problem is faced by building an aggregate function in order to assemble the objectives in a unique index that contains a weighting vector for each objective. However, the solution obtained is determined by the selection of the weighting values. An alternative option is to use a multiobjective optimization design (Miettinen, 1999). In multiobjective optimization all objectives are important, therefore all of them are optimized simultaneously. Instead of one rarely unique solution, we obtain a set of the best solutions known as *Pareto Front*. In this front, all solutions are *Pareto-optimal* and only differ from each other in the trade-off of objectives each one represents. Multiobjective optimization requires at least three fundamental steps (Miettinen et al., 2008): 1) the multiobjective problem definition (MOP), 2) the optimization process, and 3) the multicriteria decision making process (MCDM). The overall multiobjective optimization design enables us to analyze current trade-offs between the objectives and select the most suitable solutions (Reynoso-Meza et al., 2013) that reaches all of our objectives.

4.2 Multiobjective Problem Definition

As referred in (Miettinen et al., 2008), a Multiobjective Problem (MOP), can be stated as follows:

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \mathbf{J}(\boldsymbol{\theta}) = [J_1(\boldsymbol{\theta}), \ldots, J_m(\boldsymbol{\theta})] \\ &\text{subject to:} \quad \mathbf{K}(\boldsymbol{\theta}) \leq 0 \\ &\qquad\qquad \underline{\theta_i} \leq \theta_i \leq \overline{\theta_i}, \quad \forall i = [1, \ldots, n] \end{aligned} \tag{19}$$

Where $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_n]$ is the decision vector that contains the decision variables for multiobjective optimization; $\mathbf{J}(\boldsymbol{\theta})$ is the objective vector and $\mathbf{K}(\boldsymbol{\theta})$, $\mathbf{L}(\boldsymbol{\theta})$ are the inequality and equality constraint vectors, respectively, $\underline{\theta_i}$, $\overline{\theta_i}$ are the lower and upper bounds in the decision variables space $\boldsymbol{\Theta}$. The MOP (4.2) has a set of solutions whose values in the Pareto front are function of the decision variables defined as the Pareto Set $\boldsymbol{\Theta_P}$. Each solution in this set corresponds to an optimal objective vector in the Pareto Front $\mathbf{J_P}$. All solutions in the the Pareto Set are Pareto-optimal non dominated solutions, that is, they differ from each other in the objectives trade-off each one represents.

4.3 Multiobjective Problem of the Merging Pathway

Here, the objective vector $\mathbf{J}(\boldsymbol{\theta})$ has to be defined to solve the problem presented in 2.1. We maximized the desired target titer for product P while minimizing the perturbation effects on the substrate S_2 dynamics. The objective functions $\mathbf{J_1}$ and $\mathbf{J_2}$ were defined before in **Eqs. 9**, **10**, and the decision variables $\boldsymbol{\theta}$ used for our optimization are $\boldsymbol{\theta} = [a_1, k_{d20}, k_{a\sigma}, \mu]$ with their corresponding lower and upper bounds as detailed in **Table 1**.

TABLE 1 | Lower and upper bounds for the merging pathway MOP.

Bound	$k_{a\sigma}$ (min^{-1})	k_{d20} (*molec*)	a_1 (min^{-1})	μ (min^{-1})
Lower	700	100,000	90	0.005
Upper	1,500	350,000	160	0.01

Hence, the MOP in (4.2) can be stated as:

$$\begin{aligned} &\min_{\boldsymbol{\theta} \in \Re^4} J(\boldsymbol{\theta}) = [J_1(\boldsymbol{\theta}), J_2(\boldsymbol{\theta})] \in \Re^2 \\ &\text{subject to : equations } (1-8) \end{aligned} \tag{20}$$

4.4 Naringenin Pathway as a Multiobjective Problem

For the dynamic regulation of the naringenin pathway, we defined three objective functions to tune both antithetic controller together with the biosensor:

$$J_1 = \left|\text{Target} - K_{Nar} N_{\text{unp}}(T)\right|, \quad (\text{mg/L target titer error}) \tag{21}$$

$$J_2 = \left|\frac{N_{\text{unp}}(T) - N_{\text{pert}}(T)}{N_{\text{unp}}(T)}\right|, \ (\%\ \text{production loss after perturbation}) \tag{22}$$

$$J_3 = \frac{1}{2}\sum_{k=1}^{N} \left(X(k) - X(k-1)\right)^2, \quad (\text{number of oscillations}) \tag{23}$$

Where $K_{Nar} = \frac{\text{mw} \cdot x(T)}{A_v}$ is a constant that converts amounts of naringenin into grams per liter the titer units, mw = 272.25 (g/mol) is the naringenin's molecular weight, A_v is the Avogadro's number, $x(T)$ is the number of cells at time T, and $X(k) = \frac{2\ \text{sign}\left((\text{mean}(\sigma_{\text{unp}}(k)) - \sigma_{\text{unp}}(k)) + 1\right))}{2}$ is the clipped binary version of σ factor used to detect its zero-crossing and obtain the number of oscillations of the σ before the malonyl perturbation at the position k. Additionally, the constraints vector $\mathbf{K}(\boldsymbol{\theta})$ set two significant limitations for the antithetic controller performance in this pathway:

$$[\sigma] \geq 4.5\mu M \tag{24}$$

$$\sigma \geq a\sigma, \tag{25}$$

As we said in **Section 2.3**, seven decision variables from the biocontroller and biosensor kinetics in **Eqs. 13**, **14** were selected. Particularly, we considered the ones that are prone to be modified in the wet-lab:

$$\boldsymbol{\theta} = [p_{a\sigma}, C_{N_{a\sigma}}, p_h, C_{N_h}, k_{c\sigma}, k_{d20}, \mu], \tag{26}$$

Table 2 defines the lower and upper limits of the parameters selected for tuning within standard ranges for the chosen biological parts. Therefore, altogether can state the MOP of the naringenin pathway as follows:

$$\begin{aligned} &\min_{\boldsymbol{\theta} \in \Re^7} J(\boldsymbol{\theta}) = [J_1(\boldsymbol{\theta}), J_2(\boldsymbol{\theta}), J_3(\boldsymbol{\theta})] \in \Re^3 \\ &\text{subject to : equations } (11) - (15) \\ &\qquad\qquad \text{constraints } (24) - (25) \end{aligned} \tag{27}$$

TABLE 2 | Lower and upper bounds for the naringenin pathway MOP.

Bound	p_a	C_{Na}	p_h	C_{Nh}	k_c	k_{d20}	μ
	(min^{-1})	(copies)	(min^{-1})	(copies)	(min^{-1})	(molec)	(min^{-1})
Lower	0.1	1	0.1	1	0.01	1e − 2	0.0069
Upper	20	15	20	15	20	1e4	0.0231

For the rest of the parameters, a sensitivity analisys was performed in previous work. Evidently, using different values of the these parameters may have an impact on the resulting behavior of the system, however this does not invalidate the whole methodology we are presenting here.

4.5 Multiobjective Optimization Process

The multiobjective optimization process finds the best parameters $\boldsymbol{\theta_P}^\star$ producing the best Pareto front approximation $\mathbf{J_P}(\boldsymbol{\theta_P})^\star$ for each MOP. Evolutionary algorithms are one of the suggested optimization techniques to address optimization problems generally present in systems and synthetic biology (Moles et al., 2003). We used a multiobjective evolutionary algorithm based on differential evolution, which uses a spherical pruning to approximate the Pareto front. The implementation comes from the sp-MODEx[1] algorithm that improves: 1) convergence by using an external file to store solutions and include them in the evolutionary process, 2) spreading by using the spherical pruning mechanism, and 3) pertinency of solutions *via* a basic bound mechanism in the objective space (Reynoso-Meza et al., 2014).

4.5.1 Multicriteria Decision Making Process

Choosing the preferable solution according to designer's criteria takes place in an *a-posteriori* multicriteria decision making stage of the Pareto front obtained. It is desirable to have tools that simplify the visualization as well as the analysis of the trade-off among competing objectives. This could be a non-trivial task when the number of objectives is larger than three and/or the number of decision variables in the Pareto set is large enough, like in our case. We used the *Level Diagrams Toolbox* (Blasco et al., 2017) from Matlab (LD-Tool[2]) as the Pareto front visualization tool, which is freely available for designers. LD-Tool correlates the design objectives $\mathbf{J_P}(\boldsymbol{\theta})$ with their decision variables $\boldsymbol{\theta}$ by illustrating two graphs. The first graph contains each objective, where its Y-axis is the p-norm $\|J(\boldsymbol{\theta})\|_p$ of the objectives vector, and the X-axis corresponds to each objective value $J_i(\boldsymbol{\theta})$. The second graph shows $\|J(\boldsymbol{\theta})\|_p$ with respect to every $\boldsymbol{\theta}$, so a given solution will have the same y-value in all graphs.

4.6 Computational Simulations

All simulations of both the merging metabolic pathway and the naringenin pathway were performed in Matlab, using a 4 Core processor, 16 GB RAM @ 3.80 GHz. First, we defined two sets of model parameters known as nominal parameters for each system. Then, we computed the number of molecules of each species from every i − cell in the population over time. These data allow us to obtain the performance, robustness and stability of the biosensor and the antithetic controller from each system. Finally, we tuned the biosensor and the biocontroller for optimal dynamic pathway regulation following the multiobjective approach. For the merging metabolic pathway, the sp-MODEx evolutionary algorithm evaluated 1,000 the cost function, using 125 generations and taking 1 h for a simulation time. For the naringenin pathway dynamic regulation, the sp-MODEx evaluated 10,000 times the cost function, using 199 generations over 8 h of simulation time.

All the scripts of the simulations and optimization can be found in the Github repository https://github.com/sb2cl/molecular-biocontroller-tuning.

[1]Available in http://www.mathworks.com/matlabcentral/fileexchange/65145

[2]Available at http://www.mathworks.com/matlabcentral/fileexchange/24042

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed for this study can be found in the Github repository https://github.com/sb2cl/molecular-biocontroller-tuning.

AUTHOR CONTRIBUTIONS

YB JP, and AV conceived the study. YB and AV designed the experiments. YB and FS-N implemented the method and performed the experiments. YB and FS-N worked in the analysis and data visualization. AV and YB wrote the initial version of the manuscript. JP and AV provided funding and resources. All the authors edited and approved the final manuscript.

FUNDING

This research was funded by MCIN/AEI/10.13039/501100011033 grant number PID 2020-117271RB-C21.

ACKNOWLEDGMENTS

YB thanks to Universitat Politècnica de València (Grant PAID-10-21 Acceso al Sistema Español de Ciencia e Innovación), and Secretaría de Educación Superior, Ciencia, Tecnología e Innovación-Ecuador (Scholarship Convocatoria Abierta 2011). FS-N is grateful to grant PAID-01-2017 from Universitat Politècnica de València.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.801032/full#supplementary-material

REFERENCES

Aoki, S. K., Lillacci, G., Gupta, A., Baumschlager, A., Schweingruber, D., and Khammash, M. (2019). A Universal Biomolecular Integral Feedback Controller for Robust Perfect Adaptation. *Nature* 570, 533–537. doi:10.1038/s41586-019-1321-1

Blair, R. H., Kliebenstein, D. J., and Churchill, G. A. (2012). What Can Causal Networks Tell Us about Metabolic Pathways? *Plos Comput. Biol.* 8, e1002458. doi:10.1371/journal.pcbi.1002458

Blasco, X., Herrero, J. M., Reynoso-Meza, G., and Martínez Iranzo, M. A. (2017). "Interactive Tool for Analyzing Multiobjective Optimization Results with Level Diagrams," in Proceedings of the Genetic and Evolutionary Computation Conference Companion, Berlin, Germany, July 15–19, 2017, 1689–1696. doi:10.1145/3067695.3082553

Boada, Y., Picó, J., and Vignoni, A. (2021). "Multi-objective Optimization Tuning Framework for Kinetic Parameter Selection and Estimation," in *Methods in Molecular Biology* (New York, NY: Springer US), 2385. doi:10.1007/978-1-0716-1767-0)

Boada, Y., Reynoso-Meza, G., Picó, J., and Vignoni, A. (2016). Multi-objective Optimization Framework to Obtain Model-Based Guidelines for Tuning Biological Synthetic Devices: An Adaptive Network Case. *BMC Syst. Biol.* 10, 27. doi:10.1186/s12918-016-0269-0

Boada, Y., Vignoni, A., Alarcon-Ruiz, I., Andreu-Vilarroig, C., Monfort-Llorens, R., Requena, A., et al. (2019a). Characterization of Gene Circuit Parts Based on Multiobjective Optimization by Using Standard Calibrated Measurements. *ChemBioChem* 20, 2653–2665. doi:10.1002/cbic.201900272

Boada, Y., Vignoni, A., Picó, J., and Carbonell, P. (2020). Extended Metabolic Biosensor Design for Dynamic Pathway Regulation of Cell Factories. *iScience* 23, 101305. doi:10.1016/j.isci.2020.101305

Boada, Y., Vignoni, A., and Picó, J. (2017a). Engineered Control of Genetic Variability Reveals Interplay Among Quorum Sensing, Feedback Regulation, and Biochemical Noise. *ACS Synth. Biol.* 6, 1903–1912. doi:10.1021/acssynbio.7b00087

Boada, Y., Vignoni, A., and Picó, J. (2017b). Multi-Objective Optimization for Gene Expression Noise Reduction in a Synthetic Gene Circuit. *IFAC-PapersOnLine* 50, 4472–4477. doi:10.1016/j.ifacol.2017.08.376

Boada, Y., Vignoni, A., and Pico, J. (2019b). Multiobjective Identification of a Feedback Synthetic Gene Circuit. *IEEE Trans. Contr. Syst. Technol.* 28, 208–223. doi:10.1109/TCST.2018.2885694

Briat, C., Gupta, A., and Khammash, M. (2016). Antithetic Integral Feedback Ensures Robust Perfect Adaptation in Noisy Biomolecular Networks. *Cel Syst.* 2, 15–26. doi:10.1016/j.cels.2016.01.004

[Dataset] Otero-Muras, I., and Carbonell, P. (2021). Automated Engineering of Synthetic Metabolic Pathways for Efficient Biomanufacturing. *Metab. Eng.* 63, 61–80. doi:10.1016/j.ymben.2020.11.012

Dunstan, M. S., Robinson, C. J., Jervis, A. J., Yan, C., Carbonell, P., Hollywood, K. A., et al. (2020). Engineering *Escherichia coli* towards De Novo Production of Gatekeeper (2S)-Flavanones: Naringenin, Pinocembrin, Eriodictyol and Homoeriodictyol. *Synth. Biol. (Oxf)* 5, ysaa012. doi:10.1093/synbio/ysaa012

Gao, C., Xu, P., Ye, C., Chen, X., and Liu, L. (2019). Genetic Circuit-Assisted Smart Microbial Engineering. *Trends Microbiol.* 27, 1011–1024. doi:10.1016/j.tim.2019.07.005

Hartline, C. J., Schmitz, A. C., Han, Y., and Zhang, F. (2020). Dynamic Control in Metabolic Engineering: Theories, Tools, and Applications. *Metab. Eng.* 63, 126–140. doi:10.1016/j.ymben.2020.08.015

Johnson, A. O., Gonzalez-Villanueva, M., Wong, L., Steinbüchel, A., Tee, K. L., Xu, P., et al. (2017). Design and Application of Genetically-Encoded Malonyl-CoA Biosensors for Metabolic Engineering of Microbial Cell Factories. *Metab. Eng.* 44, 253–264. doi:10.1016/j.ymben.2017.10.011

Li, G., Deng, L., Xiao, G., Tang, P., Wen, C., Hu, W., et al. (2018). Enabling Controlling Complex Networks with Local Topological Information. *Sci. Rep.* 8, 4593. doi:10.1038/s41598-018-22655-5

Liu, D., Mannan, A. A., Han, Y., Oyarzún, D. A., and Zhang, F. (2018). Dynamic Metabolic Control: Towards Precision Engineering of Metabolism. *J. Ind. Microbiol. Biotechnol.* 45, 535–543. doi:10.1007/s10295-018-2013-9

Liu, D., and Zhang, F. (2018). Metabolic Feedback Circuits Provide Rapid Control of Metabolite Dynamics. *ACS Synth. Biol.* 7, 347–356. doi:10.1021/acssynbio.7b00342

Liu, L., Shan, S., Zhang, K., Ning, Z.-Q., Lu, X.-P., and Cheng, Y.-Y. (2008). Naringenin and Hesperetin, Two Flavonoids Derived fromCitrus Aurantiumup-Regulate Transcription of Adiponectin. *Phytother. Res.* 22, 1400–1403. doi:10.1002/ptr.2504

Lo-Thong, O., Charton, P., Cadet, X. F., Grondin-Perez, B., Saavedra, E., Damour, C., et al. (2020). Identification of Flux Checkpoints in a Metabolic Pathway through White-Box, Grey-Box and Black-Box Modeling Approaches. *Sci. Rep.* 10, 13446. doi:10.1038/s41598-020-70295-5

Michaelis, L., and Menten, M. (1913). Die kinetik der invertinwirkung biochem z. *Biochem. z.* Berlin: Berlin, 49, 333–369. Find this article online.

Miettinen, K. (1999). *Nonlinear Multiobjective Optimization*, 12. Boston: Kluwer Academic Publishers.

Miettinen, K., Ruiz, F., and Wierzbicki, A. P. (2008). "Introduction to Multiobjective Optimization: Interactive Approaches," in *Multiobjective Optimization*. Editors J. Branke, K. Deb, K. Miettinen, and R Słowiński (Berlin, Heidelberg: Springer), 27–57. doi:10.1007/978-3-540-88908-3_2

Moles, C. G., Mendes, P., and Banga, J. R. (2003). Parameter Estimation in Biochemical Pathways: A Comparison of Global Optimization Methods. *Genome Res.* 13, 2467–2474. doi:10.1101/gr.1262503

Otero-Muras, I., and Banga, J. R. (2017). Automated Design Framework for Synthetic Biology Exploiting Pareto Optimality. *ACS Synth. Biol.* 6, 1180–1193. doi:10.1021/acssynbio.6b00306

Oyarzún, D. A., and Stan, G.-B. V. (2013). Synthetic Gene Circuits for Metabolic Control: Design Trade-Offs and Constraints. *J. R. Soc. Interf.* 10, 20120671. doi:10.1098/rsif.2012.0671

Rahigude, A., Bhutada, P., Kaulaskar, S., Aswar, M., and Otari, K. (2012). Participation of Antioxidant and Cholinergic System in Protective Effect of Naringenin against Type-2 Diabetes-Induced Memory Dysfunction in Rats. *Neuroscience* 226, 62–72. doi:10.1016/j.neuroscience.2012.09.026

Reynoso-Meza, G., Garcia-Nieto, S., Sanchis, J., and Blasco, F. X. (2013). Controller Tuning Using Multiobjective Optimization Algorithms: a Global Tuning Framework. *IEEE Trans. Control. Syst. Technol.* 21, 445–458. doi:10.1109/tcst.2012.2185698

Reynoso-Meza, G., Sanchis, J., Blasco, X., and García-Nieto, S. (2014). Physical Programming for Preference Driven Evolutionary Multi-Objective Optimization. *Appl. Soft Comput.* 24, 341–362. doi:10.1016/j.asoc.2014.07.009

Santos-Navarro, F. N., Boada, Y., Vignoni, A., and Picó, J. (2021). Gene Expression Space Shapes the Bioprocess Trade-Offs Among Titer, Yield and Productivity. *Appl. Sci.* 11, 5859. doi:10.3390/app11135859

Schomburg, I., Jeske, L., Ulbrich, M., Placzek, S., Chang, A., and Schomburg, D. (2017). The BRENDA Enzyme Information System-From a Database to an Expert System. *J. Biotechnol.* 261, 194–206. doi:10.1016/j.jbiotec.2017.04.020

Segall-Shapiro, T. H., Sontag, E. D., and Voigt, C. A. (2018). Engineered Promoters Enable Constant Gene Expression at Any Copy Number in Bacteria. *Nat. Biotechnol.* 36, 352–358. doi:10.1038/nbt.4111

Sheng, H., Sun, X., Yan, Y., Yuan, Q., Wang, J., and Shen, X. (2020). Metabolic Engineering of Microorganisms for the Production of Flavonoids. *Front. Bioeng. Biotechnol.* 8, 1–15. doi:10.3389/fbioe.2020.589069

Siedler, S., Stahlhut, S. G., Malla, S., Maury, J., and Neves, A. R. (2014). Novel Biosensors Based on Flavonoid-Responsive Transcriptional Regulators Introduced into escherichia Coli. *Metab. Eng.* 21, 2–8. doi:10.1016/j.ymben.2013.10.011

Stevens, J. T., and Carothers, J. M. (2015). Designing RNA-Based Genetic Control Systems for Efficient Production from Engineered Metabolic Pathways. *ACS Synth. Biol.* 4, 107–115. doi:10.1021/sb400201u

Takamura, Y., and Nomura, G. (1988). Changes in the Intracellular Concentration of Acetyl-CoA and Malonyl-CoA in Relation to the Carbon and Energy Metabolism of escherichia Coli K12. *Microbiology* 134, 2249–2253. doi:10.1099/00221287-134-8-2249

Tsiantis, N., and Banga, J. R. (2020). Using Optimal Control to Understand Complex Metabolic Pathways. *BMC Bioinformatics* 21, 472. doi:10.1186/s12859-020-03808-8

Wehrs, M., Tanjore, D., Eng, T., Lievense, J., Pray, T. R., and Mukhopadhyay, A. (2019). Engineering Robust Production Microbes for Large-Scale Cultivation. *Trends Microbiol.* 27, 524–537. doi:10.1016/j.tim.2019.01.006

Wu, J., Du, G., Chen, J., and Zhou, J. (2015). Enhancing Flavonoid Production by Systematically Tuning the Central Metabolic Pathways Based on a CRISPR

Interference System in escherichia Coli. *Sci. Rep.* 5, 13477. doi:10.1038/srep13477

Xu, P., Li, L., Zhang, F., Stephanopoulos, G., and Koffas, M. (2014). Improving Fatty Acids Production by Engineering Dynamic Pathway Regulation and Metabolic Control. *Proc. Natl. Acad. Sci.* 111, 11299–11304. doi:10.1073/pnas.1406401111

Yang, L., Ebrahim, A., Lloyd, C. J., Saunders, M. A., and Palsson, B. O. (2019). DynamicME: Dynamic Simulation and Refinement of Integrated Models of Metabolism and Protein Expression. *BMC Syst. Biol.* 13, 2. doi:10.1186/s12918-018-0675-6

Yang, Y., Lin, Y., Li, L., Linhardt, R. J., and Yan, Y. (2015). Regulating Malonyl-CoA Metabolism via Synthetic Antisense RNAs for Enhanced Biosynthesis of Natural Products. *Metab. Eng.* 29, 217–226. doi:10.1016/j.ymben.2015.03.018

Zygmunt, K., Faubert, B., MacNeil, J., and Tsiani, E. (2010). Naringenin, a Citrus Flavonoid, Increases Muscle Cell Glucose Uptake via AMPK. *Biochem. Biophysical Res. Commun.* 398, 178–183. doi:10.1016/j.bbrc.2010.06.048

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Boada, Santos-Navarro, Picó and Vignoni. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 06 April 2022
doi: 10.3389/fmolb.2022.836794

Optimizing Dosage-Specific Treatments in a Multi-Scale Model of a Tumor Growth

Miguel Ponce-de-Leon[1], Arnau Montagud[1], Charilaos Akasiadis[2], Janina Schreiber[1], Thaleia Ntiniakou[1] and Alfonso Valencia[1,3]*

[1]Barcelona Supercomputing Center (BSC), Barcelona, Spain, [2]Institute of Informatics and Telecommunications, NCSR "Demokritos", Agia Paraskevi, Greece, [3]ICREA, Pg. Lluís Companys, Barcelona, Spain

The emergence of cell resistance in cancer treatment is a complex phenomenon that emerges from the interplay of processes that occur at different scales. For instance, molecular mechanisms and population-level dynamics such as competition and cell–cell variability have been described as playing a key role in the emergence and evolution of cell resistances. Multi-scale models are a useful tool for studying biology at very different times and spatial scales, as they can integrate different processes occurring at the molecular, cellular, and intercellular levels. In the present work, we use an extended hybrid multi-scale model of 3T3 fibroblast spheroid to perform a deep exploration of the parameter space of effective treatment strategies based on TNF pulses. To explore the parameter space of effective treatments in different scenarios and conditions, we have developed an HPC-optimized model exploration workflow based on EMEWS. We first studied the effect of the cells' spatial distribution in the values of the treatment parameters by optimizing the supply strategies in 2D monolayers and 3D spheroids of different sizes. We later study the robustness of the effective treatments when heterogeneous populations of cells are considered. We found that our model exploration workflow can find effective treatments in all the studied conditions. Our results show that cells' spatial geometry and population variability should be considered when optimizing treatment strategies in order to find robust parameter sets.

Keywords: multi-scale modeling, model exploration, treatment optimization, TNF, cell resistance, multi-scale modeling and simulation, agent-based model, optimization via simulation

OPEN ACCESS

Edited by:
Ernesto Perez-Rueda, Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Paola Lecca, Free University of Bozen-Bolzano, Italy
Vinicius Contessoto, Rice University, United States

****Correspondence:***
Miguel Ponce-de-Leon
miguel.ponce@bsc.es

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *15 December 2021*
Accepted: *07 March 2022*
Published: *06 April 2022*

Citation:
Ponce-de-Leon M, Montagud A, Akasiadis C, Schreiber J, Ntiniakou T and Valencia A (2022) Optimizing Dosage-Specific Treatments in a Multi-Scale Model of a Tumor Growth. Front. Mol. Biosci. 9:836794. doi: 10.3389/fmolb.2022.836794

1 INTRODUCTION

Optimizing drug treatment and efficiently screening the effect of drugs is key to improving clinical treatments and ultimately extending patients' life expectancy (Kessler et al., 2014). The emergence of resistant cancer cells is a complex phenomenon due to the inherent complexity of biological (Shaffer et al., 2017), the interplay of processes that occur at different scales, and an environment with an active role in this resistance (Lee et al., 2012; Goldman et al., 2015). Molecular mechanisms and population-level dynamics such as competition and cell–cell variability have been described as playing a key role in the emergence and evolution of cell resistances (Kim et al., 2018). For instance, high gene expression variability has been linked to aggressiveness in chronic lymphocytic leukemia (Ecker et al., 2015). Genetic heterogeneity and phenotype variability have also been related to the emergence of cell resistance (McGranahan and Swanton, 2015; Brady et al., 2017; Shaffer et al., 2017).

Furthermore, the environment has been described to have an effect on the cells' response to drugs: 2D-cultured cell line screens failed in clinical studies (Horvath et al., 2016) as cell cultures do rarely recapitulate the heterogeneity and drug sensitivity of the original tumor (Jabs et al., 2017).

Multi-scale models (MSM) are a useful tool for studying biology at very different time (no s) and spatial scales, as they can integrate different processes occurring at the molecular, cellular, and intercellular levels (Metzcar et al., 2019; Montagud et al., 2021). In the domain of cancer biology, MSMs have been used to connect cellular mechanisms underlying cancer drug resistance to population-level patient survival (Sun et al., 2016), study the role of physiologic resistance due to diffusion gradients of different nutrients and drugs (Frieboes et al., 2009), and quantitatively characterize pressure for invasion (Anderson et al., 2006), among many other applications (Metzcar et al., 2019). In general, multi-scale models provide a genotype-to-phenotype simulation framework, which is ideal for the study of *in silico* drug screenings (Flobak et al., 2015), the optimization of treatment regimens (Akasiadis et al., 2021), and the exploration of genetic or environmental perturbations (Letort et al., 2018).

Multi-scale simulation can be used to conduct *in silico* experiments and generate new experimentally testable hypotheses, accelerating the discovery of new potential treatment strategies (An, 2010). Nevertheless, due to the hybrid approaches used to describe multi-scale models (e.g., discrete, continuous, and stochastic), these models cannot be studied using formal analytical tools, and thus the analysis and exploration of simulated trajectories require complex workflows to guide the exploration of the parameter spaces associate with these models (Ozik et al., 2016). For this reason, distributed workflows to perform parallel optimization *via* simulation and model exploration are critical tools for exploiting the full potential of simulations (Ozik et al., 2018b; Reuillon et al., 2013). Model exploration workflows are required to efficiently fit parameters for which there are no available experimental measurements (Akasiadis et al., 2021; Ozik et al., 2019), explore complex and vast parameter spaces, and optimize user desirable goals, such as the space of optimal treatment strategies for a given cancer model. Optimization methods such as evolutionary algorithms have proven their usefulness in such studies for fitting unknown parameters (Akasiadis et al., 2021), as well as high-throughput hypotheses testing in cancer research (Ozik et al., 2018a).

In previous work, Letort et al. (2018) developed the multi-scale model of 3T3 fibroblast spheroids that integrates the Cell Fate Boolean network Calzone et al. (2010) inside individual cell agents. The Boolean network rules the phenotype of the cells (e.g., proliferation and apoptosis) based on the environmental conditions (e.g., drugs presence and oxygen concentration). The authors used the model to investigate the tumor response to different regimes with tumor necrosis factor supplies (TNF) and reported complex behaviors in the simulated conditions. While a set of values of pulse period, pulse duration, and TNF concentration was optimal to reduce the number of alive tumor cells, different sets of values turned the cells resistant to TNF (Letort et al., 2018). The effects of TNF in the Boolean model reported by Calzone et al. (2010) are multifaceted: TNF triggers cells to go from a naive to a proliferative state and commits cells to necrosis and apoptosis. Once the cells are committed to either survival, necrosis, or apoptosis, they cannot go back, causing resistance due to phenotypic variability. Interestingly, it has been described that prolonged TNF exposure causes the cells to be resistant to the effect of the cytokine (Lee et al., 2016).

In the present work, we use an extended hybrid multi-scale model to perform a deep exploration of the parameter space of effective treatment strategies based on TNF pulses to unravel the mechanistic details behind the complex emergent dynamics of the TNF pulses in *in silico* experiments and guide the optimization of effective treatments. We extended the multi-scale model of 3T3 fibroblast spheroid by integrating an explicit kinetic description of the TNF-receptor dynamics based on the molecular biology of the TNF receptor (Fischer et al., 2011; Li et al., 2013; Sedger and McDermott, 2014). Furthermore, we couple the TNF-receptor kinetic model with the cancer cell Boolean model from Calzone et al. (2010) to simulate the downstream propagation of the signal that induced the binding of the TNF. To explore the parameter space of effective treatments in different scenarios and conditions, we have developed an HPC-optimized model exploration workflow based on EMEWS (Ozik et al., 2018b). Our workflow includes two previously used model exploration strategies, sweep search and genetic algorithm (Akasiadis et al., 2021; Ozik et al., 2019), together with a new approach named the Covariance Matrix Adaptation Evolutionary Strategy, which exhibited good convergence in global optimization problems with continuous variables (Hansen and Ostermeier, 2001).

We applied our framework to characterize the space of effective treatments in different experimental scenarios by simulating the treatment outcome with our multi-scale model of tumor growth. We first studied the effect of the cells' spatial distribution in the values of the treatment parameters by optimizing the supply strategies in 2D monolayers and 3D spheroids of different sizes. We found that our model exploration workflow can find non-trivial *in silico* drug scheduling strategies that minimize the tumor below 1% of its initial size while avoiding the emergence of resistant cells. Our results also show that effective treatment strategies can be found in the two different cell geometries studied. We also found that the parameter spaces of effective treatments for the 2D monolayer and 3D spheroid exhibit different distributions for the parameters. We later study the robustness of the effective treatments when heterogeneous populations of cells are considered. Specifically, we model population heterogeneity by introducing different levels of cell-based variability into the kinetic parameters of the TNF-receptor models. The parameters' variability aims to mimic population-level variability in the kinetic parameters of the receptor, as well as different levels of expression in the receptor, among different cells. We found that effective treatment strategies are robust to a low level of variability, whereas, with a high level of variability, those treatment strategies optimized for populations with no variability cannot reduce tumor growth. However, when the treatments are optimized directly on a heterogeneous

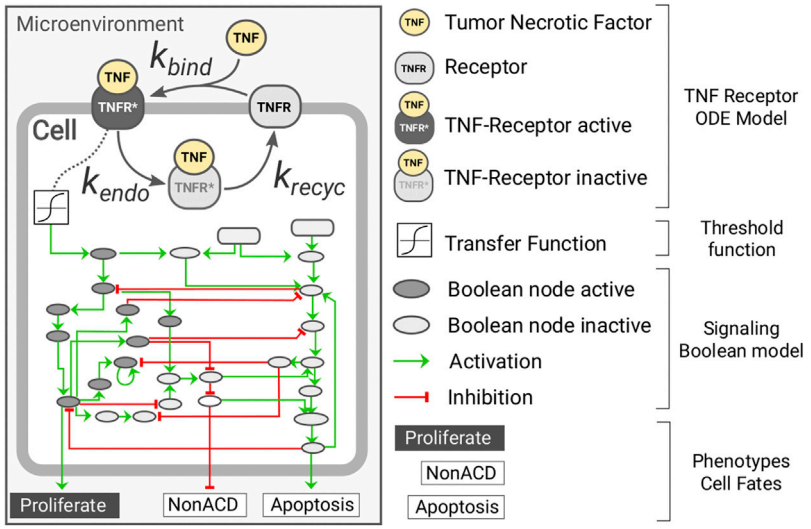


FIGURE 1 | Diagram representing the intracellular submodels of the multi-scale model of tumor growth. Each individual cell agent has a kinetic model of the TNF receptor dynamics connected to the microenvironment through the presence of surrounding TNF and coupled to the Boolean network through a transfer function. The Boolean network has three readout nodes (proliferation, NonACD, and apoptosis), which rule the fate of the cell agent.

population, we observe that the optimization algorithms can retrieve effective treatment.

Altogether, we found that our model exploration workflow can find effective treatments in all the studied conditions, showing that multi-scale simulations and model exploration are promising tools for *in silico* exploring treatment strategies. Finally, our results also show that cells' spatial geometry and population variability should be considered when optimizing treatment strategies to find robust parameter sets. In future work, we plan to extend these results by studying other experimental setups and different cancer models.

2 MATERIALS AND METHODS

2.1 Hybrid Multi-Scale Model of Cancer Cells With Signaling

Herein, we present a multi-scale model of tumor growth that considers, at the individual cell level, the dynamics of the tumor necrosis factor (TNF) receptor and its downstream effect using a hybrid approach (**Figure 1**). Our model was implemented using PhysiCell (Ghaffarizadeh et al., 2018) together with the PhysiBoSS add-on (Ponce-de-Leon et al., in preparation). The microenvironment is simulated in both the 2D and 3D domains, and it accounts for the presence of oxygen and the cytokine tumor necrosis factor (TNF). On the contrary, cells are simulated as individual agents, including intracellular submodels that account for the cell cycle, the different death models (i.e., necrosis and apoptosis), a model for TNF receptor dynamics, and a gene regulatory network.

For the cell cycle, we use PhysiCell live cell cycle with a doubling time of $22h$, and for the death models, we used PhysiCell standard ones with default parameters. The binding of the TNF to its receptor is modeled using mass-action kinetics in which TNF binds to a cell receptor TNFR at a given rate k_{bind}; the complex TNF-TNFR is internalized at a rate of k_{endo} where the TNF is degraded and the receptor recycled at a rate of $k_{recycle}$ (see **Supplementary Figure S1**). The TNF-receptor submodel was developed based on the known molecular biology of the molecular system (Fischer et al., 2011; Li et al., 2013; Sedger and McDermott, 2014). The equation below describes the submodel for the TNF-receptor dynamics:

$$\begin{aligned}\frac{[R_e]}{dt} &= -k_{bind}[R_e][TNF] + k_{recycle}[R_i^*]\\ \frac{[R_e^*]}{dt} &= k_{bind}[R_i][TNF] - k_{endo}[R_e^*]\\ \frac{[R_i^*]}{dt} &= k_{endo}[R_e^*] - k_{recycle}[R_i^*]\end{aligned} \tag{1}$$

where $[R]$, $[TNF]$, and $[R^\star]$ are the concentrations of the receptor, TNF, and TNF-TNFR complex, respectively. Furthermore, the TNF-TNFR complex $[R^\star]$ can be found in two states, in the cell membrane $(R_e^\star)$ or internalized $(R_i^\prime)$.

The gene regulatory model used is an extended version of the Boolean network (BN) reported in Calzone et al. (2010) and is simulated using the MaBoSS algorithm. The BN is coupled to the agent in two different ways (see **Supplementary Figure S2**). The BN has an input node that represents the presence of *TNF* and is coupled to the amount of active TNF-TNFR complex $(R_e^\star)$ through a transfer function that converts the continuous value of $R_e^\star$ into a Boolean one. Additionally, the BN has three mutually exclusive output nodes representing three alternative cell fates: proliferation, apoptosis, and NonAD (non-apoptotic death or necrosis). The fate or phenotype of each cell agent is ruled by the current state of the fate nodes of its internal

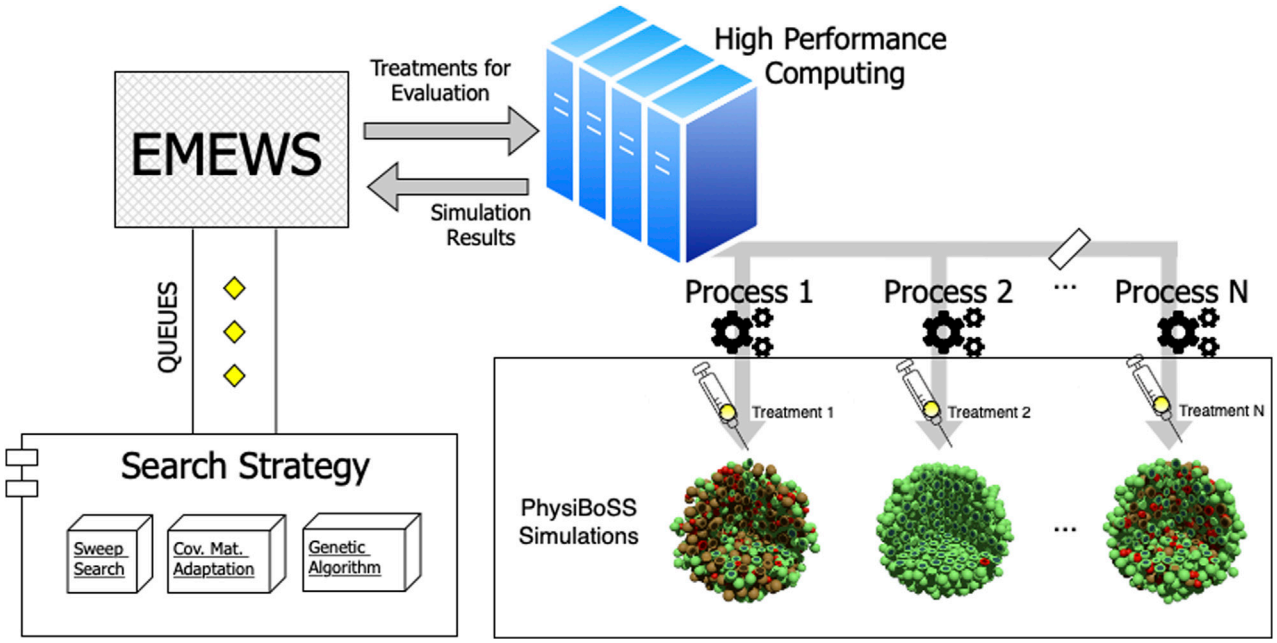


FIGURE 2 | Workflow overview. The diagram depicts the structure of the model exploration workflow. EMEWS communicates to the different search strategies using a queue system. The search strategy generates candidate parameters, and the treatments to be evaluated *via* PhysiBoSS simulations are distributed as parallel jobs to the HPC infrastructure. Upon completion, the simulation outcomes are returned to EMEWS, which, in the case of the GA and CMA-ES, sends the fitness of the evaluated parameters, so the algorithm can update its internal state and generate new candidate parameters.

regulatory network. For instance, if the proliferation node is active, the cell will grow and divide, whereas if the apoptosis becomes active, the cell agent will commit to apoptosis (see **Supplementary Figure S2**).

2.2 Model Exploration Framework

2.2.1 Workflow Overview

The parallel simulation framework used in our evaluation is a workflow that follows the Extreme-scale Model Exploration with Swift (EMEWS) paradigm. It uses the *spheroid_TNF_v2* as an example model and is publicly available in our online repository[1]. An overview of our model exploration workflow is shown in **Figure 2**. We have integrated three different search strategies: 1) a sweep search approach that evaluates a predetermined set of candidate parameters (generated from uniform sampling, or a regular grid), 2) a Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES), and 3) a genetic algorithm (GA). For the cases of CMA-ES and GA, we use the available implementations provided by the DEAP package (version 1.3.1), a mature and widely used package for evolutionary optimization (Fortin et al., 2012). Using EMEWS queues, multi-scale simulation instances are configured with the specific parameter values corresponding to the points that each exploration procedure targets and are then submitted for parallel execution in an HPC environment.

The number of each "batch" of points is relative to the number of computational nodes available in the HPC. The multi-scale simulator incorporates PhysiCell (v1.7) (Ghaffarizadeh et al., 2018) together with the PhysiBoSS 2.0 extension (Ponce-de-Leon et al., in preparation), which is an add-on version of PhysiBoSS. We merged our mass-action kinetics model, as explained in **Section 2.1**, with the multi-scale model proposed by Letort et al. (2018) that used the Boolean model from Calzone et al. (2010). Finally, the simulation results are returned, the points are evaluated according to the performance of the particular drug treatment, and the workflow iterates over the next "batch" of points. Each point is a three-dimensional vector that configures the following simulation parameters: 1) the duration of the TNF pulse; 2) the TNF pulse period; and 3) the concentration of TNF. The exploration space ranges from 5 to 800 min for the pulse period, from 5 to 200 min for the pulse duration, and from 0.001 to 1 ng/L for the TNF concentration.

We checked the number of alive tumor cells at the last time point of each simulation to evaluate the results of each particular treatment and used these values as the fitness or objective of the optimization algorithms. Note that, to ensure that the characterization is robust and not a subject of extreme randomness accruing from inherent PhysiBoSS stochasticity, we perform three replicate simulations with the same configuration parameters, using a different seed to initialize the random number generator, and calculate the average value of the final alive tumor cells count over the replicates as the final score. We now proceed to describe the different search methods we use in detail.

[1]https://github.com/bsc-life/spheroid-tnf-v2-emews

2.2.2 Sweep Search

The sweep search comprises a simple exhaustive approach that requires the user to specify a predetermined number of points to be evaluated. Our code offers a points generating script, which can be configured to choose among different distributions. In other words, points can either be selected to belong on a grid, with equal distances between each point along the dimensions, or a second option is to select random points by sampling particular probability distributions. For the purposes of the experimentation presented in this study, we have implemented the uniform distribution point selection, though this can be easily configured to use other types, such as Gaussian and Beta.

2.2.3 Covariance Matrix Adaptation Evolutionary Strategy

Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) is a stochastic, derivative-free method for numerical optimization for black-box optimization functions (Hansen and Ostermeier, 2001). This method requires, as input, a set of points, σ value that controls the range of exploration, covariance matrix C used to guide the search, the number of points population to execute the algorithm upon, and a total number of iteration or stop criteria. In a nutshell, CMA-ES generates an initial population sampling from a multivariate normal distribution, evaluates each generated point, and then calculates mutation steps of the best points to form the mutation distribution. By this, a new population of points is generated and evaluated, and the iterative process continues up to a user-defined number of times. Note that, for every update of the mutation distribution in each algorithm iteration, all past paths from previous iterations are also considered, and the most favorable points are granted a larger probability for being selected by the evolutionary strategy. This way, the length of each mutation step can be adapted to be longer in cases of greater fitness score improvement or shorter for the opposite case.

2.2.4 Genetic Algorithm

Genetic algorithm (GA) is a widely known and tested metaheuristic approach, also belonging to the family of evolutionary strategies algorithms, which mimics the evolution principles of biological organisms and operates directly on the values of points (Holland, 1975; Whitley, 1994). Similar to CMA-ES, an initial population of points is generated and evaluated, and then, following an iterative approach, a series of genetic operators are applied to each of them in order to produce the next evolved population of points. More specifically, the GA applies the selection, crossover, and mutation operators. Typically, the first operator selects the evaluated individuals in a weighted manner so that the ones with better fitness scores have an increased probability of being selected to proceed to the next generation, compared to fewer fit points. Then, the crossover mixes the point values in a principle similar to that of the gene propagation from parents to offspring as it happens in organisms. Finally, the mutation operator changes a point value (e.g., one of its dimensions) with a small probability, similarly to the process that has been observed in DNA sequences. The main idea is that combinations of points with good fitness scores would lead to even better ones, especially if the search domain is smooth. However, because the algorithm considers only the previously observed fitness scores, without having any other domain-specific knowledge, as a consequence, its search may be constrained around locally optimal points, never managing to reach the global optimal ones. Despite these shortcomings, GAs have been shown to work very well for non-smooth search spaces (Fitzpatrick and Grefenstette, 1988; Tang et al., 1996).

3 RESULTS

3.1 Multi-Scale Simulations and Model Exploration Setup

Herein, we use a multi-scale model of tumor growth to investigate different treatment strategies. The model, which is also used in Akasiadis et al. (2021), simulates the dynamics of a population of cancer cells growing under different drug treatment conditions. A treatment strategy consists of the supply of periodic pulses of the cytokine tumor necrosis factor (TNF) with fixed duration and concentration (see **Section 2**). At the molecular level, when the TNF binds to the cell's receptor TNFR forming a complex, and the TNF-TNFR complex concentration reaches a given threshold, the signal is propagated through the Boolean regulatory network, inducing cell death. However, if the stimulus is sustained for a longer period of time, cells activate the NFkB node and the survival node, becoming resistant to the death induced by the TNF. For this reason, optimal treatments should expose the cell for a sufficient time to induce death but not too much as to become resistant to it (Letort et al., 2018).

To explore the parameter space associated with the treatment, we have extended our model exploration workflow based on EMEWS (Akasiadis et al., 2021). In each *in silico* experiment, we simulate the growth of a population of cancer cells for 4,640 min (i.e., three days) subject to a given treatment strategy. In order to account for the inherent stochasticity of the model, each simulation is always run in three replicates and the average behavior is considered (see **Section 2**). We evaluate the effect of the treatment strategies by analyzing the total number of alive cells at the end of the simulations relative to the initial population size and use these values as the score or cost function associated with a treatment strategy. We define, as effective treatments, those strategies that reduce the number of the alive cell below 1% of its initial numbers in the three replicates. Based on this definition, we investigate the parameter space of the effective treatments in two different spatial arrangements of cells: a monolayer disc of radius 100 μm (151 cells) and a 3D spheroid of radius 100 μm (1,173 cells).

3.2 Effective Treatment Parameters Differs for 2D and 3D Cell Arrangements

To investigate the structure of the parameter space of the effective treatments, we perform a uniform sampling of 10,000 candidate sets of parameters corresponding to different treatment strategies. We use these sets of

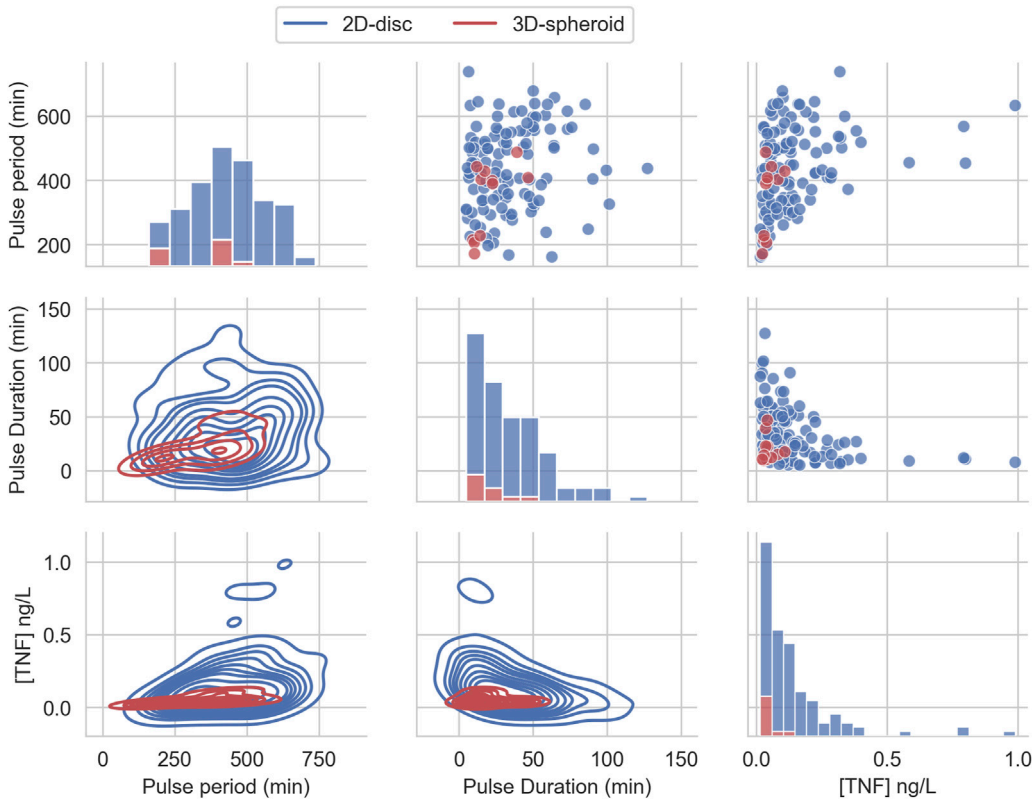


FIGURE 3 | Effective TNF treatment parameters distribution from uniform random sampling. The distribution of the three parameters of the pulse treatment, sampled from a uniform distribution and filtered to belong to the feasible region that can reduce the tumor size below 1% of its initial size.

TABLE 1 | Summary statistics for the parameters from the effective treatments found by sampling 10,000 random candidates.

Parameter	Layout	Mean	Std.	Min	Median	Max
Pulse duration (min)	2D	34.66	24.54	5.00	30.17	127.21
	3D	19.97	12.31	9.37	14.99	46.66
Pulse period (min)	2D	440.83	128.53	161.44	438.86	738.09
	3D	343.28	113.64	171.28	398.79	487.19
TNF (ng/L)	2D	0.14	0.16	0.02	0.10	0.99
	3D	0.05	0.03	0.02	0.04	0.11

parameters (sweep search) as inputs for multi-scale simulations and evaluate each treatment effect on the growth of the cancer cells in the 2D and 3D arrangements. From the 10,000 evaluated parameter sets in each condition, the results show that 113 strategies are effective treatments for the 2D setup, whereas, in the 3D case, only 11 strategies are effective treatments (see **Supplementary Table S1**). This indicates that the region containing the effective treatments for the 3D spheroid is more constrained than in the 2D disc arrangement.

Interestingly, if we restrict the definition of effective treatments to kill all cancer cells at the end of the simulation and in the three replicates, only eight sets of parameters can reach the goal for the 2D cases, whereas no effective parameter sets are found for the 3D setup (see **Supplementary Table S1** and **Supplementary Figure S3**).

We compared the distributions and summary statistics of the parameters of the effective treatments in the two arrangements (**Figure 3**). The comparison shows that the distribution values for the evaluated parameters indicate that the effective treatments of the 3D arrangement are notably more constrained than those that work in the 2D arrangement, in particular regarding the concentration of TNF and the pulse duration. In general, the effective treatments in 2D arrangements exhibit bigger values and larger ranges in the three parameters (**Table 1**).

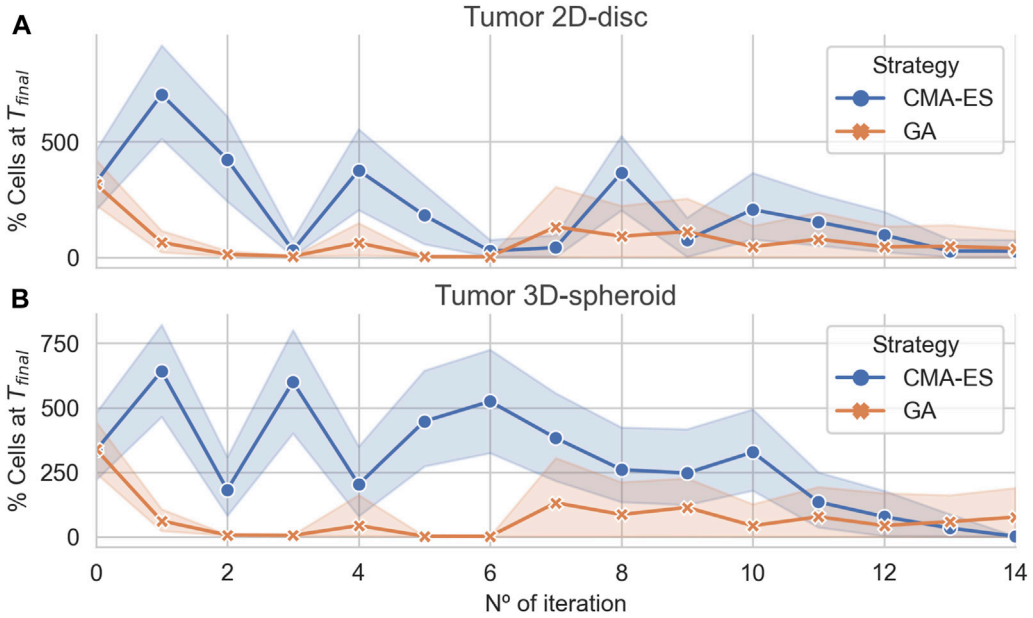


FIGURE 4 | Algorithmic convergence for the optimization of treatment parameters. The average % of alive cells at the end of each iteration step of the CMA-ES and GA optimization algorithms. **Panels (A) and (B)** show the convergence of the algorithms for simulations considering a population of cancer cells arranged in a 2D disc and a 3D spheroid, respectively.

We also found that some parameters' combinations exhibit correlations (see **Supplementary Figure S4**). On the one hand, we observe that the Pulse period positively correlates with TNF in the 2D and 3D but only correlates with pulse duration in the 2D case. On the other hand, TNF shows a negative correlation with pulse duration only for the 2D case. Altogether, these correlations indicate that the treatment parameters can compensate each other: a shorter pulse might be as effective as a longer one if it carries more TNF.

Although herein we are considering a spheroid composed only of tumor cells, more complex scenarios will also contain healthy cells. In such cases, effective treatments will also need to consider the cytotoxic effect of the drug on the healthy cells. To address this issue, we have also calculated the total concentration of TNF supplied during each treatment based on the pulse parameters as follows:

$$TNF_{total} = Pulse_{duration} * Pulse_{concentration} * n,$$

where n is the total number of pulses supplied calculated by dividing the total treatment (simulation) duration (min) over the pulse period (min). Using the calculated values, we ranked the feasible solutions to find the ones that minimize the total concentration of TNF used during the whole treatment (see **Supplementary Table S1**). The results show that the distribution of total TNF is biased to lower values of the total TNF supplied in both the 2D and 3D (see **Supplementary Figure S5**). Strikingly, when we analyze the effective treatments that minimize the total amount of TNF supplied in the 2D and 3D, we found a very similar value of 5 *ng/L*.

3.3 Optimal Treatment Parameters Differs for 2D and 3D Cell Arrangements

To further investigate the structure of the parameter space of the effective treatments, we conducted an optimization *via* simulation to find the set of treatment parameters that minimizes the number of alive cells at the end of the application of the treatment. We performed the parameter optimization in both cell arrangements (i.e., 2D disc and 3D spheroid of radius 100 μm), focusing on the same parameters as in **Section 3.2**: pulse duration, pulse period, and TNF. These optimizations were run using two evolutionary algorithms: Covariance Matrix Adaptation Evolutionary Strategy (CMA-ES) and genetic algorithm (GA) (see **Section 2** for details).

The results showed that both algorithms converge to optimal (leaving no remaining cells) or near-optimal sets of parameters for the 2D and 3D case, respectively (**Figure 4**). Interestingly, in both cases, the GA found effective treatments after two iterations, whereas the CMA-ES only does after 15 iterations. Nonetheless, both algorithms are capable of finding effective treatments. In addition, the CMA-ES algorithm also estimated a multivariate normal distribution for the region of effective treatments parameters that is updated at each iteration. In the last iteration, the population sampled by the CMA-ES showed a very low variance (**Figure 4**) in the 2D and the 3D arrangements, indicating that the estimated distribution captures, at least, part of the structure of the parameters associated with the effective treatments.

We compared the parameter sets for the effective treatments predicted by both algorithms and uncovered that each one converges to different regions of the parameter space (see **Supplementary Figure S6**). The CMA-ES found effective treatments parameter distributions different for the 2D and 3D. In both cases, the parameter ranges were narrower than those found in uniform sampling. Moreover, the CMA-ES converged to distributions for the pulse period different for

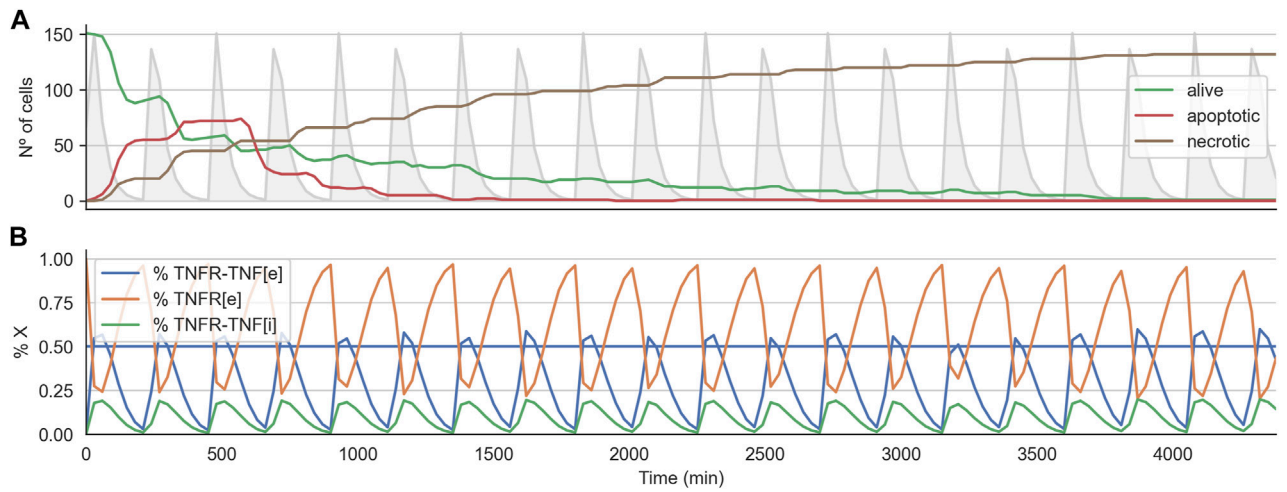


FIGURE 5 | Time course for effective treatment in 2D cell arrangement. **Panel (A)** shows the time course for the number of alive, apoptotic, and necrotic cells. The light grey curve shows the TNF pulses. **Panel (B)** shows the average state of the TNFR receptor model across all the alive cells. The horizontal line indicates the threshold at which the signal induced by the binding of the TNF propagates downstream to the Boolean network.

the 2D and the 3D arrangement. In addition, the parameters corresponding to effective treatments found by the GA were more scattered showing wider ranges of values. In both cases, the pulse duration seemed to be the less critical or constrained parameter.

We analyzed the time course of the total number of alive, apoptotic, and necrotic cells for one of the optimal treatments found. Furthermore, for this simulation, we also checked the internal state of the TNF-receptor model (i.e., the values for R_e, R_e^*, $R_i^!$ variables) at each time step of the simulation and averaged these values over all the cells to investigate the coarse grain dynamics of the submodel. **Figure 5A** shows how the number of alive cells periodically drops until it reaches zero as a consequence of the TNF pulses. When analyzing the average dynamics of the TNFR receptor model in effective treatments, we found that the TNF pulse needs to trigger the activation of the receptor of 50% of the population for short periods of time to be effective (5b). If the average number of cells that got activated is lower than this threshold, then the rate of cells entering necrosis will be lower than the population growth rate, and therefore the number of alive cancer cells will steadily grow. On the contrary, if the average number of activated cells is above this threshold for extended periods of time, many cells will become resistant to the treatment. We found similar results for the case of the 3D spheroid (data not shown).

3.4 Robustness Analysis of the Effective Treatments in Heterogeneous Populations

In the previous section, we showed that effective treatments could be found for 2D and 3D cell arrangements, using either the CMA-ES or the GA algorithms. Nonetheless, those treatments were optimized on monoclonal or homogeneous tumors, that is, the population of cells with identical parameters. In this section, we study the robustness of effective treatments by studying heterogeneous populations of cells. We forced the population heterogeneity by introducing variability into the three kinetic parameters of the TNF receptor, that is, the TNFR binding rate, the TNFR endocytosis rate, and the TNFR recycling rate. The variability is applied by considering a normal distribution centered in each parameter's default value and with a standard deviation, the control parameter that varies from zero (homogeneous population) to one (almost uniformly distributed random parameters). Then, when the population is initialized, the kinetic parameters of each cell are sampled from the corresponding distribution.

To evaluate the robustness of the effective treatments in heterogeneous populations, we considered the top 30 effective treatments parameter sets that had no final tumor cells in any of their replicates for the 2D and for the 3D cell arrangements. Then, for each set of parameters, we run the simulations with different levels of variability from 0 to 1. As expected, we observed that, for low values of the variability control parameter (< 0.2), most of the evaluated effective treatments still can reduce the initial tumor size to the 1% of the initial size. However, for higher variability values (> 0.2), most of the evaluated treatments could not reduce the tumor size below its initial size (**Figure 6**). This indicates that, as the level of variability on the TNFR receptor kinetic parameters increases, the effectiveness of the treatments dramatically decreases (**Figure 6**).

Interestingly, the critical value at which most effective treatments were no longer effective is different for the 2D and the 3D cell arrangements. In the case of the 2D disc, the critical value is close to 0.25, whereas, in the 3D spheroid cases, this value is around 0.15 (**Figure 6**). When the variability value is above this threshold, some of the sampled kinetic parameters make the cell insensitive to the treatment, and thus it can grow even in the presence of TNF. The differences in the critical threshold were possibly due to the different

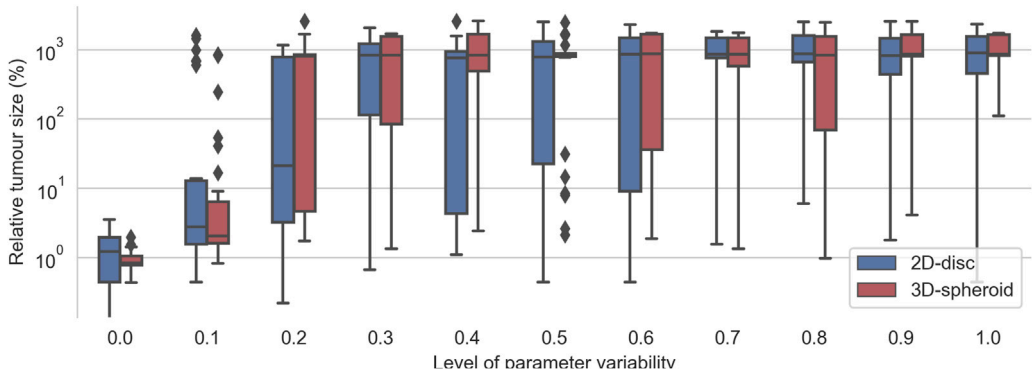


FIGURE 6 | Evaluation of the top 30 best effective treatments in heterogeneous populations of cancer cells with different degrees of variability in the kinetic parameters of the TNFR. For a given value kinetic parameters' variability (*x*-axis), each pair of boxes depicts the distribution % of alive cells at the end of the simulations obtained after evaluating the top 30 best effective treatments founded when zero variability was considered.

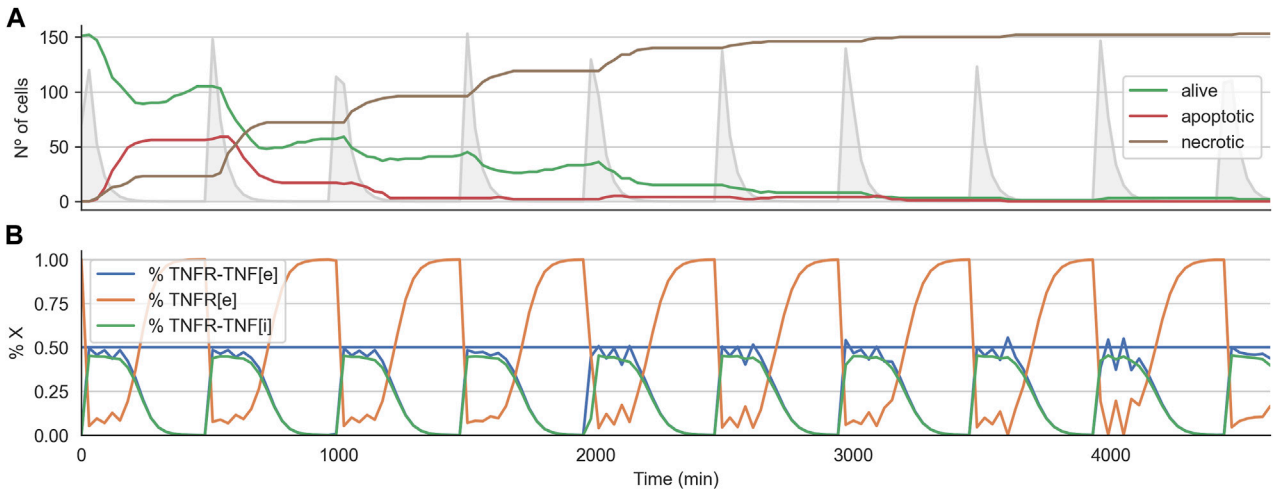


FIGURE 7 | Time course for effective treatment in the 2D cell arrangement with variability of 0.25. **Panel (A)** shows the time course for the number of alive, apoptotic, and necrotic cells. The light grey curve shows the TNF pulses. **Panel (B)** shows the average state of the TNFR receptor model across all the alive cells. The horizontal line indicates the threshold at which the signal induced by the binding of the TNF propagates downstream to the Boolean network.

number of initial cells considered in the 2D and the 3D cell arrangements. To assess this, we tested the effect of the radius size of the 2D simulations in this robustness analysis.

We evaluated radius sizes of 50, 275, and 500 μm corresponding to initial tumor sizes of 37, 1,069, and 3,559 cells, respectively. We tested three levels of variability with each radius size in three replicates and averaged their results. As already discussed, we can observe that the more variability, the worse the outcome (see **Supplementary Figure S7**). Interestingly, we did not see a clear correlation between the radius length and the decrease of the effectiveness of the treatments, with the 50 μm being the one with worse outcomes with the higher variability level.

3.5 Optimization *via* Simulation Can Find Effective Treatments in Heterogeneous Populations of Cancer Cells

To evaluate the performance of model exploration workflow in more complex scenarios, we investigated the optimization of treatment strategies in tumors with different levels of heterogeneity. This use case was considered as a way to

evaluate what would happen when using this methodology in a less ideal situation as it can be the drug screens in cell lines or tumors with heterogeneous non-clonal cells. For this purpose, we conducted the optimization *via* simulation to find effective treatments in two conditions with different degrees of variability in the kinetic parameters.

At first, we set the variability value on the kinetic parameters of the TNFR receptor model to 0.25 and then run the GA and CMA-ES to find treatments that minimize the total number of alive cells. We found that, with this level of variability, neither of the algorithms could converge (see **Supplementary Figure S8**). Nonetheless, this did not prevent the algorithms from finding effective treatments for both arrangements: the 2D disc and the 3D spheroid. While, for the 2D disc, several candidate sets of parameters were found, only a few candidate effective treatments were found by the CMA-ES for the 3D spheroid.

Interestingly, for the 2D cell arrangement, the CMA-ES algorithm was able to find two optimal treatment strategies, that is, a set of parameters for the TNF pulse that kills every tumor cell in all three replicates. These two sets of treatment parameters are similar to those effective treatments that worked when variability was not introduced (see the previous section). **Figure 7A** shows the time course for the effective treatments optimized in the 2D cell arrangement. The plot shows how the number of alive cells steadily decreases until zero (**Figure 7B**). It also shows the average internal dynamic of the receptor model exhibiting some noisy behavior due to the heterogeneity present in the population.

We also performed a similar experiment with a higher variability value (0.50) on the kinetic parameters of the TNFR receptor model. We ran the treatment optimization using the GA and CMA-ES and found that, as expected, with this level of variability, the convergence of both algorithms was even worse than for the case of 0.25 (see **Supplementary Figure S9**). Furthermore, the algorithms could only find effective treatments for the 2D disc arrangement. Nevertheless, the number of different parameter sets found were fewer than in the previous scenario, with a variability value of 0.25, as expected for a more complex landscape. For the 3D spheroid case, the best treatment reduced the initial tumor size to 1.05% of its initial size when averaged over the three replicates. If we relax the definition of effective treatment to a threshold of 2% and compare the distribution of the effective parameters between the cases with variability set to 0.25 and 0.50, we found that the ranges of values were wider in the second case. Altogether the results presented in this section indicate that the higher the variability in the population, the harder to find effective treatments. Nonetheless, the results also show that even with high values of parameters variability, it is still possible to find very effective treatment strategies.

4 DISCUSSION

In this work, we used a hybrid multi-scale model that merges a mass-action kinetics model of the TNF receptor with a cancer cell Boolean model of different signaling pathways. Moreover, these models are embedded in an agent-based framework that allows considering populations of cells in a defined microenvironment. By performing a model exploration, we have shown that the effective treatments parameter can be found in different cells' geometries, including 2D monolayers and 3D spheroids. Furthermore, by performing a uniform random sampling of the effective treatment spaces, we found that the parameters for 2D and 3D arrangements exhibit different distributions. These differences are more pronounced in the case of the TNF concentration and the pulse duration, where the effective treatments for the 3D spheroid case are notably more constrained than those found for the 2D disc. We hypothesize that the 3D configuration imposes spatial constraints in the diffusion of the TNF, which restraint the space of values for the candidate's effective treatment.

We also found that some parameters' combinations exhibit correlations, indicating that one parameter change can be compensated by adjusting another one. For instance, we observe that the pulse period positively correlates with TNF in the 2D and 3D. This means that increasing the period between pulses can be compensated by increasing the pulse concentration. Interestingly, the correlation between these two parameters is stronger in the 3D spheroid than in the 2D monolayer, showing how the former case is more constrained than the latter. For the 3D spheroid, we also found a strong positive correlation between the pulse period and its duration, showing that these two parameters can also compensate for each other. Nonetheless, the correlation between these two parameters is very low and does not show statistical significance in the 2D monolayer. Finally, we found a negative correlation between the pulse concentration and its duration in the 2D monolayer showing that, in these cases, an increase in the concentration of the pulse can be compensated by a reduction of its duration. Altogether these results indicate that the structure of parameters spaces of effective treatment depends on the spatial cell distribution. Therefore, treatment strategies that work in a 2D monolayer may not work in a 3D spheroid.

Although our model only considers tumor cells, the total drug supplied will be critical when healthy cells are also present. For this reason, we analyzed the total TNF supplied during each experiment as a way to estimate the cytotoxicity associated with the effective treatments. Our results showed that, in general, the distribution of this value tends to be skewed to lower values. We also found that the minimum value of the total TNF for the 2D and 3D are quite similar. Nonetheless, the specific treatment's parameters are very different; in the 2D case, the effective treatments that use the minimum TNF values has a period around twice times larger than the one in 3D, but the duration of the pulse has half of the duration; the pulse concentration is around four times higher in the 2D case. This analysis also indicates that spatial cell distribution is important for the design of efficient strategies.

We later investigate treatment optimization using two different evolutionary algorithms: GA and CMA-ES. Our results showed that both algorithms could quickly converge to effective treatments, but while the GA can find candidates in the first iteration, the CMA-ES converge to a more robust region of the parameter space. Furthermore, the CMA-ES also finds a statistical distribution for

the region of effective treatments. Strikingly, both algorithms converge to slightly different regions of the parameter space in both the 2D and the 3D arrangements. This suggests that the fitness landscape of effective treatments is very rough, exhibiting several valleys of effective treatments regions.

In order to unravel the molecular mechanism behaving like the effective treatments, we analyzed the coarse-grained dynamics of the TNF receptor models for working and none working treatment strategies. The non-working strategies can be grouped into two classes. On the one hand, there are those sets of treatment parameters that do not allow the receptor to reach the threshold needed to propagate the signal downstream of the Boolean model. On the other hand, there are treatment strategies that keep the activation of the receptor for long periods enough to induce cell resistance. In this context, if we define drug resistance as the inability of a cell to respond to a given drug, we find that resistance can come from two aspects: from the dynamics of the Boolean model in response to TNF and from the characteristics of the TNF receptors. The effects of TNF in this Boolean model reported by Calzone et al. (2010) are multifaceted: TNF triggers cells to go from a naive to a survival state but to commit cells to necrosis and apoptosis. Once the cells are committed to either survival, necrosis, or apoptosis, they cannot go back, causing a resistance that can be due to phenotypic variability. As this model was studied using a stochastic Boolean simulator (Stoll et al., 2012), it was possible to capture its dynamics and see that these commitments were not equally fast, or even, that there was a window of activation that allowed controlling the commitment to survival and commiting the cells to necrosis (Letort et al., 2018).

In addition to the Boolean model, our hybrid model also has a mass-action kinetics model that can cause another type of resistance. As we see in **Section 3.2** and **Section 3.3**, there are values for the receptor's kinetic parameters that prevent the cell from the regulatory effects produced by the binding of TNF. Therefore, when we consider heterogeneous populations by introducing variability in the kinetic parameters of the TNF receptor, we observed that beyond a critical value of the parameter that controls variability, most of the effective treatments that work in the homogeneous population fail to reduce tumor growth. Our hypothesis is that, with high variability, some of the cells could have kinetic parameters that make them insensitive to the treatment, and thus they will produce the relapse after the sensitive ones have been killed by the treatment. We found that this is the main cause of the non-optimal parameters sets found by the optimization techniques within heterogeneous populations.

Regarding the critical value for the control parameter, it is different for the 2D and the 3D cell arrangements, with a lower value in the latter case. We hypothesize that this difference may be due to two factors. The first is because the space of effective treatments strategies is more constrained in the 3D case. The second reason we propose is due to the differences in the total number of cells simulated in the 2D and 3D. While we set the same radius for the disc and the spheroid, the numbers of initial cells are ~150 and ~1,000, respectively. Because variability is generated by the sampling of random parameters, a larger number of cells increases the probability of getting a set of kinetic parameters that make the cell insensitive to the TNF. We have shown that population variability can cause resistance. The higher the variability, the harder to find effective treatments. However, even in the cases of the maximum variability analyzed, the algorithms can find a few sets of candidate effective treatments.

5 CONCLUSION

Multi-scale modeling allows for gaining mechanistic insights in dynamic drug dosages and predicting novel strategies for treatments. Even though, in the last few years, it has been great progress in the field (Montagud et al., 2021), it is known that virtual drug screens seldom match with clinical trials results. Thus, we need to acknowledge that we are far from using these models at the patient's bedside (Horvath et al., 2016). One of the improvements that would help close this gap would be to have simulations and optimizations that account for and embrace uncertainty. We have hereby presented a free-to-use, open-source framework that allows optimizing treatment strategies with varying levels of uncertainty. We tested the framework using a multi-scale model of cancer growth in different cell arrangements introducing population variability to show that population heterogeneity is critical, either caused by the cells' state, their parameters, or the population size, affecting the optimal parameter sets. We found that our model exploration workflow can find effective treatments in all the studied conditions. Most importantly, our results show that cells' spatial geometry and population variability should be considered when optimizing treatment strategies to find robust parameter sets.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/bsc-life/spheroid-tnf-v2-emews.

AUTHOR CONTRIBUTIONS

MP-L and AV conceived the work. MP-L and AM implemented the model. MP-L, CA, TD and JS implemented the workflow and the code for analyzing simulation results. MP-L, CA and AM prepared the manuscript. All authors edited, commented, and agreed on the final version of the manuscript.

FUNDING

This research has received funding from the Horizon 2020 INFORE Project, GA n° 825070 and the Horizon 2020 PerMedCoE Project, GA n° 951773.

ACKNOWLEDGMENTS

The authors acknowledge the technical expertise and assistance provided by the Spanish Supercomputing

Network (Red Española de Supercomputación) and the computer resources used: the LaPalma Supercomputer (BCV-2019-2-0008), at the Instituto de Astrofísica de Canarias and MareNostrum4 (BCV-2020-3-0016), located at the Barcelona Supercomputing Center.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.836794/full#supplementary-material

REFERENCES

Akasiadis, C., Ponce-de-Leon, M., Montagud, A., Michelioudakis, E., Atsidakou, A., Alevizos, E., et al. (2021). Parallel Model Exploration for Tumor Treatment Simulations. *Computational Intelligence*. doi:10.1111/coin.12515

An, G. (2010). Closing the Scientific Loop: Bridging Correlation and Causality in the Petaflop Age. *Sci. Transl. Med.* 2, 41ps34. doi:10.1126/scitranslmed.3000390

Anderson, A. R. A., Weaver, A. M., Cummings, P. T., and Quaranta, V. (2006). Tumor Morphology and Phenotypic Evolution Driven by Selective Pressure from the Microenvironment. *Cell* 127, 905–915. doi:10.1016/j.cell.2006.09.042

Brady, S. W., McQuerry, J. A., Qiao, Y., Piccolo, S. R., Shrestha, G., Jenkins, D. F., et al. (2017). Combating Subclonal Evolution of Resistant Cancer Phenotypes. *Nat. Commun.* 8, 1231. doi:10.1038/s41467-017-01174-3

Calzone, L., Tournier, L., Fourquet, S., Thieffry, D., Zhivotovsky, B., Barillot, E., et al. (2010). Mathematical Modelling of Cell-Fate Decision in Response to Death Receptor Engagement. *Plos Comput. Biol.* 6, e1000702. doi:10.1371/journal.pcbi.1000702

[Dataset] Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. ann arbor: University of michigan press.

Ecker, S., Pancaldi, V., Rico, D., and Valencia, A. (2015). Higher Gene Expression Variability in the More Aggressive Subtype of Chronic Lymphocytic Leukemia. *Genome Med.* 7, 8. doi:10.1186/s13073-014-0125-z

Fischer, R., Maier, O., Naumer, M., Krippner-Heidenreich, A., Scheurich, P., and Pfizenmaier, K. (2011). Ligand-induced Internalization of TNF Receptor 2 Mediated by a Di-leucin Motif Is Dispensable for Activation of the NFκB Pathway. *Cell Signal.* 23, 161–170. doi:10.1016/j.cellsig.2010.08.016

Fitzpatrick, J. M., and Grefenstette, J. J. (1988). Genetic Algorithms in Noisy Environments. *Mach Learn.* 3, 101–120. doi:10.1007/bf00113893

Flobak, Å., Baudot, A., Remy, E., Thommesen, L., Thieffry, D., Kuiper, M., et al. (2015). Discovery of Drug Synergies in Gastric Cancer Cells Predicted by Logical Modeling. *Plos Comput. Biol.* 11, e1004426. doi:10.1371/journal.pcbi.1004426

Fortin, F.-A., De Rainville, F.-M., Gardner, M.-A., Parizeau, M., and Gagné, C. (2012). DEAP: Evolutionary Algorithms Made Easy. *J. Machine Learn. Res.* 13, 2171–2175. doi:10.5555/2503308.2503311

Frieboes, H. B., Edgerton, M. E., Fruehauf, J. P., Rose, F. R. A. J., Worrall, L. K., Gatenby, R. A., et al. (2009). Prediction of Drug Response in Breast Cancer Using Integrative Experimental/Computational Modeling. *Cancer Res.* 69, 4484–4492. doi:10.1158/0008-5472.CAN-08-3740

Ghaffarizadeh, A., Heiland, R., Friedman, S. H., Mumenthaler, S. M., and Macklin, P. (2018). PhysiCell: An Open Source Physics-Based Cell Simulator for 3-D Multicellular Systems. *Plos Comput. Biol.* 14, e1005991. doi:10.1371/journal.pcbi.1005991

Goldman, A., Majumder, B., Dhawan, A., Ravi, S., Goldman, D., Kohandel, M., et al. (2015). Temporally Sequenced Anticancer Drugs Overcome Adaptive Resistance by Targeting a Vulnerable Chemotherapy-Induced Phenotypic Transition. *Nat. Commun.* 6, 6139. doi:10.1038/ncomms7139

Hansen, N., and Ostermeier, A. (2001). Completely Derandomized Self-Adaptation in Evolution Strategies. *Evol. Comput.* 9, 159–195. doi:10.1162/106365601750190398

Horvath, P., Aulner, N., Bickle, M., Davies, A. M., Nery, E. D., Ebner, D., et al. (2016). Screening Out Irrelevant Cell-Based Models of Disease. *Nat. Rev. Drug Discov.* 15, 751–769. doi:10.1038/nrd.2016.175

Jabs, J., Zickgraf, F. M., Park, J., Wagner, S., Jiang, X., Jechow, K., et al. (2017). Screening Drug Effects in Patient-derived Cancer Cells Links Organoid Responses to Genome Alterations. *Mol. Syst. Biol.* 13, 955. doi:10.15252/msb.20177697

Kessler, D. A., Austin, R. H., and Levine, H. (2014). Resistance to Chemotherapy: Patient Variability and Cellular Heterogeneity. *Cancer Res.* 74, 4663–4670. doi:10.1158/0008-5472.CAN-14-0118

Kim, E., Kim, J.-Y., Smith, M. A., Haura, E. B., and Anderson, A. R. A. (2018). Cell Signaling Heterogeneity Is Modulated by Both Cell-Intrinsic and -extrinsic Mechanisms: An Integrated Approach to Understanding Targeted Therapy. *Plos Biol.* 16, e2002930. doi:10.1371/journal.pbio.2002930

Lee, M. J., Ye, A. S., Gardino, A. K., Heijink, A. M., Sorger, P. K., MacBeath, G., et al. (2012). Sequential Application of Anticancer Drugs Enhances Cell Death by Rewiring Apoptotic Signaling Networks. *Cell* 149, 780–794. doi:10.1016/j.cell.2012.03.031

Lee, R. E. C., Qasaimeh, M. A., Xia, X., Juncker, D., and Gaudet, S. (2016). NF-κB Signalling and Cell Fate Decisions in Response to a Short Pulse of Tumour Necrosis Factor. *Sci. Rep.* 6, 39519. doi:10.1038/srep39519

Letort, G., Montagud, A., Stoll, G., Heiland, R., Barillot, E., Macklin, P., et al. (2018). PhysiBoSS: A Multi-Scale Agent-Based Modelling Framework Integrating Physical Dimension and Cell Signalling. *Bioinformatics* 35, 1188–1196. doi:10.1093/bioinformatics/bty766

Li, J., Yin, Q., and Wu, H. (2013). Structural Basis of Signal Transduction in the TNF Receptor Superfamily. *Adv. Immunol.* 119, 135–153. doi:10.1016/B978-0-12-407707-2.00005-9

McGranahan, N., and Swanton, C. (2015). Biological and Therapeutic Impact of Intratumor Heterogeneity in Cancer Evolution. *Cancer Cell* 27, 15–26. doi:10.1016/j.ccell.2014.12.001

Metzcar, J., Wang, Y., Heiland, R., and Macklin, P. (2019). A Review of Cell-Based Computational Modeling in Cancer Biology. *JCO Clin. Cancer Inform.* 3, 1–13. doi:10.1200/CCI.18.00069–

Montagud, A., Ponce-de-Leon, M., and Valencia, A. (2021). Systems Biology at the Giga-Scale: Large Multiscale Models of Complex, Heterogeneous Multicellular Systems. *Curr. Opin. Syst. Biol.* 28, 100385. doi:10.1016/j.coisb.2021.100385

Ozik, J., Collier, N., Heiland, R., An, G., and Macklin, P. (2019). Learning-accelerated Discovery of Immune-Tumour Interactions. *Mol. Syst. Des. Eng.* 4, 747–760. doi:10.1039/C9ME00036D

Ozik, J., Collier, N. T., Wozniak, J. M., Macal, C. M., and An, G. (2018b). Extreme-scale Dynamic Exploration of a Distributed Agent-Based Model with the Emews Framework. *IEEE Trans. Comput. Soc. Syst.* 5, 884–895. doi:10.1109/TCSS.2018.2859189

Ozik, J., Collier, N. T., Wozniak, J. M., and Spagnuolo, C. (2016). "From Desktop to Large-Scale Model Exploration with Swift/T," in *2016 Winter Simulation Conference* (Washington, DC, United States: WSC), 206–220. doi:10.1109/WSC.2016.7822090

Ozik, J., Collier, N., Wozniak, J. M., Macal, C., Cockrell, C., Friedman, S. H., et al. (2018a). High-throughput Cancer Hypothesis Testing with an Integrated PhysiCell-EMEWS Workflow. *BMC Bioinformatics* 19, 483. doi:10.1186/s12859-018-2510-x

Reuillon, R., Leclaire, M., and Rey-Coyrehourcq, S. (2013). Openmole, a Workflow Engine Specifically Tailored for the Distributed Exploration of Simulation Models. *Future Generation Comput. Syst.* 29, 1981–1990. doi:10.1016/j.future.2013.05.003

Sedger, L. M., and McDermott, M. F. (2014). TNF and TNF-Receptors: From Mediators of Cell Death and Inflammation to Therapeutic Giants - Past, Present and Future. *Cytokine Growth Factor. Rev.* 25, 453–472. doi:10.1016/j.cytogfr.2014.07.016

Shaffer, S. M., Dunagin, M. C., Torborg, S. R., Torre, E. A., Emert, B., Krepler, C., et al. (2017). Rare Cell Variability and Drug-Induced Reprogramming as a Mode of Cancer Drug Resistance. *Nature* 546, 431–435. doi:10.1038/nature22794

Stoll, G., Viara, E., Barillot, E., and Calzone, L. (2012). Continuous Time Boolean Modeling for Biological Signaling: Application of gillespie Algorithm. *BMC Syst. Biol.* 6, 116. doi:10.1186/1752-0509-6-116

Sun, X., Bao, J., and Shao, Y. (2016). Mathematical Modeling of Therapy-Induced Cancer Drug Resistance: Connecting Cancer Mechanisms to Population Survival Rates. *Sci. Rep.* 6, 22498. doi:10.1038/srep22498

Tang, K. S., Man, K. F., Kwong, S., and He, Q. (1996). Genetic Algorithms and Their Applications. *IEEE Signal. Process. Mag.* 13, 22–37. doi:10.1109/79.543973

Whitley, D. (1994). A Genetic Algorithm Tutorial. *Stat. Comput.* 4, 65–85. doi:10.1007/bf00175354

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Ponce-de-Leon, Montagud, Akasiadis, Schreiber, Ntiniakou and Valencia. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 11 April 2022
doi: 10.3389/fmolb.2022.807228

Cyclic Attractors Are Critical for Macrophage Differentiation, Heterogeneity, and Plasticity

Manuel Azaid Ordaz-Arias[1,2], *Laura Díaz-Alvarez*[3,4†], *Joaquín Zúñiga*[5,6†], *Mariana Esther Martinez-Sánchez*[5*†‡] *and Yalbi Itzel Balderas-Martínez*[1*†‡]

[1]*Laboratorio de Biopatología Pulmonar, Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas, Mexico City, Mexico,* [2]*Licenciatura en Ciencias Genómicas, UNAM, Cuernavaca, Mexico,* [3]*Departamento de Inmunología, Instituto de Investigaciones Biomédicas, UNAM, Mexico City, Mexico,* [4]*Posgrado en Ciencias Biológicas, UNAM, Mexico City, Mexico,* [5]*Instituto Nacional de Enfermedades Respiratorias Ismael Cosío Villegas, Mexico City, Mexico,* [6]*Tecnológico de Monterrey, Escuela de Medicina y Ciencias Biomédicas, Mexico City, Mexico*

Adaptability, heterogeneity, and plasticity are the hallmarks of macrophages. How these complex properties emerge from the molecular interactions is an open question. Thus, in this study we propose an actualized regulatory network of cytokines, signaling pathways, and transcription factors to survey the differentiation, heterogeneity, and plasticity of macrophages. The network recovers attractors, which in regulatory networks correspond to cell types, that correspond to M0, M1, M2a, M2b, M2c, M2d, M2-like, and IL-6 producing cells, including multiple cyclic attractors that are stable to perturbations. These cyclic attractors reproduce experimental observations and show that oscillations result from the structure of the network. We also study the effect of the environment in the differentiation and plasticity of macrophages, showing that the observed heterogeneity in macrophage populations is a result of the regulatory network and its interaction with the micro-environment. The macrophage regulatory network gives a mechanistic explanation to the heterogeneity and plasticity of macrophages seen *in vivo* and *in vitro*, and offers insights into the mechanism that allows the immune system to react to a complex dynamic environment.

Keywords: boolean network, macrophage, oscillation, cycles, differentiation, heterogeneity, plasticity, regulatory network

OPEN ACCESS

Edited by:
Alberto Jesus Martin, Facultad de Ciencias, Chile

Reviewed by:
Fransisca Leonard, Houston Methodist Research Institute, United States
Mohit Kumar Jolly, Indian Institute of Science (IISc), India

***Correspondence:**
Mariana Esther Martinez-Sánchez
maresther.martinez@iibiomedicas.unam.mx
Yalbi Itzel Balderas-Martínez
yalbibalderas@gmail.com

†ORCID:
Manuel Azaid Ordaz-Arias
https://orcid.org/0000-0002-5963-2630
Laura Díaz-Alvarez
https://orcid.org/0000-0002-8604-5714
Joaquín Alejandro Zúñiga Ramos
https://orcid.org/0000-0002-7143-0281
Mariana Esther Martinez-Sánchez
https://orcid.org/0000-0002-6115-1449
Yalbi Itzel Balderas-Martínez
https://orcid.org/0000-0001-6400-0526

‡*These authors have contributed equally to this work and share last authorship*

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *01 November 2021*
Accepted: *11 March 2022*
Published: *11 April 2022*

Citation:
Ordaz-Arias MA, Díaz-Alvarez L, Zúñiga JD, Martinez-Sánchez ME and Balderas-Martínez YI (2022) Cyclic Attractors Are Critical for Macrophage Differentiation, Heterogeneity, and Plasticity. Front. Mol. Biosci. 9:807228. doi: 10.3389/fmolb.2022.807228

INTRODUCTION

The balance between inflammatory and anti-inflammatory immune responses is crucial to maintain homeostasis in the face of the diverse immune challenges an organism meets. Macrophages are cells essential to immunity. They recognize pathogens and pathogen-derived molecules, collaborate with other cells of the innate and adaptive immune system, and are critical players both in chronic inflammation and in tissue regeneration (Sica and Mantovani, 2012; Varol et al., 2015; Vannella and Wynn, 2017; Li et al., 2019; Locati et al., 2020). Macrophages are characterized by their diversity and plasticity. Depending on the signals received, non-polarized M0 macrophages can be polarized into two main types: classically activated macrophages or M1, characterized by a pro-inflammatory profile, and alternatively activated macrophages or M2, which promote proliferation and repair (Mendoza-Coronel and Ortega, 2017; Funes et al., 2018).

M0 macrophages are usually monocytes differentiated into M0 macrophages in the presence of GM-CSF that have not been exposed to any pro or anti-inflammatory stimulus or environment that promotes their activation, cytokine production, and functional polarization (Kumar, 2019). M1 polarization is generally triggered by the stimulation of TLRs, or by cytokines such as IFNγ and GM-CSF, which lead to high production of pro-inflammatory cytokines such as IL-1β, IL-6, IL-12, and IL-23, in addition to a low expression of IL-10 (Lehtonen et al., 2002; Park et al., 2009; Weber et al., 2010; At et al., 2011; Lawrence and Natoli, 2011; Liu et al., 2014; Bally et al., 2015; Funes et al., 2018; Hamilton, 2019; Wang et al., 2019; Petrina et al., 2021). M2 polarization has been subdivided into M2a, M2b, M2c, and M2d macrophages, due to diverse transcriptional programs and stimuli involved (Huang et al., 2018). The M2a macrophages are derived from M0 cells stimulated by IL-4 and IL-13, they release high levels of IL-10 and TGF-β using transcriptional factors (TFs) such as STAT6 and IRF4, and are involved in proliferation and tissue repair functions (Bouhlel et al., 2007; Chawla, 2010; Gordon and Martinez, 2010; Ma et al., 2015; Arora et al., 2018; Wang et al., 2019). A combination of TRL ligands generates M2b macrophages and immune complexes (IC) as well as IL-1R ligands, yielding both pro-inflammatory cytokines, such as IL-1β, IL-6, TNF-α, and IL-12, and anti-inflammatory cytokines such as IL-10. The signaling pathways stimulated involve MAPKs, PI3K/Akt, and, ultimately, NF-κB. M2b macrophages have a role in the regulation of inflammatory responses (Lucas et al., 2005; Park et al., 2009; Zhang et al., 2009; Luo et al., 2010; Weber et al., 2010; Foey, 2014; Liu et al., 2014; Bally et al., 2015; Wang et al., 2019). M2c macrophages arise upon IL-10 stimulation. They express high levels of IL-10, which induces the phosphorylation of STAT3, thus negatively regulating the production of pro-inflammatory cytokines (Hutchins et al., 2013; Ma et al., 2015; Wang et al., 2019). M2d macrophages are induced by the costimulation of the adenosine A2 receptor and TLR, expressing levels of IL-10 as well as IL-12, and characterized by presenting properties of tumor-associated macrophages that carry out angiogenesis and tumor progression (Leibovich et al., 2002; Grinberg et al., 2009; Park et al., 2009; Colin et al., 2014; Arora et al., 2018; Wang et al., 2019; Anders et al., 2021). However, these are not the only possible expression patterns, as variations have been found *in vitro* and *in vivo*.

Given th e plasticity, heterogeneity, and adaptability of macrophages and their role in the immune system, it is important to understand their phenotypic landscape, the conditions in which they originate, and the possible transitions between subsets. Macrophage differentiation can be seen as a continuum between M1 and M2 phenotypes, where these cells can express different profiles and concentrations of cytokines, receptors, and transcription factors (Sica and Mantovani, 2012; Palma et al., 2018). At the same time, not all combinations of key molecules like IL-12, IL-10, IL-6, or VEGF are possible. For example, the IFNγ-induced and IL-4-induced programs inhibit each other in the cell, leading to heterogeneous populations in environments with mixed signals (Munoz-Rojas et al., 2021). Furthermore, it is known that the signaling pathways have inhibitory mechanisms that lead to self-regulation, causing oscillations in the expression of cytokines like IL-6, which are expected as part of the physiological behavior of macrophages, but that can also act against the host in pathological scenarios (Wang et al., 2013).

Regulatory networks are a valuable tool to bridge the molecular regulation of a cell with its phenotype and have been used to study the differentiation of hematopoietic cells (Saez-Rodriguez et al., 2007; Naldi et al., 2010; Martinez-Sanchez et al., 2015; Liquitaya-Montiel and Mendoza, 2018; Ramırez and Mendoza, 2018; Palma et al., 2018; Ramirez et al., 2019; Avila-Ponce de León et al., 2021) and the plasticity of macrophages (Palma et al., 2018; Ramirez et al., 2019; Avila-Ponce de León et al., 2021). Boolean networks integrate qualitative data about the interactions between cytokines, signaling pathways, and transcription factors to predict differentiation and plasticity. They allow testing different hypotheses and determine how the regulatory structure impacts complex cellular behaviors, all of this, with only a few parameters (Kauffman, 1969; Albert and Thakar, 2014).

This paper presents a Boolean model of the regulatory network that underlies macrophage differentiation, extending previous approaches (Palma et al., 2018; Ramirez et al., 2019). The model recovers the M0, M1, M2a, M2b, M2c, M2d, and other M2-like cell types, including several cyclic attractors that reproduce known experimental data. Then, we use the model to study how the classic polarizing environments and mixed combinations of extrinsic signals affect the stability of these cells. We show that the plasticity, heterogeneity, adaptability, and variable levels of expression of key cytokines in macrophages result from the structure of the regulatory network.

MATERIALS AND METHODS

All the datasets, scripts, tables, and images used in this study can be found in the repository https://github.com/mar-esther23/Macrophage_Differentiation.

A Boolean network consists of nodes representing molecular components (i.e., cytokines, signaling pathways, transcription factors) and edges representing the interactions between them. The value of the nodes is a discrete variable: one if the node is functional and 0 if it is not functional. The value of a node i at the time *t+1* depends on the value of its regulators at time *t*, according to a logical function that recapitulates available biological information. The state of the network at *x(t)* depends on the values of all its nodes and will evolve through time as the regulatory functions are evaluated. Eventually, the system will arrive at an attractor, which corresponds to a cell type. These attractors can be steady states when $x_t = x_{t+1}$ or cycles when $x_t = x_{t+\tau}$ (Kauffman, 1969; Albert and Thakar, 2014).

We constructed the macrophage regulatory network according to previous models (Palma et al., 2018) and available information (Leibovich et al., 2002; Yoshimura et al., 2007; Chang et al., 2013; Wang et al., 2013; Liu et al.,

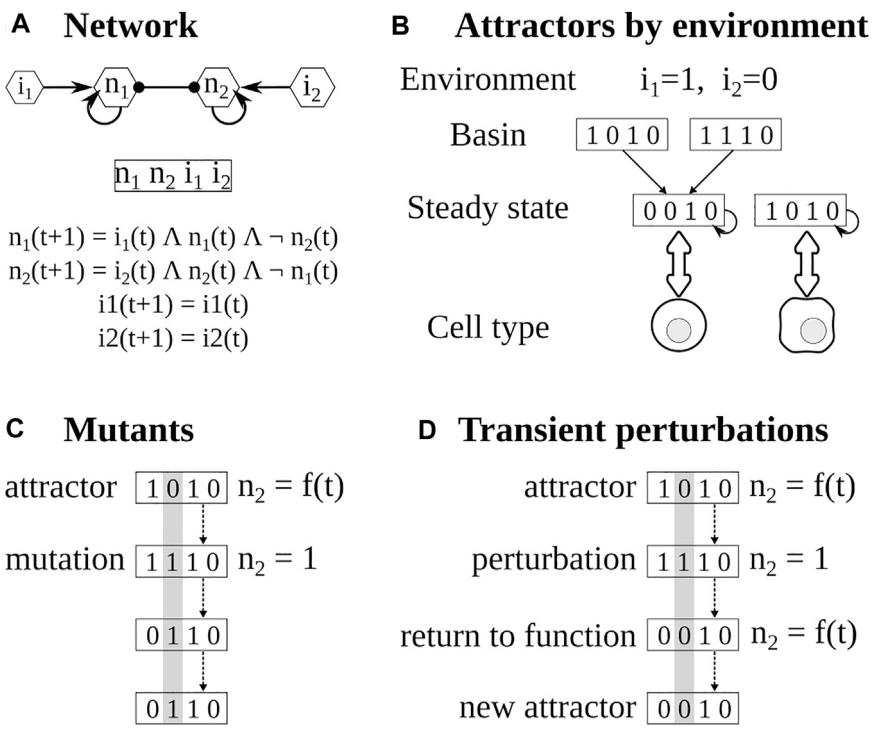


FIGURE 1 | Pipeline for the analysis of Boolean networks. **(A)** The network is constructed using available experimental information. The state of the network depends on the values of each node. The value of each node depends on its regulators. **(B)** The attractors of the network are calculated using BoolNet, the attractors depend on the inputs or environment and the functions. **(C)** Mutants are obtained by fixing the value of the target node for the whole simulation. **(D)** In transient perturbations, the target node is changed (bitflip) for one time step, and then the perturbation is relaxed. Eventually, the network may stay, return to the original attractor, or reach a different one.

2014; Wilson, 2014) among others which can be seen in (**Supplementary Table S1**, **Figure 1A**). The dynamical analysis of the network was done using the packages BoolNet (Mussel et al., 2010) and BoolNetPerturb (Martinez-Sanchez et al., 2018). Synchronous updating was used in all simulations.

We determined attractors of the network and classified them depending on the expression of both the characteristic transcription factor and cytokine. In every case, we focused on the presence and absence of the nodes that correspond to common cell type markers, and ignored the value of the other nodes (Kauffman, 1969; Albert and Thakar, 2014; Martinez-Sanchez et al., 2015). The basin of attraction of a network is the set of states that lead to an attractor (steady state or cycle) in the simulation.

To determine the effect of the micro-environment, we first determined the cytokines present in each polarizing environment (Kauffman, 1969; Albert and Thakar, 2014; Martinez-Sanchez et al., 2015), then, we fixed the corresponding input nodes according to the cytokines present or absent in that environment and then determined the resulting attractors (**Figure 1B**). This is modeled by changing the input function $i_{t+1} = i_t$ to $i_{t+1} = = 0$ or $i_{t+1} = 1$ depending on the presence or absence of the cytokines in the environment.

To further verify the model, we simulated the knock-out and overexpression of target nodes by setting their values to 0 or 1 and comparing the resulting attractors with known mutants (Kauffman, 1969; Albert and Thakar, 2014; Martinez-Sanchez et al., 2015) (**Figure 1C**). Furthermore, we checked that the attractors found with synchronous updating could be found using asynchronous updating.

The expression pattern of a cell can change in response to changes in both internal and environmental factors. We focused on the effect of small transient perturbations. For example, due to stochastic effects, a transcription factor may not be activated because the polymerase fails to bind to its DNA sequence for a time, even if the rest of the regulators are present. This can be modeled as the corresponding node having a value of zero for a time step, and then the perturbation will be relaxed and the node will acquire a new value depending on its regulators (**Figure 1D**). On the other hand, a cell may be subjected to a small peak of a cytokine in its environment. This can be modeled as the extrinsic cytokine node having a value of one for a time step and then returning to its original value. The attractor of the system, which corresponds to the cell type, may change or not depending on the regulatory network, the original state of the network, and the perturbed node. To study the stability and plasticity of the system for each microenvironment we took its attractors and modified for one time step the value of the node (bitflip), then the perturbation was relaxed, and the resulting attractor was determined (Martinez-Sanchez et al., 2018, 2015).

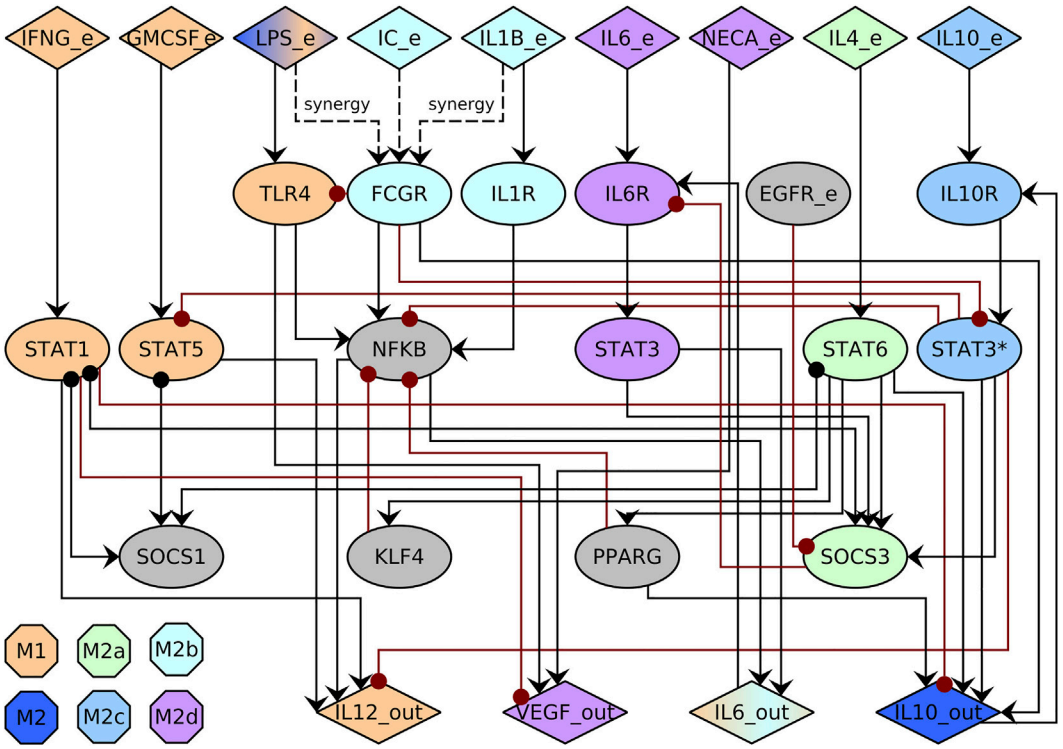


FIGURE 2 | Macrophage regulatory network. The network includes cytokines in the environment (_e) and produced by the macrophage (_out), signaling pathways, and transcription factors (ellipses). Activations are represented with black arrows, and inhibitions with red dotted arrows. We use STAT3 for IL-6 dependent signaling and STAT3* for IL-10 dependent signaling. The color of the node corresponds to the associated cell type.

RESULTS

Macrophage Differentiation Patterns Emerge From Feedback Between Transcription Factors, Cytokines, and Signaling Pathways

We expanded the previously published macrophage regulatory networks (Palma et al., 2018; Ramirez et al., 2019). In this network, we included multiple molecules like transcription factors, STAT proteins, cytokine receptors, SOCS proteins, and cytokines, among others. We only included direct interactions that have been experimentally validated (**Supplementary Table S1**, **Supplementary Table S2**) to include Ie IL6 (Chang et al., 2013; Wang et al., 2013; Liu et al., 2014; Wilson, 2014), NECA (Leibovich et al., 2002), EGFR (Wang et al., 2013), and SOCS3 (Yoshimura et al., 2007; Wang et al., 2013; Wilson, 2014). Then, we simplified the network using GINSIM(Gonzalez et al., 2006). The resulting network has 29 nodes and 52 interactions (**Figure 2**, **Supplementary Table S3**). We assumed that different pathways mediate IL-6 and IL-10 signaling by STAT3 and marked them as STAT3 for IL-6 dependent signaling and STAT3* for IL-10 dependent signaling. The state of a node represents whether the biological component is active 1) or inactive (0). A node is active if it can alter the regulation of other nodes.

Then, we determine the macrophage cell types by calculating the attractors, steady states, and cycles of the network (Kauffman, 1969) and label them. An attractor corresponds to a cell type if the characteristic signaling pathways, transcription factors, and produced cytokines are present (**Supplementary Table S4**). The network recovers 44 steady-state attractors and 358 cyclic attractors of size 2, 3, and 6 which correspond to 'M1', 'M2b', 'M2a', 'M2c', 'M2d', 'M2' (M2-like), 'M0', and 'il6' cell types (**Figure 3**, **Supplementary Figure S2**, **Supplementary Table S5**).

M1 macrophages produce IL-12 and may produce IL-6 and activate the pathways for STAT1, STAT5, or NFKB. M0 and cyclic M0* macrophages do not produce any cytokines and correspond to naive macrophages. They are usually found in simulated environments with neither extrinsic cytokines nor contradictory extrinsic signals that inhibit each other through SOCS proteins. The attractors labeled 'il6' produce IL-6 but no other extrinsic cytokines. The steady states il6 macrophages are only found when there is EGFR_e in the microenvironment and may correspond to inflammatory pathogenic states as those seen in cancer (Wang et al., 2013) and severe COVID-19 (Matsuyama et al., 2020; Merad and Martin, 2020). In contrast, the cyclic il6* attractors may correspond to pathogenic states or be non-fully differentiated macrophages.

M2 macrophages produce IL-10 or VEGF, and they can be classified into different subtypes depending on the cytokines produced and active signaling pathways. The M2a subtype

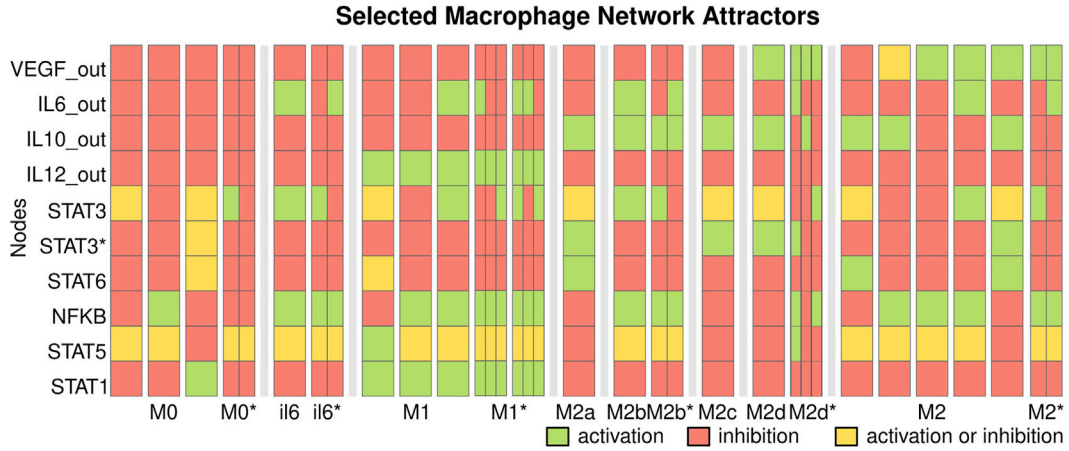


FIGURE 3 | Selected macrophage regulatory attractors. The attractors of the macrophage regulatory network correspond to cell types. Each column corresponds to key nodes of a state; attractors are separated by white spaces and cell types by black bars. We include the cyclic attractors as narrow columns that represent the oscillation. Each node can be active (green) or inactive (red), or active or inactive (yellow). The network recovers the attractors corresponding to M1, M2a, M2b, M2c, M2d, and M2-like cell types.

produces IL-10 and activates the STAT6 pathway, M2b produces IL-10 and IL-6, M2c produces IL-10 and activates the IL-10 dependent STAT3* pathway, M2d macrophages produce VEGF, IL-10 and no IL-12. The model recovers steady states corresponding to these cell types, including cyclic attractors for M2b with an IL-6 dependent STAT3 oscillation. We also recover M2-like subsets that produce IL-10 and VEGF or IL-6.

Most cyclic attractors present oscillations in the IL6/STAT3/SOCS3 circuit, which may affect the downstream production of IL6_out. The self-inhibition of the IL6 pathway causes these oscillations: STAT3 induces SOCS3 expression, which inhibits IL6R and STAT3 phosphorylation, causing a repressed circuit and oscillations. The IL10/STAT3* pathway also presents oscillations that may affect SOCS1, NFKB, and STAT5. These oscillations can be caused by inhibition by other pathways, for example, IL6. The self-inhibition caused by SOCS3 and the other cycles may have a role in limiting the production of IL-6 by macrophages and the associated hyperinflammation. Given the hypothesis that macrophage differentiation is a continuum (Sica and Mantovani, 2012; Palma et al., 2018) these M2-like and cyclic states may be a mechanism to regulate the production of cytokines by macrophages.

The number of attractors associated with a cell type does not necessarily correspond to the number of states that reach the attractors of that cell type (basin of attraction). The biggest basin is M0, followed by M2, M2a, and M1, while il6, M0*, and il6* had the smaller basins. In general, the cell types labeled as cyclic (*) had smaller basins.

To validate the model we verified whether the attractors were robust to asynchronous updating. All the steady states were robust to the change in update schema. While most of the attractors of sizes 2 and 3 were unstable we found asynchronous attractors of size 6 or more that correspond to M0*, il6*, M2b*, and M2*, but lost the M1* and M2d*.

To further validate the model, we compared the knock-out and over-expression simulations with experimental data (**Supplementary Figure S3**, **Supplementary Table S6**). In general, the predictions made by the model correspond to the observed biological data (**Supplementary Table S7**). For instance, in STAT1-null macrophages stimulated with IFN-γ and Pam3CSK4, a dose-dependent decrease in IL-12 has been experimentally observed compared to wild-type macrophages (Kim et al., 2015). This phenomenon is recovered by a network simulation of a STAT1 knock-out, where we see that attractors completely lose IL-12 production, causing the disappearance of M1. In the same way, it has been experimentally obtained that inhibition of PPARγ-dependent gene expression significantly decreases the production of IL-10 mediated by LPS, which is recovered in the simulations in which a knock-out of PPARγ was set, obtaining a decrease in IL-10-producing attractors (Majai et al., 2007). Furthermore, regarding overexpression, there is the experimental case where *in vitro* STAT6 has been overexpressed, causing a promotion of M2 macrophages (Gong et al., 2017). An overexpression simulation in STAT6 recovers this last, which causes a higher proportion of M2 attractors and a decrease in M1. However, these simulations do not recover the expected behavior in the case of the NFkB mutant. The NF-kB pathway is a highly complex protein, but our network simplifies it to a single node, so this discrepancy is probably the result of the modeling decisions. In this mutations experiments, the most stable states were the M0*, il6*, and M2d macrophages. On the other hand, the more sensitive cell types are M0 and M1. The nodes that tended to cause more changes in differentiation if mutated are IL12_out and IL10_out, which affect the cytokine profile, followed by STAT1, STAT6, and IL-10 mediated STAT3 activation (STAT3*). Furthermore, this analysis predicts the effect of knock-out and over-expression mutants that have not been tried experimentally.

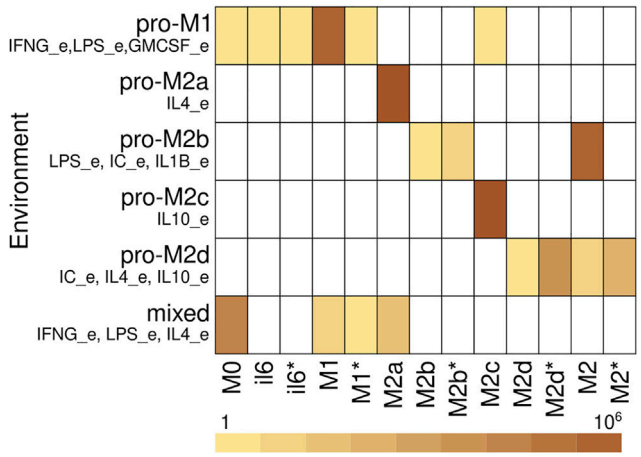


FIGURE 4 | Macrophage differentiation in response to the micro-environment. Cell types recover depending on the cytokines in the micro-environment. The rows correspond to the environment and the columns to the cell type. The color corresponds to the basin size of the attractors that correspond to each cell type. If a cell type was not recovered on a microenvironment, it is represented with white.

Role of the Micro-Environment in Macrophage Differentiation and Stability

Macrophage differentiation does not occur in a vacuum but in response to the micro-environmental signals (**Supplementary Table S8**). To determine the role of the microenvironment, we determined the attractors associated with a cell type and their combined basin size in different microenvironments (**Figure 4**, **Supplementary Table S9**). The pro-M1 microenvironment contains IFNG_e, GMCSF_e, and LPS_e, and the model mainly recovers the presence of M1 attractors with a small number of M1*, M0, il6, il6* attractors. The pro-M2a environment contains IL4_e, and the model recovers only M2a attractors. The pro-M2b environment contains LPS_e, IC_e, and IL1B_e, and the model recovers M2, M2b, and M2b*. The pro-M2c environment contains IL10_e, and the model recovers only M2c attractors. The pro-M2d environment contains IC_e, IL4_e, and IL10_e, and the model recovers M2d, M2d*, M2, and M2* attractors. The mixed environment contains LPS_e, IFNG_e, and IL4_e, which are associated with M1 and M2a polarization, and recovers M0, M1, M1*, and M2a attractors, which are in accordance with the heterogeneity observed in macrophage populations subjected to *in vitro* co-stimulation with these same cues (Munoz-Rojas et al., 2021).

Macrophage differentiation is not a wholly deterministic process, but it can be affected by transient changes in the environment, stochastic noise during transcription, traduction and signaling events, and other types of noise. Furthermore, the environment and the internal state of the cell can have small changes in response to the progress of a pathological state. To study this for each of the six environments, we took the recovered attractors and perturbed each node one by a time step, and determined if that changed the resulting attractor and cell type (**Figure 5**, **Supplementary Table S10**). In the pro-M1 environment, most of the transitions are between M1 and M1*, with a bias towards the cyclic M1* attractors. Most M1* attractors have oscillations in the IL6/STAT3/SOCS3 pathway, and some of them have oscillations in IL6_out. There is a small number of transitions towards M0 and il6/il6* that increase in percentage, which may have a role *in vivo* by limiting the production of IL-12 as the infection progresses toward resolution. In the pro-M2b environment, there is a small number of transitions between M2 and M2b/M2b*, with a slight bias towards M2. In the pro-M2d environment, there are also transitions between M2/M2* and M2d/M2d*. In the pro-M2a and pro-M2c environments, there are only M2a and M2c attractors, so these are stable. In the mixed environment, there are transitions from M0 to and from all differentiated environments but not between M1 and M2a attractors, which may indicate that the plasticity between these cell types requires longer signals, especially in mixed environments, and that a temporal cease of cytokine production precedes a transition between M1 and M2 cell types.

In general, we can say that each microenvironment favors the differentiation and stability of the associated cell type, even in some cases where there is a small number of attractors associated with an additio'nal cell type. The exceptions are the pro-M2b and pro-M2d environments, where there is a strong presence of M2-like attractors; however, this can be seen as part of the phenotypic plasticity of macrophages. If we consider all possible combinations of cytokines, most of the cell types were highly stable, with only a small proportion of transitions between subsets. The cyclic attractors associated with a cell type were highly stable, as the oscillations seem to be the result of the network topology and not a dynamical artifact.

To determine the key nodes for the dynamic stability of the model, we determined which nodes caused more changes between cell types when transiently perturbed (**Figure 6**, **Supplementary Table S10**). The nodes that caused more transitions between cell types were STAT1, IL-10 mediated STAT3*, and STAT6, which are associated with the signaling pathways of key cytokines in macrophage differentiation. IC_e and FCGR also had an essential role in the stability of the model, as they regulate both NFKB, STAT3*, and IL10_out. STAT1, STAT3*, STAT6, and FCGR have a higher number of out-going edges and directly or indirectly modulate the activation and inhibition of different circuits of the network. The activation of SOCS1 also has a relevant role, as its activation inhibits the STAT1, STAT5, and STAT6 nodes. The nodes that cause fewer transitions between cell types are IL12_out, VEGF_out, NECA_e, IL1R, and TLR4.

In general, of the single state transient perturbations 21% resulted in a change of cell type and 2.96% were transitions between M1 and M2 states. We also simulated all possible double node perturbations 34.44% resulted in a change of cell type and 5.02% were transitions between M1 and M2 states. We also realized a Derrida curve (Derrida and Pomeau, 1986) to determine how sensible the system was to perturbations in the

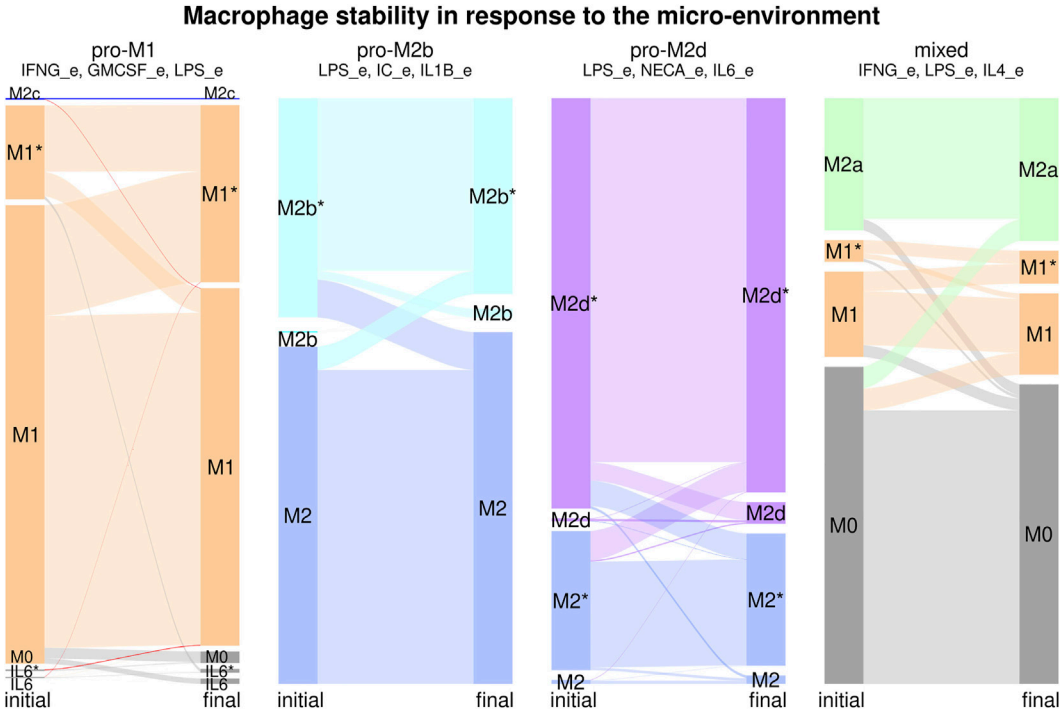


FIGURE 5 | Macrophage stability in response to the microenvironment. For each environment we calculated the attractors, then, for each attractor, we transiently perturbed every node independently for one time step to determine the stability of the different cell types. Each stability experiment is represented by a flux diagram, where the colored boxes correspond to each cell type. The initial state is on the left of the diagram, and the final state is on the right. The height of the bar corresponds to the basin size of the attractor. The width of the lines between boxes represents the transitions between attractors.

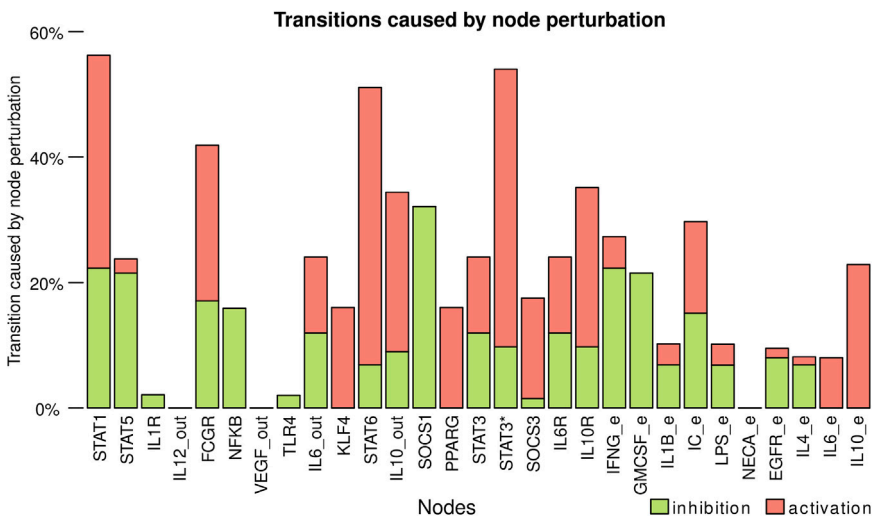


FIGURE 6 | Transitions caused by node perturbation. Each bar corresponds to the percentage of perturbations of a node that caused a transition between cell types. Activation of the nodes is represented with green and inhibitions with red. https://github.com/mar-esther23/Macrophage_Differentiation/blob/master/images/MP_transitionnode_clean.png.

value of the states. For perturbations of Hamming distance 1 on average the change was of 4.18 nodes and increased towards 8.47 nodes on average as the number of perturbed nodes increased (**Supplementary Figure S3**). It is worth taking into account that the high number of input nodes, that represent the micro environment, heavily influence these results. For example, on a

micro-environment with only M2 attractors it is impossible for there to be a transition towards M1, as the cell type is not stable. Furthermore, not all possible micro-environments can be found *in vivo*, which implies that while a transition may be possible in the system it might not be observed *in vivo*.

DISCUSSION

In this paper, we propose an actualized regulatory network of cytokines, signaling pathways, and transcription factors to study the differentiation, heterogeneity, and plasticity of macrophages. This network allows us to give a mechanistic explanation of the dynamic behavior of these cells in response to different micro-environments. Furthermore, the network recovers multiple cyclic attractors that are both stable to perturbations and in accordance with previous experimental observations (Wang et al., 2013), showing that these oscillations are the result of the structure of the network and suggesting that they may have a biological function. In fact, the biological relevance of oscillatory behavior seems to lie in exquisitely regulated phenomena like the activation and nuclear translocation of NFKB. Cheng et al. (2021) recently demonstrated that macrophages respond to different proinflammatory micro-environments with oscillatory or non-oscillatory activity of NFKB. ChIP-seq data and computational modeling analysis reveal that NFKB presents oscillatory behavior in the majority of stimulated cells, however, only non-oscillatory behavior leads to sufficiently prolonged chromatin accessibility, favoring gene expression. This phenomenon is regulated by the NFKB inhibitor IkBa, which may allow sensing of the micro-environment while refraining the cell from secreting inflammatory mediators until the concentration of a certain cue reaches a specific threshold. All of it without changing the cell phenotype. Therefore, we can also speculate that inflammatory disorders may arise or be sustained by macrophages whose oscillation is skewed towards non-oscillatory behavior. The macrophage regulatory network recovers steady state and cyclic attractors that correspond to M0, M1, M2b, M2d, M2-like, and IL-6 producing macrophages. Cyclic attractors represent 89% of the total attractors, but their combined basins of attraction are only 11% of the total state space. However, it is worth noting that most of these attractors are stable, and when perturbed, most perturbations lead to a cycle of the same cell type. Oscillatory activation of STAT3 with its downstream effect in IL-6 production in macrophages has been previously reported (Wang et al., 2013). The oscillations in the macrophage regulatory network are the result of the IL-6R/STAT3/SOCS3 pathway, and the crosstalk with other signaling pathways, like those mediating the activation of STAT3* via STAT6, STAT5, or IL-10, which are circuits commonly observed in immune cell regulatory networks.

In vivo and *in vitro* macrophages express marker molecules and cytokines in an expression range, which can be observed in a flow cytometric analysis as the spread of the population on a dot plot or the width of a histogram (Munoz-Rojas et al., 2021). The expression range can vary depending on the cell type, the molecule being measured, and the pathological state. For example, the levels of IL-6 produced by macrophages in cancer and COVID-19 are associated with the severity of the disease (Wang et al., 2013; Matsuyama et al., 2020; Merad and Martin, 2020). How do macrophages, and other immune cells, generate and regulate the expression range is an open question. Oscillations, and their associated cyclic attractors, could be a mechanism to create variability in the expression range of a molecule. For example, when averaged over time, the oscillations in STAT3 activation, and their downstream targets in M1 macrophages could create diverse expression levels that depend on the structure of the network. These oscillatory circuits could be further finely tuned by other mechanisms like the induction of signaling pathways, transcriptional regulation or stochastic effects like variations in local cytokine concentration, noise in signaling pathways, transcription factors binding, etc. Cyclic attractors are usually ignored when studying Boolean dynamics in hematopoiesis (Alvarez-Buylla Roces et al., 2018), but the relevance of the oscillatory dynamics *in vivo* and in this network (Wang et al., 2013; Munoz-Rojas et al., 2021) indicates that more methods should be developed to study cyclic attractors and determine when and how they have a functional role.

The macrophage regulatory network also allows us to study the effect of the environment on the heterogeneity and plasticity of macrophages. All the polarizing environments favor the attractor associated with it, for example, in a pro-M2a environment, we found M2a attractors. At the same time, in most of the environments we studied (pro-MI, pro-M2b, pro-M2d, and mixed), there was more than one possible cell type, which implies that the heterogeneity in macrophages populations is a result of the regulatory network. This was especially notable in the mixed environment (LPS + IFNγ + IL4), where we recover M0, M1, and M2a macrophages (Munoz-Rojas et al., 2021). The specific differentiation pathway a cell follows is also a result of the regulatory network, the internal state of the cell, and stochastic events. Studying the basins of attraction of the different cell types and their sensibility to stochastic events may allow us to understand the heterogeneity of macrophage populations better. For example, *in vitro* stimulation with a combination of LPS, IFN-γ and IL-4 produces heterogeneous populations with M1 and M2 sub-populations (Munoz-Rojas et al., 2021). In this study Muñoz-Rojas et al. use a combination of molecular and cell biology techniques, including single-cell RNA sequencing (scRNA-seq), to ascertain the global transcriptional programs that lead to the observed heterogeneity. Similar results have been observed *in vivo*, where scRNA-seq of macrophage populations has also shown the coexistence of two clearly defined subpopulations in adipose tissue that do not follow the classic M1/M2 paradigm and whose proportions vary depending on the micro-environment (Grosjean et al., 2021). The differentiation of each individual cell depends on the initial state of the cell (transcription factors expressed and active signaling pathways) and stochastic events (local cytokine concentration, noise in signaling pathways, transcription factors binding, etc.), which generates an initial variability. Such variability determines which pathways of the regulatory network activate and which to inhibit, to polarize the individual cells into clearly defined

subpopulations, thus maintaining a heterogeneous population. These results coincide with our findings that in mixed environments M0, M1, and M2 macrophages coexist, implying that the design and performance of our network are appropriate to recover the outcomes of complex scenarios reported *in vivo* and *in vitro* after extensive analyses.

The model also allowed us to study macrophage plasticity. The environment determines plasticity because it limits the accessible cell types and modulates the effect of perturbations. In general, most perturbations did not cause changes in the labeled cell type. However, transitions between a steady state and a cyclic attractor of the same cell type were common, which means that cells that may be classified as the same cell type given their membrane markers may have different internal states, creating a hidden source of heterogeneity to respond to changes in the micro-environment. In polarizing environments (pro-M1 and pro-M2), most of the transitions were towards the favored cell type. There was a high level of transitions among the different subtypes of M2 attractors but limited transitions towards M1. This seems to indicate that M2 attractors are more closely related to one another than to M1 attractors, fine tuning their regulatory activity and creating a continuum of M2-like states. The multiple inhibitions between M1 and M2 transcriptional programs help the system maintain a stable inflammatory or regulatory program, making these the two poles of macrophage differentiation. This is especially relevant when taking into account the key role of macrophage differentiation and plasticity in COVID-19 and cancer (Wang et al., 2013; Li et al., 2019; Matsuyama et al., 2020; Merad and Martin, 2020). Perturbations that favor M2 macrophages can favor transitions towards more aggressive cancers, even in situations where perturbing the cancer cells may not be enough to change the steady state behavior of the system (Li et al., 2019). This seems to imply that there are a series of feedback loops between the tissues, the environment, and immune cell populations that are crucial to understanding complex diseases. Understanding these feedback loops will require us to conceive disease as a system where the cytokine and cellular environment play a key role.

The mixed environment (LPS + IFN-γ + IL4) had a high number of non-differentiated M0 attractors, because the mixed signals most likely inhibited each other, as reported by Munoz-Rojas et al. (2021), where LPS + IFN-γ, and IL-4 give rise to orthogonal global transcriptional programs We observed only M0-M1 and M0-M2a transitions but no M1-M2a direct transitions. M1-M2a transitions are possible if they pass through an intermediate M0-like state with no cytokine production and require more than one perturbation. Such a phenomenon was indeed described by Tarique et al. (2015) for LPS + IFN-γ (M1), and IL-4 + IL-13 (M2) polarized macrophages, where the depletion of cytokines in the culture medium causes the cells to revert to the M0 phenotype, and by supplying the appropriate stimuli the macrophages can be re-polarized to the alternative phenotype. This could be a mechanism to warrant stability in the different cell types while allowing for plasticity if the environment changes past a certain time threshold. It also shows once more the power of experimental recapitulation of our network.

Traditionally, differentiation of hematopoietic cells has been considered a hierarchical process with clear differentiation pathways and well-defined cellular types, for example, M1 and M2 macrophages. However, as our understanding of macrophages in particular, and immune cells in general, has advanced, it is becoming increasingly clear that this is a highly dynamic process. Attempts to classify macrophages in subpopulations have proven intricate, as they seem to be both a continuum and a heterogeneous mix of subpopulations that do not always coincide with the M1/M2 paradigm (Sica and Mantovani, 2012; Wang et al., 2013; Mendoza-Coronel and Ortega, 2017; Grosjean et al., 2021; Munoz-Rojas et al., 2021). The analysis of our regulatory network suggests that there are independent circuits composed of receptors, transcription factors, and cytokines that activate in response to the signals in the environment. Some of these circuits inhibit each other (IFN-γ and IL-10), others are mostly independent (VEGF and IL-10), and others have more complex relationships (IL-6). Furthermore, these circuits can have dynamically stable oscillations, which affect not only the production of downstream cytokines, but also the crosstalk with other pathways. We propose that, on the one hand, when the circuits inhibit each other, we can expect a clear separation in the expression levels of the molecules involved (IL-12 and IL-10) and almost no plasticity, which creates pseudo-populations for those specific markers. On the other hand, when the circuits are independent or modulate each other in context-specific ways, the result is a continuum of expression for those markers, as seen in the M2-like family of attractors. In this case, given that the circuits are mostly independent, we should expect a higher level of "plasticity" as the circuits are activated or inhibited depending on the environmental signals. These circuits are further modulated by the microenvironment, the initial state of the cell, and stochastic effects. Focusing on the active regulatory circuits could give us a framework to comprehend the biological functions of macrophages in specific conditions while considering the environment, heterogeneity, and plasticity of these cells. This could have a profound impact on our understanding of the pathogenic mechanisms in certain diseases. For example, in patients with Crohn's disease there is a clear difference between macrophage populations from the intestinal mucosa and from the mesenteric fat tissue. In the former, TLR-4, IL-1b and IL-6 protein levels are higher compared to those in patients with non-inflammatory disease; while in the latter there is no such increase. The authors of the study attribute these differences to the micro-environment, which in the case of intestinal macrophages is largely determined by the interaction with the microbiota. As a consequence, there is an anomalous up-regulation of the signaling pathways that result in the production of inflammatory mediators. Hence, a network like the one we devised could be of great value to understand this type of heterogeneous scenarios, helping improve medical care towards the design of treatments with side effects noticeably reduced in comparison to the ones currently prescribed.

The model also allowed us to determine the key nodes of the network. When subjected to knock-out or over-expression

experiments, IL12_out, IL10_out, STAT1, STAT6, and IL-10 mediated STAT3 activation (STAT3*) had the most notable effect, especially IL-12 and IL-10, as they are cell type markers. Also, *in vivo* cells are subjected to transient changes in extrinsic cytokine levels or stochastic effects in signaling pathways and transcriptional regulation, which we simulated as transient perturbations. The nodes that caused more transitions between cell types were: STAT1, STAT3*, STAT6, IC_e, FCGR, and SOCS1. On the other hand, IL12_out, VEGF_out, NECA_e, IL1R, and TLR4 had the least effect. STAT1, STAT6, and STAT3* activation has a higher number of out-going edges and directly or indirectly modulates the activation and inhibition of different network circuits, which explains their key roles within the network dynamics.

While the Boolean nature of the model favors the study of how the structure of the regulatory network determines cellular behavior, it also limits the scope of the analysis. The model uses discrete values for the nodes, severely restricting our understanding of how the range of expression levels observed in macrophages is generated. Furthermore, the model uses discrete time steps and synchronous actualization for very different processes like signaling, which can take minutes, and transcription, which can take hours. Most cyclic attractors of size two or three were unstable, but we did recover asynchronous cyclic attractors of size six or bigger, however understanding their biological implications is still an open question. The model is also deterministic, and the perturbation analysis, while sufficient to determine the possible transitions, is not a true stochastic analysis. This is particularly important as the internal state of the cell and random noise probably have an important role in the emergence of heterogeneous populations. Further models using differential equations or stochastic methods are warranted.

The model also oversimplifies the NF-kβ pathway to a degree where the predicted and experimental mutants are not in accordance. Additionally, our network would benefit from the inclusion of multiple molecules like TNF-α, TGF-β, TLRs, NODs, and MyD88. In fact, it will be necessary to incorporate these molecules to integrate the model to other cell types and create integrative immune models. Finally, the number of environments used was limited to reported polarizing conditions and mixed environments (Sica and Mantovani, 2012; Wang et al., 2013; Mendoza-Coronel and Ortega, 2017; Grosjean et al., 2021; Munoz-Rojas et al., 2021). Thus, it will be interesting to see if the model can be extended to study diseases with complex immune profiles, like cancer, *tuberculosis*, or COVID-19.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/mar-esther23/Macrophage_Differentiation.

AUTHOR CONTRIBUTIONS

Design of the research study: MM-S, YB-M, JZ. Verified the model: LD-A. Performed analyses of the data: AO-A. All authors wrote the manuscript, contributed to the article, and approved the submitted version.

FUNDING

MMS received funding from CONACYT: Estancias Posdoctorales por México CVU 400554. LDA received funding from CONACYT: Doctorado Nacional CVU 399345.

ACKNOWLEDGMENTS

We acknowledge Miguel Ángel Pérez León for his technical support at "Laboratorio de cómputo de alto rendimiento, Departamento de Matemáticas, Facultad de Ciencias, UNAM", and the Laboratorio Nacional de Soluciones Biomiméticas para Diagnóstico y Terapia -LaNSBioDyT.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.807228/full#supplementary-material

REFERENCES

Albert, R., and Thakar, J. (2014). Boolean Modeling: A Logic-Based Dynamic Approach for Understanding Signaling and Regulatory Networks and for Making Useful Predictions. *Wires Syst. Biol. Med.* 6, 353–369. doi:10.1002/wsbm.1273

Alvarez-Buylla Roces, M. E., Martınez-Garcıa, J. C., Davila-Velderrain, J., Domınguez-Huttinger, E., and Martınez-Sanchez, M. E. (2018). Medical Systems Biology. *Adv. Exp. Med. Biol.* 1069, 1. doi:10.1007/978-3-319-89354-9_1

Anders, C. B., Lawton, T. M. W., Smith, H. L., Garret, J., Doucette, M. M., and Ammons, M. C. B. (2021). Use of Integrated Metabolomics, Transcriptomics, and Signal Protein Profile to Characterize the Effector Function and Associated Metabotype of Polarized Macrophage Phenotypes. *J. Leukoc. Bio* 111, 667–693. doi:10.1002/JLB.6A1120-744R

Arora, S., Dev, K., Agarwal, B., Das, P., and Syed, M. A. (2018). Macrophages: Their Role, Activation and Polarization in Pulmonary Diseases. *Immunobiology* 223, 383–396. doi:10.1016/J.IMBIO.2017.11.001

At, W., Hl, J., Pj, C., and Am, W. (2011). Tight Control of STAT5 Activity Determines Human CD34-Derived Interstitial Dendritic Cell and Langerhans Cell Development. *J. Immunol.* 186, 7016–7024. doi:10.4049/JIMMUNOL.1003977

Avila-Ponce de León, U., Vázquez-Jiménez, A., Matadamas-Guzman, M., Pelayo, R., and Resendis-Antonio, O. (2021). Transcriptional and Microenvironmental Landscape of Macrophage Transition in Cancer: A Boolean Analysis. *Front. Immunol.* 12, 2110. doi:10.3389/fimmu.2021.642842

Bally, A. P. R., Lu, P., Tang, Y., Austin, J. W., Scharer, C. D., Ahmed, R., et al. (2015). NF-κB Regulates PD-1 Expression in Macrophages. *J.I.* 194, 4545–4554. doi:10.4049/jimmunol.1402550

Bouhlel, M. A., Derudas, B., Rigamonti, E., Dièvart, R., Brozek, J., Haulon, S., et al. (2007). PPARγ Activation Primes Human Monocytes into Alternative M2

Macrophages with Anti-inflammatory Properties. *Cell Metab.* 6, 137–143. doi:10.1016/j.cmet.2007.06.010

Chang, Q., Bournazou, E., Sansone, P., Berishaj, M., Gao, S. P., Daly, L., et al. (2013). The IL-6/JAK/Stat3 Feed-Forward Loop Drives Tumorigenesis and Metastasis. *Neoplasia* 15, 848–IN45. doi:10.1593/NEO.13706

Chawla, A. (2010). Control of Macrophage Activation and Function by PPARs. *Circ. Res.* 106, 1559–1569. doi:10.1161/CIRCRESAHA.110.216523

Cheng, Q. J., Ohta, S., Sheu, K. M., Spreafico, R., Adelaja, A., Taylor, B., et al. (2021). NF-κB Dynamics Determine the Stimulus Specificity of Epigenomic Reprogramming in Macrophages. *Science* 372 (6548), 1349–1353. doi:10.1126/science.abc0269

Colin, S., Chinetti-Gbaguidi, G., and Staels, B. (2014). Macrophage Phenotypes in Atherosclerosis. *Immunol. Rev.* 262, 153–166. doi:10.1111/imr.12218

Derrida, B., and Pomeau, Y. (1986). Random Networks of Automata: a Simple Annealed Approximation. *Europhys. Lett.* 1 (2), 45–49. doi:10.1209/0295-5075/1/2/001

Foey, A. D. (2014). Macrophages - Masters of Immune Activation, Suppression and Deviation. *Immune Response Activation*, 276 Chapter 5. https://www.intechopen.com/books/3811. doi:10.5772/57541

Funes, S. C., Rios, M., Escobar-Vera, J., and Kalergis, A. M. (2018). Implications of Macrophage Polarization in Autoimmunity. *Immunology* 154, 186–195. doi:10.1111/IMM.12910

Gong, M., Zhuo, X., and Ma, A. (2017). STAT6 Upregulation Promotes M2 Macrophage Polarization to Suppress Atherosclerosis. *Med. Sci. Monit. Basic Res.* 23, 240–249. doi:10.12659/msmbr.904014

Gonzalez, a. G., Naldi, A., Sánchez, L., Thieffry, D., and Chaouiya, C. (2006). GINsim: a Software Suite for the Qualitative Modelling, Simulation and Analysis of Regulatory Networks. *Biosystems* 84, 91–100. doi:10.1016/j.biosystems.2005.10.003

Gordon, S., and Martinez, F. O. (2010). Alternative Activation of Macrophages: Mechanism and Functions. *Immunity* 32, 593–604. doi:10.1016/j.immuni.2010.05.007

Grinberg, S., Hasko, G., Wu, D., and Leibovich, S. J. (2009). Suppression of PLCβ2 by Endotoxin Plays a Role in the Adenosine A2A Receptor-Mediated Switch of Macrophages from an Inflammatory to an Angiogenic Phenotype. *Am. J. Pathol.* 175, 2439–2453. doi:10.2353/ajpath.2009.090290

Grosjean, A., Venteclef, N., and Dalmas, E. (2021). Understanding the Heterogeneity and Functions of Metabolic Tissue Macrophages. *Semin. Cell Developmental Biol.* 119, 130–139. doi:10.1016/j.semcdb.2021.09.002

Hamilton, J. A. (2019). GM-CSF-Dependent Inflammatory Pathways. *Front. Immunol.* 10. doi:10.3389/fimmu.2019.02055

Huang, X., Li, Y., Fu, M., and Xin, H.-B. (2018). Polarizing Macrophages *In Vitro*. *Methods Mol. Biol.* 1784, 119–126. doi:10.1007/978-1-4939-7837-3_12

Hutchins, A. P., Diez, D., and Miranda-Saavedra, D. (2013). The IL-10/STAT3-mediated Anti-inflammatory Response: Recent Developments and Future Challenges. *Brief. Funct. Genomics* 12, 489–498. doi:10.1093/bfgp/elt028

Kauffman, S. A. (1969). Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets. *J. Theor. Biol.* 22, 437–467. doi:10.1016/0022-5193(69)90015-0

Kim, H. S., Kim, D. C., Kim, H. M., Kwon, H. J., Kwon, S. J., Kang, S. J., et al. (2015). STAT1 Deficiency Redirects IFN Signalling toward Suppression of TLR Response through a Feedback Activation of STAT3. *Sci. Rep.* 5 (1), 13414–13415. doi:10.1038/srep13414

Kumar, V. (2019). Macrophages: The Potent Immunoregulatory Innate Immune Cells. *Macrophage Activation - Biol. Dis.* doi:10.5772/INTECHOPEN.88013

Lawrence, T., and Natoli, G. (2011). Transcriptional Regulation of Macrophage Polarization: Enabling Diversity with Identity. *Nat. Rev. Immunol.* 11, 750–761. doi:10.1038/nri3088

Lehtonen, A., Matikainen, S., Miettinen, M., and Julkunen, I. (2002). Granulocyte-macrophage colony-stimulating Factor (GM-CSF)-induced STAT5 Activation and Target-Gene Expression during Human Monocyte/macrophage Differentiation. *J. Leukoc. Biol.* 71, 511–519. doi:10.1189/jlb.71.3.511

Leibovich, S. J., Chen, J.-F., Pinhal-Enfield, G., Belem, P. C., Elson, G., Rosania, A., et al. (2002). Synergistic Up-Regulation of Vascular Endothelial Growth Factor Expression in Murine Macrophages by Adenosine A2A Receptor Agonists and Endotoxin. *Am. J. Pathol.* 160, 2231–2244. doi:10.1016/S0002-9440(10)61170-4

Li, X., Jolly, M. K., George, J. T., Pienta, K. J., and Levine, H. (2019). Computational Modeling of the Crosstalk between Macrophage Polarization and Tumor Cell Plasticity in the Tumor Microenvironment. *Front. Oncol.* 9, 10. doi:10.3389/fonc.2019.00010

Liquitaya-Montiel, A. J., and Mendoza, L. (2018). Dynamical Analysis of the Regulatory Network Controlling Natural Killer Cells Differentiation. *Front. Physiol.* 9, 1029. doi:10.3389/fphys.2018.01029

Liu, L., Zhao, Y., Xie, K., Sun, X., Jiang, L., Gao, Y., et al. (2014). Estrogen Inhibits LPS-Induced IL-6 Production in Macrophages Partially via the Nongenomic Pathway. *Immunological Invest.* 43, 693–704. doi:10.3109/08820139.2014.917095

Locati, M., Curtale, G., and Mantovani, A. (2020). Diversity, Mechanisms, and Significance of Macrophage Plasticity. *Annu. Rev. Pathol. Mech. Dis.* 15, 123–147. doi:10.1146/annurev-pathmechdis-012418-012718

Lucas, M., Zhang, X., Prasanna, V., and Mosser, D. M. (2005). ERK Activation Following Macrophage FcγR Ligation Leads to Chromatin Modifications at the IL-10 Locus. *J. Immunol.* 175, 469–477. doi:10.4049/jimmunol.175.1.469

Luo, Y., Pollard, J. W., and Casadevall, A. (2010). Fcγ Receptor Cross-Linking Stimulates Cell Proliferation of Macrophages via the ERK Pathway. *J. Biol. Chem.* 285, 4232–4242. doi:10.1074/jbc.M109.037168

Ma, X., Yan, W., Zheng, H., Du, Q., Zhang, L., Ban, Y., et al. (2015). Regulation of IL-10 and IL-12 Production and Function in Macrophages and Dendritic Cells. *F1000Res* 4, 1465 41465. doi:10.12688/f1000research.7010.1

Majai, G., Sarang, Z., Csomós, K., Zahuczky, G., and Fésüs, L. (2007). Pparγ-dependent Regulation of Human Macrophages in Phagocytosis of Apoptotic Cells. *Eur. J. Immunol.* 37 (5), 1343–1354. doi:10.1002/eji.200636398

Martinez-Sanchez, M. E., Enciso, J., and Arias Del Angel, J. (2018). BoolNetPerturb: Perturb Boolean Networks. *R package version 0.2. GitHub*. Available at: https://github.com/mar-esther23/boolnet-perturb.git

Martinez-Sanchez, M. E., Mendoza, L., Villarreal, C., and Alvarez-Buylla, E. R. (2015). A Minimal Regulatory Network of Extrinsic and Intrinsic Factors Recovers Observed Patterns of CD4+ T Cell Differentiation and Plasticity. *Plos Comput. Biol.* 11, e1004324. doi:10.1371/journal.pcbi.1004324

Matsuyama, T., Kubli, S. P., Yoshinaga, S. K., Pfeffer, K., and Mak, T. W. (2020). An Aberrant STAT Pathway Is central to COVID-19. *Cell Death Differ* 27, 3209–3225. doi:10.1038/s41418-020-00633-7

Mendoza-Coronel, E., and Ortega, E. (2017). Macrophage Polarization Modulates FcγR- and CD13-Mediated Phagocytosis and Reactive Oxygen Species Production, Independently of Receptor Membrane Expression. *Front. Immunol.* 8, 303. doi:10.3389/fimmu.2017.00303

Merad, M., and Martin, J. C. (2020). Pathological Inflammation in Patients with COVID-19: a Key Role for Monocytes and Macrophages. *Nat. Rev. Immunol.* 20, 355–362. doi:10.1038/s41577-020-0331-4

Muñoz-Rojas, A. R., Kelsey, I., Pappalardo, J. L., Chen, M., and Miller-Jensen, K. (2021). Co-stimulation with Opposing Macrophage Polarization Cues Leads to Orthogonal Secretion Programs in Individual Cells. *Nat. Commun.* 12, 301. doi:10.1038/s41467-020-20540-2

Müssel, C., Hopfensitz, M., and Kestler, H. A. (2010). BoolNet-an R Package for Generation, Reconstruction and Analysis of Boolean Networks. *Bioinformatics* 26, 1378–1380. doi:10.1093/bioinformatics/btq124

Naldi, A., Carneiro, J., Chaouiya, C., and Thieffry, D. (2010). Diversity and Plasticity of Th Cell Types Predicted from Regulatory Network Modelling. *Plos Comput. Biol.* 6, e1000912. doi:10.1371/journal.pcbi.1000912

Palma, A., Jarrah, A. S., Tieri, P., Cesareni, G., and Castiglione, F. (2018). Gene Regulatory Network Modeling of Macrophage Differentiation Corroborates the Continuum Hypothesis of Polarization States. *Front. Physiol.* 9, 1659. doi:10.3389/fphys.2018.01659

Park, B. S., Song, D. H., Kim, H. M., Choi, B.-S., Lee, H., and Lee, J.-O. (2009). The Structural Basis of Lipopolysaccharide Recognition by the TLR4-MD-2 Complex. *Nature* 458, 1191–1195. doi:10.1038/nature07830

Petrina, M., Martin, J., and Basta, S. (2021). Granulocyte Macrophage colony-stimulating Factor Has Come of Age: From a Vaccine Adjuvant to Antiviral Immunotherapy. *Cytokine Growth Factor. Rev.* 59, 101–110. doi:10.1016/j.cytogfr.2021.01.001

Ramírez, C., and Mendoza, L. (2018). Phenotypic Stability and Plasticity in GMP-Derived Cells as Determined by Their Underlying Regulatory Network. *Bioinformatics* 34, 1174–1182. doi:10.1093/bioinformatics/btx736

Ramirez, R., Herrera, A. M., Ramirez, J., Qian, C., Melton, D. W., Shireman, P. K., et al. (2019). Deriving a Boolean Dynamics to Reveal Macrophage Activation

with *In Vitro* Temporal Cytokine Expression Profiles. *BMC Bioinformatics* 20, 725. doi:10.1186/s12859-019-3304-5

Saez-Rodriguez, J., Simeoni, L., Lindquist, J. A., Hemenway, R., Bommhardt, U., Arndt, B., et al. (2007). A Logical Model Provides Insights into T Cell Receptor Signaling. *Plos Comput. Biol.* 3, e163–1590. doi:10.1371/journal.pcbi.0030163

Sica, A., and Mantovani, A. (2012). Macrophage Plasticity and Polarization: *In Vivo* Veritas. *J. Clin. Invest.* 122, 787–795. doi:10.1172/JCI59643

Tarique, A. A., Logan, J., Thomas, E., Holt, P. G., Sly, P. D., and Fantino, E. (2015). Phenotypic, Functional, and Plasticity Features of Classical and Alternatively Activated Human Macrophages. *Am. J. Respir. Cell Mol Biol* 53 (5), 676–688. doi:10.1165/rcmb.2015-0012oc

Vannella, K. M., and Wynn, T. A. (2017). Mechanisms of Organ Injury and Repair by Macrophages. *Annu. Rev. Physiol.* 79, 593–617. doi:10.1146/annurev-physiol-022516-034356

Varol, C., Mildner, A., and Jung, S. (2015). Macrophages: Development and Tissue Specialization. *Annu. Rev. Immunol.Annual Reviews)* 33, 643–675. doi:10.1146/annurev-immunol-032414-112220

Wang, L. x., Zhang, S. x., Wu, H. j., Rong, X. l., and Guo, J. (2019). M2b Macrophage Polarization and its Roles in Diseases. *J. Leukoc. Biol.* 106, 345–358. doi:10.1002/JLB.3RU1018-378RR

Wang, Y., Van Boxel-Dezaire, A. H. H., Cheon, H., Yang, J., and Stark, G. R. (2013). STAT3 Activation in Response to IL-6 Is Prolonged by the Binding of IL-6 Receptor to EGF Receptor. *Proc. Natl. Acad. Sci. U.S.A.* 110, 16975–16980. doi:10.1073/pnas.1315862110

Weber, A., Wasiliew, P., and Kracht, M. (2010). Interleukin-1 (IL-1) Pathway. *Sci. Signal.* 3, cm1. doi:10.1126/scisignal.3105cm1

Wilson, H. M. (2014). SOCS Proteins in Macrophage Polarization and Function. *Front. Immunol.* 5, 357. doi:10.3389/fimmu.2014.00357

Yoshimura, A., Naka, T., and Kubo, M. (2007). SOCS Proteins, Cytokine Signalling and Immune Regulation. *Nat. Rev. Immunol.* 7, 454–465. doi:10.1038/nri2093

Zhang, Y., Liu, S., Liu, J., Zhang, T., Shen, Q., Yu, Y., et al. (2009). Immune Complex/Ig Negatively Regulate TLR4-Triggered Inflammatory Response in Macrophages through FcγRIIb-dependent PGE2 Production. *J. Immunol.* 182, 554–562. doi:10.4049/jimmunol.182.1.554

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Ordaz-Arias, Díaz-Alvarez, Zúñiga, Martinez-Sánchez and Balderas-Martínez. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 27 April 2022
doi: 10.3389/fmolb.2022.855735

Combination of Genome-Scale Models and Bioreactor Dynamics to Optimize the Production of Commodity Chemicals

Jorge Lázaro[1], Giorgio Jansen[2], Yiheng Yang[3], Mario A. Torres-Acosta[3], Gary Lye[3], Stephen G. Oliver[2] and Jorge Júlvez[1]**

[1]Department of Computer Science and Systems Engineering, University of Zaragoza, Zaragoza, Spain, [2]Cambridge Systems Biology Centre and Department of Biochemistry, University of Cambridge, Cambridge, United Kingdom, [3]Department of Biochemical Engineering, University College London, London, United Kingdom

OPEN ACCESS

Edited by:
Daniel Garrido,
Pontificia Universidad Católica de Chile, Chile

Reviewed by:
Francis Mairet,
Institut Français de Recherche pour l'Exploitation de la Mer (IFREMER), France
Daniel Gamermann,
Federal University of Rio Grande do Sul, Brazil

***Correspondence:**
Stephen G. Oliver
sgo24@cam.ac.uk
Jorge Júlvez
julvez@unizar.es

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: 15 January 2022
Accepted: 07 March 2022
Published: 27 April 2022

Citation:
Lázaro J, Jansen G, Yang Y, Torres-Acosta MA, Lye G, Oliver SG and Júlvez J (2022) Combination of Genome-Scale Models and Bioreactor Dynamics to Optimize the Production of Commodity Chemicals. Front. Mol. Biosci. 9:855735. doi: 10.3389/fmolb.2022.855735

The current production of a number of commodity chemicals relies on the exploitation of fossil fuels and hence has an irreversible impact on the environment. Biotechnological processes offer an attractive alternative by enabling the manufacturing of chemicals by genetically modified microorganisms. However, this alternative approach poses some important technical challenges that must be tackled to make it competitive. On the one hand, the design of biotechnological processes is based on trial-and-error approaches, which are not only costly in terms of time and money, but also result in suboptimal designs. On the other hand, the manufacturing of chemicals by biological processes is almost exclusively carried out by batch or fed-batch cultures. Given that batch cultures are expensive and not easy to scale, technical means must be developed to make continuous cultures feasible and efficient. In order to address these challenges, we have developed a mathematical model able to integrate in a single model both the genome-scale metabolic model for the organism synthesizing the chemical of interest and the dynamics of the bioreactor in which the organism is cultured. Such a model is based on the use of Flexible Nets, a modeling formalism for dynamical systems. The integration of a microscopic (organism) and a macroscopic (bioreactor) model in a single net provides an overall view of the whole system and opens the door to global optimizations. As a case study, the production of citramalate with respect to the substrate consumed by *E. coli* is modeled, simulated and optimized in order to find the maximum productivity in a steady-state continuous culture. The predicted computational results were consistent with the wet lab experiments.

Keywords: flexible nets, modeling formalisms, model integration, multi-scale models, genome-scale models, commodity chemicals, continuous culture, citramalate production

1 INTRODUCTION

Methyl methacrylate (MMA) is a volatile synthetic chemical used mainly in the preparation of acrylic emulsion and extrusion resins. Polymers and co-polymers containing methyl methacrylate are used as solvents, adhesives, sealants, leather and paper coatings, inks, textiles, dental prothesis, etc.

There are 17 different routes widely used in industry that end up synthesizing MMA. The main problem with these routes is that all the precursor molecules (ethylene, propyne, propylene, tert-butyl alcohol, isobutene and isobutane) have their origin in non-renewable sources such as petroleum and natural gas whose extraction is highly damaging for ecosystems (Sugiyama et al., 2009).

An alternative approach consists of considering citramalate, a precursor for the synthesis of MMA, which is produced by *Methanocaldococcus jannaschii*. The production of citramalate in *M. jannaschii* is due to the presence of the gene *cimA* which encodes the enzyme citramalate synthase (EC: 2.3.1.182). This enzyme catalyses the reaction in which one molecule of acetyl-CoA, one molecule of pyruvate and one molecule of water react to produce one molecule of (3 R)-citramalate, one molecule of CoA and liberating a proton (UniProt Consortium, 2019):

$$acetyl - CoA + pyruvate + H_2O \rightarrow (3R)citramalate + CoA + H^+ \quad (1)$$

One of the most difficult tasks when carrying out the design of a biological experiment is setting the conditions and parameters that have to be tracked during the experiment. Computational models can help overcome these difficulties by providing the researchers with guidance when designing experiments in the wet lab, thus avoiding costly trial-and-error approaches.

In Webb et al. (2018), researchers could reach an efficient bioproduction of citramalic acid by a genetically engineered *E. coli* strain which included the gene *CimA*. The fact that the cell culture operated in fed-batch mode suggests that the production could be optimized by changing to continuous culture. In a continuous culture, a steady state is reached when the macroscopic variables of the tank remain constant over time. The complexity of continuous cultures lies in the fact that identical macroscopic conditions may trigger multiple steady states. The potential steady states of an *E. coli* continuous culture are characterized in Fernandez-de Cossio-Diaz et al. (2017).

Flux Balance Analysis (FBA) Orth et al. (2010) has proven to be an extremely useful approach to analyze steady states of genome-scale constraint-based models. FBA assumes the attainment of a steady state of intracellular metabolite concentrations to compute reactions fluxes by means of a linear programming problem. In addition to analyzing the potential steady states, a challenging problem when modeling and optimizing a continuous culture consists of linking the microscopic variables of the genome-scale model with the macroscopic variables of the bioreactor, e.g., metabolite concentrations out of the cells.

A variation of the FBA approach, called *Dynamic Flux Balance Analysis* (DFBA) (Mahadevan et al., 2002a), can be used to couple intracellular metabolism with the dynamics of the extracellular metabolite concentrations. DFBA has been applied for the production of several metabolites. In Flassig et al. (2016), the production of β-carotene in green microalgae was optimized for a fed-batch continuous culture. The work in Hanly et al. (2013) validated and optimized a yeast dynamic flux balance model in order to determine the optimum conditions that maximize the production of ethanol in a batch culture of *S. cerevisiae*. Two approaches to predict batch growth of *E. coli* based on DFBA are introduced in Mahadevan et al. (2002b). DFBA was applied as well in (Meadows et al. (2010)) with the aim of simulating simultaneous acetate and glucose consumption and evaluate the behaviour of *E. coli* cells in different types of media. Although DFBA has been applied successfully in many areas, it has some limitations as it assumes quasi-steady-state conditions (Reimers and Reimers,2016) and has been used almost exclusively on batch and fed-batch cultures.

Other methods not based on DFBA, such as k-OptForce, have been used to integrate kinetics in constraint-based models. For instance, the optimization of the production of L-serine in mutant *E. coli* and triacetic acid lactone in mutant *S. cerevisae* were performed in Chowdhury et al. (2014). K-OptForce uses kinetic rate expressions to redistribute fluxes in the metabolic network, instead of relying on surrogate fitness functions such as biomass maximization. For additional information on this topic, a review of efforts to integrate kinetic information in constraint-based models can be found in Kim et al. (2018).

In contrast to the previous works, we propose the use of Flexible Nets (FNs) Júlvez et al., 2018), a modeling framework that produces analytical models that can be represented graphically and that are well suited for analysis and optimization, in order to design an overall computational model that combines both the bioreactor dynamics and the metabolic network of the cultured organism. In addition to facilitating the integration of a macroscopic and a microscopic model, FNs can accommodate uncertain parameters and can approximate non-linear dynamics. In particular, constraint-based models Sigmarsdóttir et al. (2020) of metabolic networks can be straightforwardly mimicked and analyzed by FNs. Notice that such models do not account for the concentration of species, and loose flux bounds are usually associated with the reactions. Moreover, FNs can also model the differential equations that determine the dynamics of the bioreactor variables, e.g., cell density, nutrient supply, and metabolite concentration. In this way, FNs can integrate, in a seamless model, both the genome-scale model of the cultured microorganism and the bioreactor dynamics. We show how such an integrated model can be developed and exemplify the process through the modeling and simulation of a system that produces citramalate by a genetically modified *E. coli* culture. The optimization of the model obtained with respect to citramalate productivity provides the optimal settings, i.e., intracellular fluxes and bioreactor parameters, that maximize the productivity of citramalate in a steady-state continuous culture.

2 MATERIALS AND METHODS

2.1 Flexible Nets

Flexible Nets (FNs) is a modeling formalism for dynamic systems inspired by Petri Nets, see Murata (1989); Silva, 1993) for a gentle introduction. FNs aim to capture the

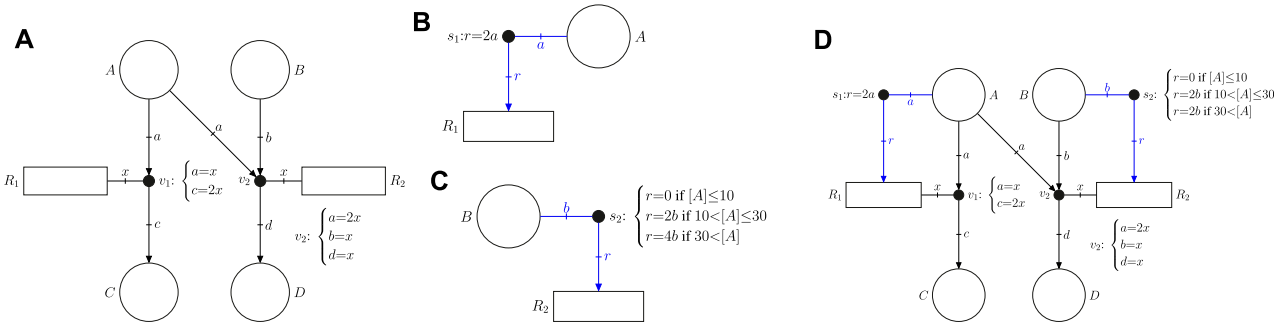


FIGURE 1 | (A) Event net modeling the stoichiometry of reactions R_1: $A \rightarrow 2\ C$ and $R_2 2A + B \rightarrow D$. **(B)** Intensity net producing a speed in R_1 proportional to the concentration [A]. **(C)** Guarded intensity net producing a speed in R_2 equal to 0 if [A] is below 10, 2 [B] if [A] is between 10 and 30, and 4 [B] otherwise. **(D)** Flexible net combining the event net in **(A)** and the intensity nets in **(B)** and **(C)**.

relationship between the state and the processes of a given dynamic system by means of two interconnected nets: the event net and the intensity net. On the one hand, the event net models how the processes modify the state variables. On the other hand, the intensity net models how the state variables determine the speeds of the processes. In contrast to Petri nets, both the event and the intensity nets are tripartite graphs which have three types of vertices: places, transitions and handlers. The handlers of the event net are called event handlers, and the handlers of the intensity net are called intensity handlers. Places (which are depicted as circles) are associated with metabolites and transitions (which are depicted as rectangles) are associated with reactions. Event handlers (which are depicted as dots) capture the change of concentration of metabolites produced by reactions. Intensity handlers (also depicted as dots) model how the concentrations of metabolites determine the speeds of reactions. Although event and intensity handlers can be distinguished by the net elements to which they are connected, for clarity the arcs and edges of event handlers will be drawn in black and those of intensity handlers in blue.

As an example, the event net in **Figure 1A** has four places {A, B, C, D}, two reactions {R_1, R_2} and two event handlers {v_1, v_2}. Such a net models the following reactions:

$$R_1: \quad A \rightarrow 2\ C$$
$$R_2: \quad 2A + B \rightarrow D$$

The stoichiometry of the reactions is modeled by the equalities associated with the event handlers. In particular, the equalities $a = x$ and $c = 2x$ of v_1 imply that each occurrence of reaction R_1 consumes one unit of metabolite A and produces two units of metabolite C (such units usually refer to concentrations). On the other hand, the equalities $a = 2x$, $b = x$, and $d = x$ of v_2 mean that each occurrence of R_2 consumes 2 units of A, 1 unit of B and produces 1 unit of D.

The event net in **Figure 1A** does not establish any dynamics, it just models the stoichiometry. The dynamics of reactions can be specified by the intensity net. For instance, the intensity net in **Figure 1B** specifies the speed of reaction R_1 as twice that of the concentration of A, see equation $r = 2a$ associated with the intensity handler s_1. In addition to equalities, intensity handlers can be associated with inequalites to model uncertainty, e.g., if $1.8a \leq r \leq 2.2a$ was associated with s_1 then the speed of R_1 could be any value in the interval [1.8 [A], 2.2 [A]] where [A] is the concentration of A.

Moreover, several sets of equalities and inequalities can be associated with the same intensity handler. If this is the case, the set of inequalities that rules the reaction dynamics is determined by the concentrations of the system. The intensity net in **Figure 1C** associates three different equalities with the intensity handler s_2 which imply that the speed of R_2 is 0 if [A] is below 10, 2 [B] if [A] is between 10 and 30, and 4 [B] otherwise. Intensity handler s_2 is said to be guarded, it has three guards (or regions) that can determine the speed of R_2. Guarded handlers can be exploited to approximate non-linear kinetics of reactions.

The event net (**Figure 1A**) can be combined with the intensity nets (**Figure 1B** and **Figure 1C**) to produce an FN (**Figure 1D**) which models both the stoichiometry and dynamics of the system.

2.2 FNs to Model Constraint-Based Models

A *constraint-based model* (Varma and Palsson, 1994) can be expressed as a tuple {$\mathcal{R}$, $\mathcal{M}$, $\mathcal{S}$, L, U} where $\mathcal{R}$ is the set of reactions, $\mathcal{M}$ is the set of metabolites, $\mathcal{S} \in \mathbb{R}^{|\mathcal{M}|\times|\mathcal{R}|}$ is the stoichiometric matrix, and $L, U \in \mathbb{R}^{|\mathcal{R}|}$ are lower and upper flux bounds of the reactions (notice that very loose flux bounds can be assigned when no kinetic information is available). The concentrations of metabolites are usually disregarded in constraint-based models, being the main focus of most analyses on the fluxes of reactions. This section shows how constraint-based models can be expressed graphically and analyzed numerically in a straightforward way by FNs.

Consider the constraint-based model defined by **Table 1**. It consists of four reactions together with their corresponding flux bounds. The FN in **Figure 2** models such a constraint-based model. The net has one place per metabolite, one transition per reaction, and one event handler per reaction. The equalities associated with the event handlers model the stoichiometry of the reactions. Since constraint-based models do not account for the concentrations of metabolites, the fluxes of reactions cannot depend on concentrations. Hence, the corresponding FN does not have intensity handlers. The range of potential fluxes of reactions is modeled by a parameter λ_0 that is associated with each transition, e.g. λ_0 [R_1] = 5 in R_1 means that the flux of R_1 will always be equal to 5 $mmol\ gDW^{-1}h^{-1}$, and $0 \leq \lambda_0$ [R_2] ≤

TABLE 1 | Simple constraint-based model of four reactions with lower and upper flux bounds ($mmol\ gDW^{-1}h^{-1}$). The model is represented graphically by the FN in **Figure 2**

Reaction	Lower bound	Upper bound
R_1: $\varnothing \rightarrow A$	5	5
R_2: $\varnothing \rightarrow C$	0	20
R_3: $A \leftrightarrow 2B$	–500	1,000
R_4: $B + C \rightarrow \varnothing$	0	1,000

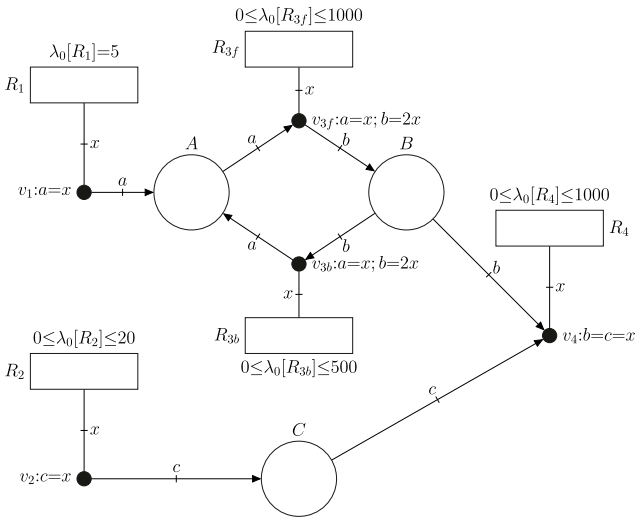


FIGURE 2 | FN modeling the constraint based model expressed by the reactions in **Table 1**. The flux bounds of the reactions are modeled by the default intensities (or speeds) λ_0 of transitions. Reversible reactions like R_3 are unfolded into a forward and a backward reaction with non-negative flux.

20 in R_2 implies that the flux of R_2 can be any quantity between 0 and 20 $mmol\ gDW^{-1}h^{-1}$.

The fluxes of transitions must be non-negative in FNs. Thus, the modeling of reversible reactions like R_3 requires its unfolding into two reactions, a forward reaction R_{3f} and a backward reaction R_{3b} with appropriate limits for their parameters λ_0 (Júlvez and Oliver, 2020b). If no λ_0 is explicitly associated with a reaction r, then it is assumed that $\lambda_0\,[r] = 0$. In general, the speed of r is equal to $\lambda_0\,[r]$ plus the intensities provided by the intensity handlers to which it is connected (see **Figure 1B**).

FNs can be analyzed by building a set of mathematical constraints that the state of the system necessarily satisfies (Júlvez et al., 2018). The association of such constraints with an objective function of interest results in a programming problem whose solution yields a theoretical optimum. For instance, if the objective function for the FN in **Figure 2** is the maximization of the flux of R_4 in the steady state, then the solution of the programming problem would be 10 $mmol\ gDW^{-1}h^{-1}$ which is the theoretical maximum steady-state flux of R_4. In addition to the flux of R_4, fluxes for the rest of reactions are obtained. In this particular case, this approach is equivalent to performing Flux Balance Analysis (FBA) Orth et al. (2010) on the constraint-based model.

For the production of citramalate, the constraint-based model of the organism *Escherichia coli* strain K-12 MG1655 (Webb et al. (2018) was considered. The model is named iJO1366 in the BiGG repository database (Orth et al. (2010)) and has 1805 metabolites, 2,583 reactions, and 1,367 genes. The reaction in **Eq. 1** was added to this model, simulating a transgenic *E. coli* strain capable of synthesizing citramalate.

The transformation of the resulting constraint-based model into an FN can be carried out by following the approach to obtain the net in **Figure 2**. Such an approach is performed automatically by the *cobra2fn* module of the *Python* tool *fnyzer* (Júlvez and Oliver, 2020a).

2.3 FNs to Model Bioreactor Dynamics

The macroscopic model of the bioreactor consists of three parts (see **Figure 3**) the "Reservoir", which contains the fresh sterile medium and supplies the cell culture with the essential nutrients for cell survival; the "Tank", where the cell culture is placed, and the "Effluent" which clears away the accumulated products and some of the cells in the tank.

The dynamics of the bioreactor variables, which are named macroscopic variables, are determined by differential equations (Fernandez-de Cossio-Diaz et al., 2017). The equation that expresses the evolution of the cell density in the tank is:

$$\frac{dX}{dt} = (\mu - D)X \tag{2}$$

where X ($gDWL^{-1}$) is the cell density in the tank, μ is the effective cell growth rate (h^{-1}), and D is the dilution rate (h^{-1}), which is the

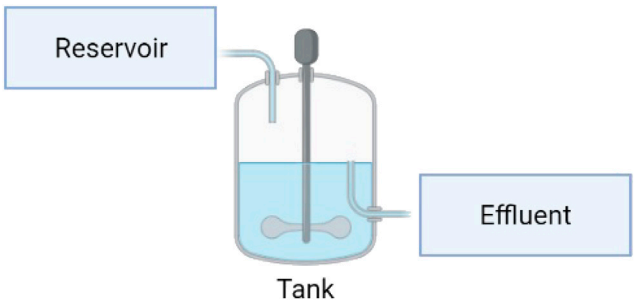


FIGURE 3 | Sketch of a bioreactor in continuous culture mode. The nutrient supply and the removal of toxic and cell products are executed at the same time uninterruptedly. The three main compartments are: reservoir, tank and effluent.

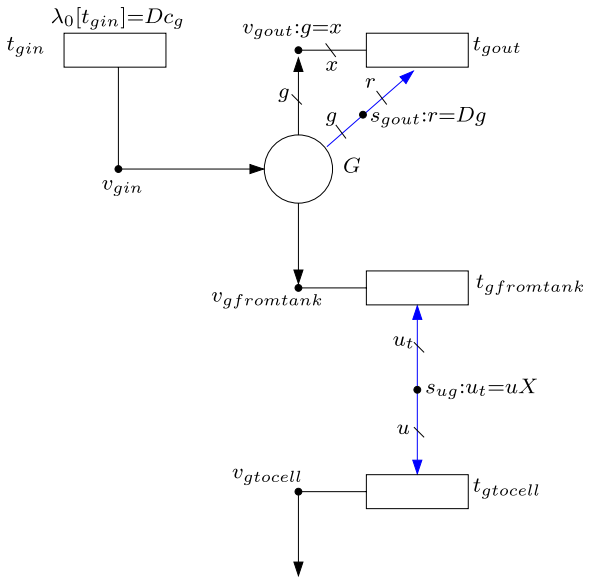


FIGURE 4 | FN modeling the differential **Eq. 4**

rate at which culture fluid is replaced divided by the culture volume.

The evolution of the concentration of a given metabolite, *i*, in the tank is given by:

$$\frac{ds_i}{dt} = (c_i - s_i)D - u_i X \tag{3}$$

where c_i is the concentration of the metabolite s in the medium (mM), s_i is the concentration of the metabolite in the tank (mM), D is the dilution rate (h^{-1}), u_i is the specific uptake rate of the metabolite by the cells ($mmol\ gDW^{-1}\ h^{-1}$), and X is the cell density in the tank ($gDW\ L^{-1}$). If $u_i > 0$ the metabolite is consumed by the cell, otherwise ($u_i < 0$) the metabolite is secreted from the cell.

For the particular case of a system in which glucose, denoted as metabolite *g*, is consumed by an *E. coli* culture, **Eq. 3** for the concentration of glucose in the tank becomes:

$$\frac{ds_g}{dt} = \left(c_g - s_g\right)D - u_g X \tag{4}$$

where c_g is the concentration of glucose in the supply medium, s_g is the concentration of glucose in the tank and, u_g is the glucose uptake flux by the cell (which is a positive value).

The variation of a given product, e.g., citramalate, denoted as *c*, is derived from **Eq. 3** as:

$$\frac{ds_c}{dt} = -s_c D - u_c X \tag{5}$$

where s_c is the citramalate concentration in the tank, and u_c is the citramalate secretion flux. Notice that, since citramalate is secreted from the cell, u_c is negative, and hence, $-u_c X$ is a positive contribution of citramalate to the tank.

The above differential equations can be modeled by FNs.[1] For instance, **Eq. 4** can be modeled by the FN in **Figure 4** where place

[1]For clarity, it is assumed that if the labels and equations of a handler are omitted (see handlers $v_{gfromtank}$ and $v_{gtocell}$ in **Figure 4**), then an equality among all the connected elements holds. For instance if the labels *b*, *c*, *x* and the equation $b = c = x$ are omitted in v_4 in **Figure 2**, the implicit meaning is that the stoichiometric weights of all metabolites involved in reaction R_4 are 1.

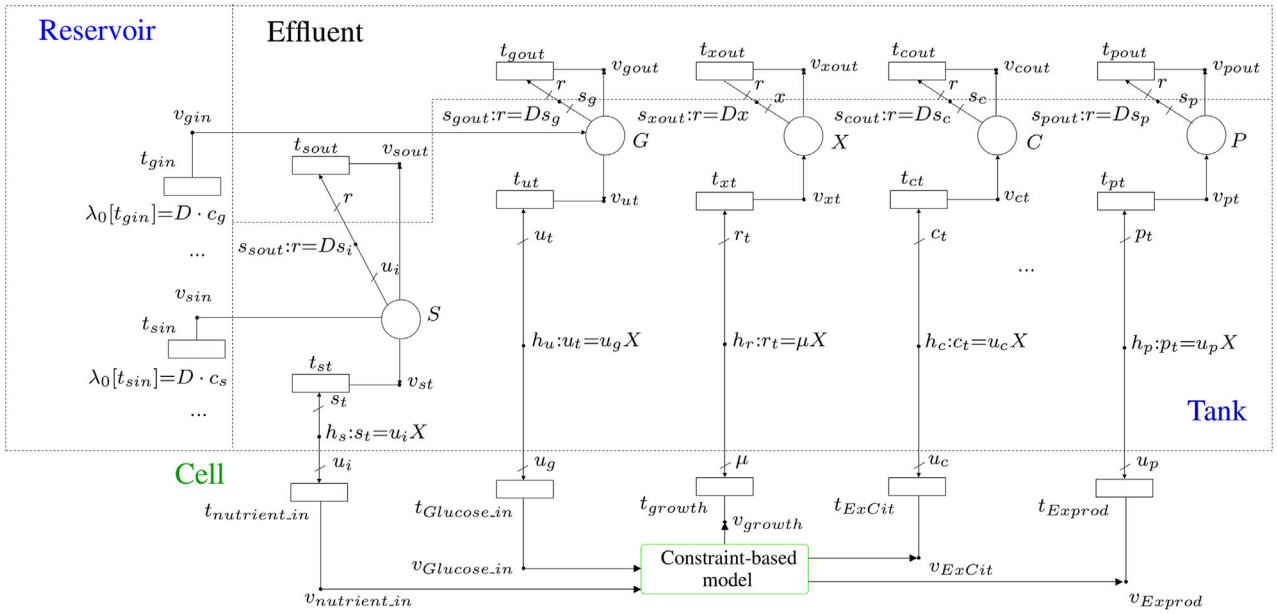


FIGURE 5 | FN modeling the overall production system by integrating the macroscopic model of the bioreactor and the microscopic model of the constrained-based model, MODEL1108160000 Orth et al. (2011) of the BioModels database Malik-Sheriff et al. (2020), of the cultured organism.

G accounts of the concentration of glucose in the tank. As established by **Eq. 4**, place G has one input flux and two output fluxes. The input flux comes from the reservoir, it is modeled by transition t_{gin}, and it is equal to $D \cdot c_g$. As this is a constant amount, no intensity handlers are needed, and the flux is modeled by the λ_0 associated with t_{gin}. The output flux modeled by $t_{gfromtank}$ represents the uptake rate of glucose by the cell culture, and it is equal to u_gX where u_g is the specific uptake rate and X is the cell biomass. This flux is produced by the intensity handler s_{ug}. Such an intensity handler scales by X the amount of glucose that is consumed by the cells, see equation $u_t = uX$ associated with s_{ug}. The output flux modeled by t_{gout} represents the glucose that leaves the tank without being captured by the cells. Such an output flux is equal to the dilution rate times the concentration of glucose in the tank, see equation Ds_g associated with s_{gout}: $r = Ds_g$.

Eq. 5 can be modeled similarly by FNs, see place C and the elements connected to it in **Figure 5**. In this case, there is one input flux and one output flux. The input flux comes from the cell (citramalate is produced by the culture), it is modeled by transition t_{ct} and it is equal to u_cX, see equation associated with h_c. The output flux corresponds to the amount of citramalate in the tank, once it has been released by the cell, that forms part of the effluent. Such a flux is modeled by t_{cout} and it is equal to s_cD, see equation associated with s_{cout}.

Finally, **Eq. 2** is modeled by the place X and the net elements connected to it. The input flux of X, i.e. the rate at which X increases, is modeled by t_{xt} and it is equal to the specific growth rate of the culture times the cell density, rX, see equation associated with h_r. The output flux of X, i.e. the rate at which X decreases, corresponds to the cells that are cleared away in the continuous culture, it is modeled by t_{xout} and is equal to Dx, see equation associated with s_{gout}.

2.4 Model Integration

The merger of the FNs that model the dynamics of metabolite concentrations in the bioreactor, see **Section 2.3**, and the FNs that model the constraint-based model of the metabolic network of the cultured organism, see **Section 2.2**, results in a single FN that models the overall production system, see **Figure 5**. The nutrients in the medium can be introduced by the transitions located in the reservoir compartment. In this compartment, as many transitions as metabolites in the medium are required. For example, transition t_{gin} accounts for the presence of glucose in the medium, and transition t_{sin} represents the presence of a given metabolite s in the medium.

The concentrations of metabolites in the bioreactor are modeled by the places in the tank compartment. In a general system, there will be as many places in the tank compartment as there are metabolites being tracked. In **Figure 5**, place S represents a generic nutrient, i.e., it is provided by the medium, it is consumed by the culture (see arrow going from S to v_{st}) and forms part of the effluent. On the other hand, place P represents a generic product, i.e., it is not provided by the medium, it is produced by the culture (see arrow going from S to v_{st}) and forms part of the effluent. Although the cell density, X, is not a metabolite, its evolution, see **Eq. 2**, can be modeled exactly as it was a product of the culture, see place X and the net elements connected to it (the increase of X is due to the biomass production which is determined by the growth rate).

For the particular system that produces citramalate by a genetically modified *E. coli* culture, glucose-limited conditions are assumed. Thus, in addition to the cell density modeled by place X with units $gDWL^{-1}$, the focus will be on the concentration of one nutrient (glucose which is modeled by place G with units mM), and the concentration of one product

(citramalate which is modeled by place *C* with units *mM*). Recall that the dynamics of *X* is ruled by **Eq. 2**, while *G* and *C* are ruled by **Eqs 4** and **5**, respectively. The microscopic model, i.e., constrained-based model, for the citramalate production system is given by the genome-scale metabolic model of *E. coli* strain K-12 (iJO1366), MODEL1108160000 Orth et al. (2011) of the BioModels database (Malik-Sheriff et al., 2020). This model was converted to an FN by the *fnyzer* tool (Júlvez and Oliver, 2020a).

The integration of the macroscopic and the microscopic models was possible thanks to the elements involved in the interface between the tank and the cell compartments. The main elements taking part in this connection are the intensity handlers h_u, h_r and h_c. These intensity handlers relate the macroscopic variables of the bioreactor with the exchange fluxes of the cell in such a way that each macroscopic flux equals X times the exchange flux of the cell, X being the cell density in the tank. The equations which model the interface between the cell and the tank are: $u_t = u_g X$, $r_t = \mu X$, and $c_t = u_c X$, which are associated with the intensity handlers h_u, h_r and h_c, respectively.

The aforementioned intensity handlers are graphically located at the interface between the cell and tank compartments, and each one acts as a bridge between two transitions: h_u connects $t_{Glucose_in}$ and t_{ut}, h_r connects t_{growth} and t_{xt}, and h_c connects t_{ExCit} and t_{ct}. In our model, which contains all the metabolic reactions of *E. coli* strain K-12 (iJO1366) combined with the reactions that allow citramalate production, $t_{Glucose_in}$ represented the glucose exchange reaction, t_{growth} is used for the biomass reaction and t_{ExCit} defined the citramalate exchange reaction.

2.5 Model Optimization

This section discusses the approximations that must be applied to the FN in **Figure 5** prior to its optimization (**Subsection 2.5.1**), as well as the type of objective function that is considered (**Subsection 2.5.2**).

2.5.1 Tackling Non-linearities

Notice that the equations associated with the intensity handlers at the interface between the macroscopic model of the tank and the microscopic model of the cell are not linear. For instance, the equation associated with h_u is $u_t = u_g X$ where both u_g (the uptake rate of glucose) and X (the cell density) are real variables. The optimization of a non-linear system is, in general, very demanding from a computational point of view. To overcome such a computational burden, non-linear equations can be approximated by piece-wise linear inequalities that are associated with intensity handlers. This approximation results in a guarded FN (see **Section 2.1**).

A non-linear equation such as $u_t = u_g X$ of h_u can be approximated piece-wise linearly by partitioning the state space of one of the real variables, e.g., X, into a number of regions and associating a linear inequality with each of the regions. Thus, h_u: $u_t = u_g X$ can be approximated by:

$$h_u: \begin{cases} X_{min} \cdot u_g \le u_t \le X_1 \cdot u_g & \text{if } X_{min} \le X < X_1 \\ X_1 \cdot u_g \le u_t \le X_2 \cdot u_g & \text{if } X_1 \le X < X_2 \\ X_2 \cdot u_g \le u_t \le X_3 \cdot u_g & \text{if } X_2 \le X < X_3 \\ \ldots, & \\ X_{n-1} \cdot u_g \le u_t \le X_{max} \cdot u_g & \text{if } X_{n-1} \le X \le X_{max} \end{cases} \quad (6)$$

where $X_{\min}$ and $X_{\max}$ are lower and upper bounds for the cell density, i.e., the cell density is known to be in the interval $[X_{\min}, X_{\max}]$ (notice that these bounds do not need to be tight).

The above approximation considers n regions, the first region is active if the cell density X is in the interval $[X_{\min}, X_1]$ (in general, the *ith* region is active if the cell density X is in the interval $[X_{i-1}, X_i]$). The values $X_1, \ldots, X_{n-1}$ do not need to be evenly separated, the only condition they must satisfy is $X_{\min} < X_1 < \ldots, < X_{n-1} < X_{\max}$. This way, one and only one region is active at any particular time. The region that is active determines the linear inequality that is used to produce intensity, i.e., if region i is active then the intensity produced by h_u can be any value in the interval $[X_{i-1} \cdot u_g, X_i \cdot u_g]$. Clearly, the higher the number of regions (and hence, the smaller the regions), the better the approximation to the original non-linear equation. Since a higher number of regions involves a longer run time, there is a trade-off between accuracy and computational cost. As discussed below, the number of regions was determined experimentally so that both the computational burden and the obtained precision are acceptable.

Notice that the previously defined regions can also be used to approximate the non-linear equations of the other handlers in the interface between the macroscopic and microscopic models because all include X in their equations. Given that the number of regions has a direct impact on the complexity of the programming problem which needs to be solved (the number of binary variables is linear in the number of regions), partitioning X instead of u_g is advantageous from a computational point of view.

The overall procedure to optimize an FN that integrates a bioreactor and a metabolic network is outlined in **Figure 6**. After integrating both models in a single FN, a set of mathematical constraints that represent necessary reachability conditions for the state of the system are derived. Such a set of constraints can be derived automatically by the *Python* tool *fnyzer* (Júlvez and Oliver, 2020a). The addition of an objective function to the constraints results in a mixed-integer linear programming (MILP) problem whose solution represents the theoretical optimum state that the system can achieve.

In order to speed up the model optimization, a non-guarded FN has been defined for each of the above regions, and each of these nets has been solved separately. For instance, region i determines an FN in which the value of X is constrained to the interval $[X_{i-1}, X_i]$ (this constraint will be part of the programming problem) and the inequalities $X_2 \cdot u_g \le u_t \le X_3 \cdot u_g$ are associated with h_u (similar inequalities are associated with the rest of handler in the interface). The programming problem associated with each of these nets is linear, and hence, can be solved very efficiently. The optimum solution of the original

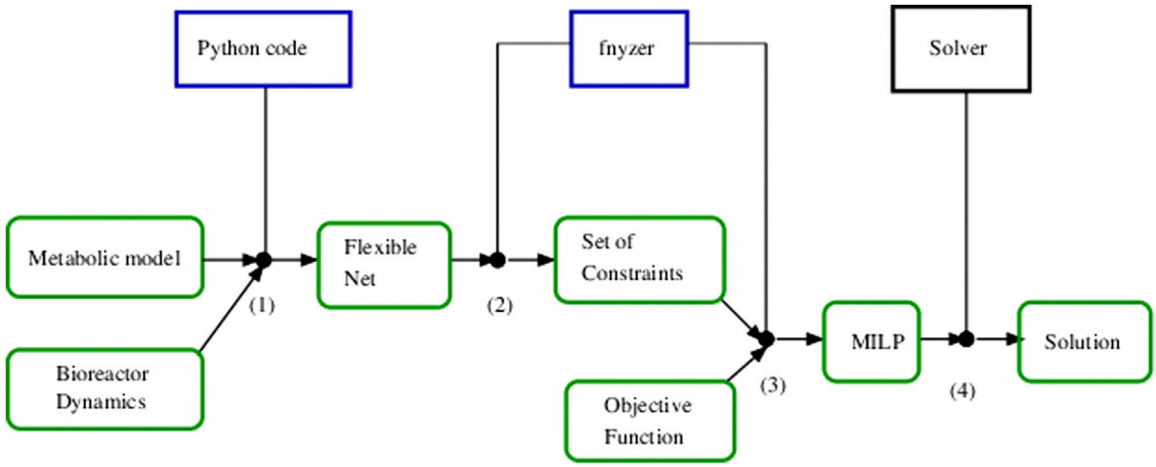


FIGURE 6 | Pipeline showing the steps performed to optimize the integrated model. **(1)** The metabolic model and the bioreactor dynamics are combined to generate a Flexible Net that integrates the macroscopic (dilution rate, substrate concentration, cell density) and microscopic variables (intracellular metabolite fluxes). **(2)** A set of mathematical constraints is derived from the net specification and **(3)** the objective function is selected. The tool fnyzer performs the generation of a mixed-integer linear programming problem according to the set of constraints and the objective function. **(4)** Finally, the MILP problem is easily solved by using a solver (e.g., CPLEX, Gurobi, GLPK) that computes the mathematical solution.

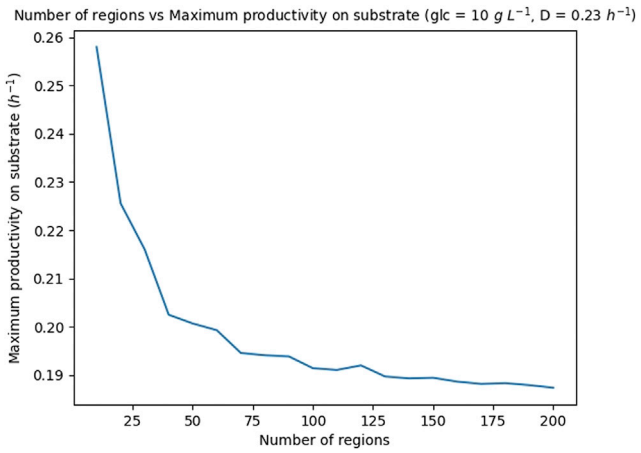


FIGURE 7 | Theoretical maximum productivity on substrate *PS* with respect to the number of regions in which the biomass is partitioned. The glucose concentration is set to 10 $g\ L^{-1}$ and the dilution rate to 0.23 h^{-1}.

guarded FN can be obtained straightforwardly by taking the maximum of all the computed objective values of the particular non-guarded FNs.

In order to partition the cell concentration X in an appropriate number of regions, the productivity on substrate (*PS*), see **Subsection 2.5.2**, was calculated repeatedly for different number of regions, ranging in the interval [10, 200], and fixed values of glucose concentration in the medium, 10 gL^{-1}, and dilution rate, 0.23 h^{-1}. The obtained maximum *PS* are shown in **Figure 7** and the CPU run-times are reported in **Supplementary Data S3**. Notice that after an initial sharp decrease, the productivity converges to a given value. On the other hand, the run-time of the simulations increases linearly with the number of regions (see reported run-times). Based on these results, it was decided to set the number of regions for the optimizations to 100, as this number provided a good trade-off between accuracy and run-time (the run-time to optimize the FN for a given glucose concentration and a given dilution rate is 564 s (9.4 min), see hardware features in **Supplementary Data S3**.

2.5.2 Optimizing the Productivity

Among the different objective functions that can be considered, we focus on two measures for the productivity of a culture in continuous mode: 1) volumetric productivity, and 2) productivity on substrate.

The volumetric productivity, a. k.a. space-time yield, accounts for the amount of product produced per liter and per hour, it will

be denoted *VP* and is expressed in grams of product per liter per hour, i.e., $g \cdot L^{-1} \cdot h^{-1}$. In terms of the FN in **Figure 5**, the volumetric productivity corresponds to the outgoing flux of product $flux(P)$, i.e. intensity of transition t_{pout}, expressed in $g \cdot L^{-1} \cdot h^{-1}$. The volumetric productivity would be the primary concern of a chemical engineer designing an industrial process, which needs to be both feasible and economically viable.

The productivity on substrate takes into account both the specific growth rate of the culture, μ (units h^{-1}), and the product yield coefficient, $Y_{P/S}$, where $Y_{P/S}$ denotes the number of grams of product that are produced per Gram of substrate fed into the tank, i.e. $Y_{P/S} = \frac{Product\ (g)}{Substrate\ (g)}$, and hence $Y_{P/S}$ is unitless. In particular, the productivity on substrate, which will be denoted *PS*, for a given net $\mathcal{N}$ is defined as:

$$PS(\mathcal{N}) = \mu \cdot Y_{P/S} \tag{7}$$

given that $Y_{P/S}$ is unitless, $PS(\mathcal{N})$ is expressed in h^{-1}. Notice that **Eq. 7** entails a trade-off between biomass formation and product production. Although the productivity on substrate is usually a secondary concern, it becomes more relevant as the cost of the substrate increases with respect to the selling price of the product. The productivity on substrate is also a useful metric for a synthetic biologist comparing the performance of different genetically engineered microbial strains.

In the following, the mathematical relation between volumetric productivity and productivity on substrate is explored. In a continuous culture, the yield $Y_{P/S}$ is equal to the grams of product produced per Gram of substrate per time unit, i.e. it can be expressed as $\frac{flux(P)}{flux(S)}$ where $flux(P)$ is the flux of produced product in $g \cdot L^{-1} \cdot h^{-1}$, and $flux(S)$ is the flux of provided substrate in $g \cdot L^{-1} \cdot h^{-1}$. Notice that $flux(S)$ is equal to Dc_s where D is the dilution rate in h^{-1} and c_s is the concentration of the substrate in the fresh medium in $g \cdot L^{-1}$. This way, the yield can be expressed as:

$$Y_{P/S} = \frac{flux(P)}{flux(S)} = \frac{flux(P)}{D \cdot c_s} \tag{8}$$

On the other hand, the cell density is assumed to be constant in a continuous culture, i.e., $\frac{dX}{dt} = 0$, and hence, **Eq. 2** implies that $\mu = D$. Thus, **Eq. 7** can be rewritten as:

$$PS(\mathcal{N}) = \mu \cdot Y_{P/S} = D \cdot \frac{flux(P)}{D \cdot c_s} = \frac{flux(P)}{c_s} = \frac{VP(\mathcal{N})}{c_s} \tag{9}$$

Thus, the volumetric productivity is equal to the productivity on substrate times the concentration of substrate in the medium. Therefore, for a given fixed concentration of substrate in the medium, c_s, optimizing the model to maximize productivity on substrate is equivalent to optimizing the model to maximize volumetric productivity. Moreover, for a given c_s, the linear objective function $flux(P)$ can be used to perform such an optimization. For the particular case of citramalate production, $flux(P)$ is given by the value of $\lambda[t_{cout}]$ (see **Figure 5**), and hence, the objective function will be the maximization of $\lambda[t_{cout}]$.

2.5.3 Fermentation Experiments

Continuous cultures were grown in a DASbox® Mini Bioreactor System (Eppendorf, Stevenage, UK). The *E. coli* strain used for all fermentation experiments was BW25113 Δ*ldhA* pET29a-*Cer*-BBaJ23119-RFP-*cimA3.7*. The *E. coli ldhA* deletion prevents lactate formation, improving flux towards citramalate (Webb et al. (2018)). The plasmid contains both the *cimA* gene, to enable citramalate production, and the *cer* gene to reduce loss of the plasmid through mis-segregation (Green et al. (2018)). Glucose-limited chemostat cultures (150 ml working volume) were grown at 37°C, with pH controlled to 7 and dissolved oxygen to >30%, in modified MS (Stephens and Dalton (1987)) medium (2 g L^{-1} KH2PO4, 2 ml L^{-1} trace metals solution (Vishniac and Santer (1957)), 0.25 ml L^{-1} antifoam polypropylene glycol, 4 g L^{-1} NH4Cl, 0.4 g L^{-1} MgSO4.7H2O). Biomass concentrations were determined by centrifuging measured samples from the fermenter into pre-weighed tubes, washing the pellets, and drying to constant weight at 100°C for 48 h. The supernatants from these samples were used to measure glucose, citramalate, and acetate concentrations. These analyses were performed using an UltiMate 3000 HPLC system (Thermo Fisher Scientific, Loughborough, UK) equipped with an Aminex HPX-87H ion-exclusion column (Bio-Rad, Hertfordshire, UK) and a RefractoMax520 RI detector (Knauer, Berlin, Germany). The mobile phase used was 0.1% (v/v) trifluoroacetic acid (TFA) in Milli-Q water.

3 RESULTS

The validation of the model was carried out by comparing several simulation runs of the designed FN, see **Figure 5**, to previously obtained experimental results. These experiments were carried out under glucose-limited conditions, using three different concentrations of glucose in the supplied medium (5, 10 and 50 gL^{-1}), and different dilution rates.

Table 2 reports the numerical results obtained both *in vivo*, and *in silico* by the FN model. The columns of the table are divided in three parts: the first part sets the experimental parameters, i.e., the glucose concentration in the medium and the dilution rate for the glucose-limited continuous cultures; the second part reports the *in vivo* experimental results; and the third part reports the *in silico* results. For each pair of experimental parameters (glucose in medium and dilution rate), the columns corresponding to the "*In vivo* results" report the cell density (column "Biomass"), the concentration of citramalate (column "Citramalate in tank"), and the concentration of glucose (column "Residual glucose") measured in the tank. In order to validate the model, the concentrations of citramalate and glucose in the tank are computed for each set of experimental parameters (glucose in medium and dilution rate) and measured biomass. The columns corresponding to the "*In silico* results" report the computed concentration of citramalate in the tank (column "Citramalate in tank"), the relative error of such predicted concentrations with respect to the measured *in vivo*

TABLE 2 | Data obtained after running the code that simulates the FN model implementing the experimental conditions and previous results.

Experimental parameters		*In vivo* results			*In silico* results		
Glucose in medium (gL^{-1})	**Dilution rate (h^{-1})**	**Biomass ($gDWL^{-1}$)**	**Citramalate in tank (*in vivo*) (gL^{-1})**	**Residual glucose (*in vivo*) (gL^{-1})**	**Citramalate in tank (*in silico*) (gL^{-1})**	**Citramalate relative error (%)**	**Residual glucose (*in silico*) (gL^{-1})**
5	0.1	1.88	1.79	0.0	1.32	35.60	0.0
10	0.03	3.22	2.49	0.0	1.87	33.16	0.0
10	0.1	3.41	3.09	0.0	3.46	−10.7	0.0
10	0.17	3.54	3.34	0.0	3.55	−5.92	0.0
10	0.23	3.55	3.32	0.0	3.67	−9.54	0.0
50	0.17	16.32	13.27	0.0	20.75	−36.05	0.0
50	0.23	16.86	13.11	0.0	20.22	−35.16	0.0

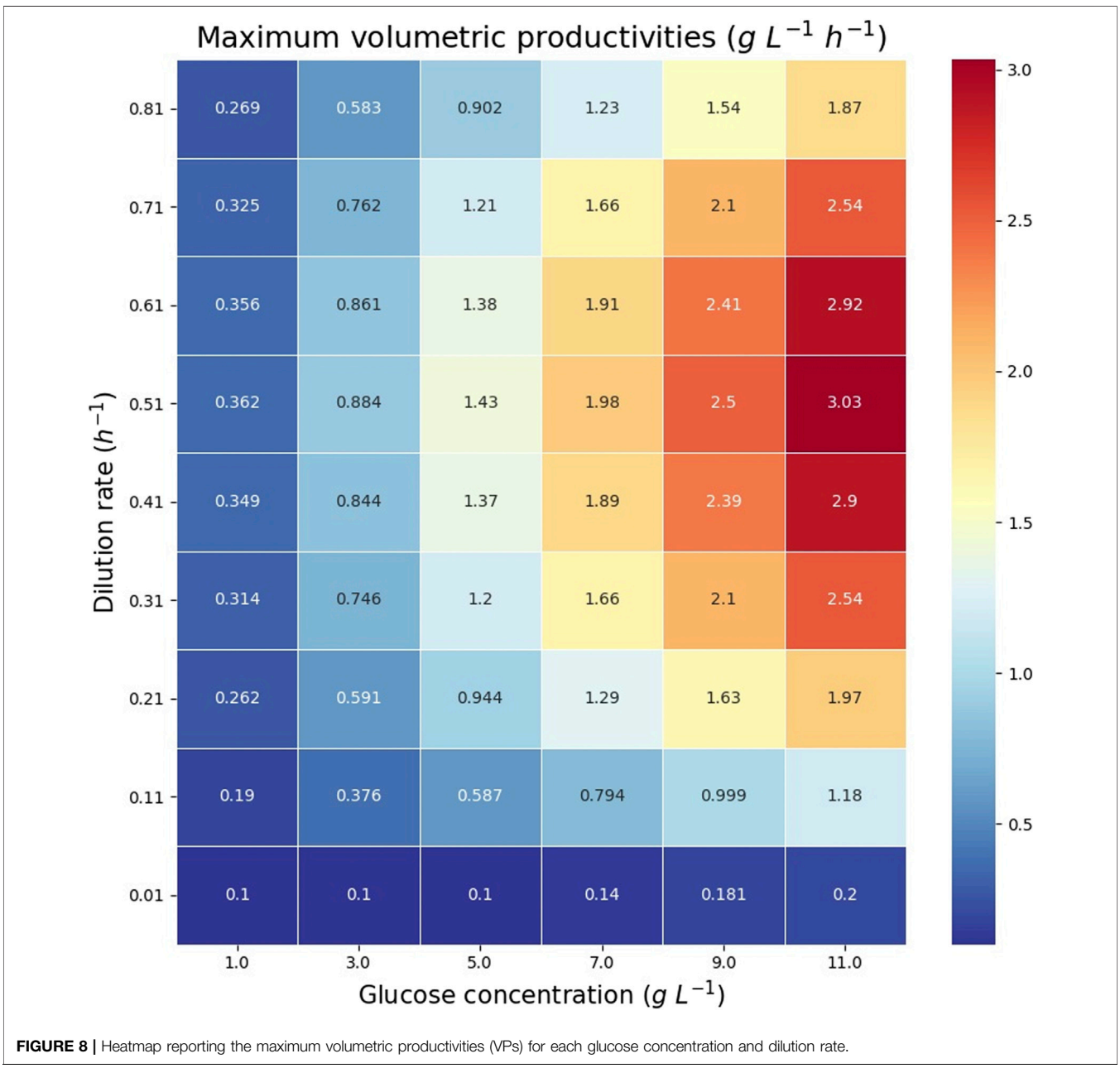


FIGURE 8 | Heatmap reporting the maximum volumetric productivities (VPs) for each glucose concentration and dilution rate.

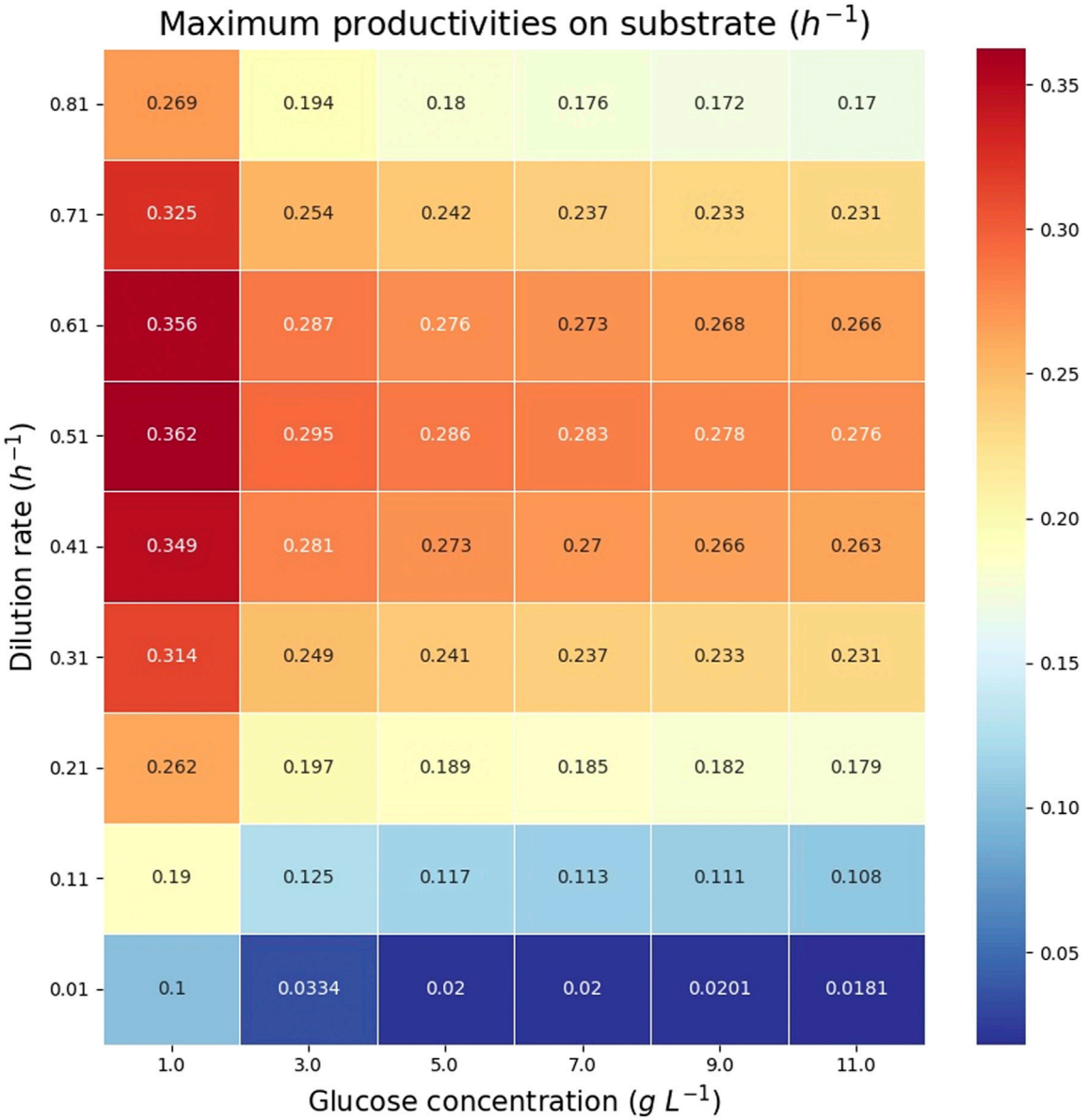


FIGURE 9 | Heatmap reporting the maximum productivities on substrate (*PS*) for each glucose concentration and dilution rate.

concentrations (column "Citramalate relative error"), and the computed concentration of glucose in the tank (column "Residual glucose").

For a specific concentration of glucose and dilution rate, the biomass and the citramalate concentration reached the values showed in columns "Biomass" and "Citramalate in tank *in vivo*" of **Table 2**. From these results, it can be confirmed that the higher the concentration of glucose and dilution rate, the greater the amount of citramalate and biomass that will be produced.

The results obtained for a set of experiments (column "Citramalate in tank *in vivo*") are consistent with the results obtained by the simulation of the FN model (column "Citramalate in tank *in silico*") for the production of citramalate. The most similar outcome occurs when the glucose concentration was 10 gL^{-1} as shown in the "Citramalate relative error" column in **Table 2**.

Notice that, in all cases, the supplied glucose is used up by the culture, i.e. the concentration of residual glucose is 0 gL^{-1} (column "Residual glucose (*in vivo*)"). This fact is correctly predicted by the FN model (column "Residual glucose (*in silico*)").

Once the model was validated, it was exploited to estimate the theoretical maximum productivity (see **Subsection 2.5.2**) of citramalate as well as the optimum biomass that produces its associated productivity.

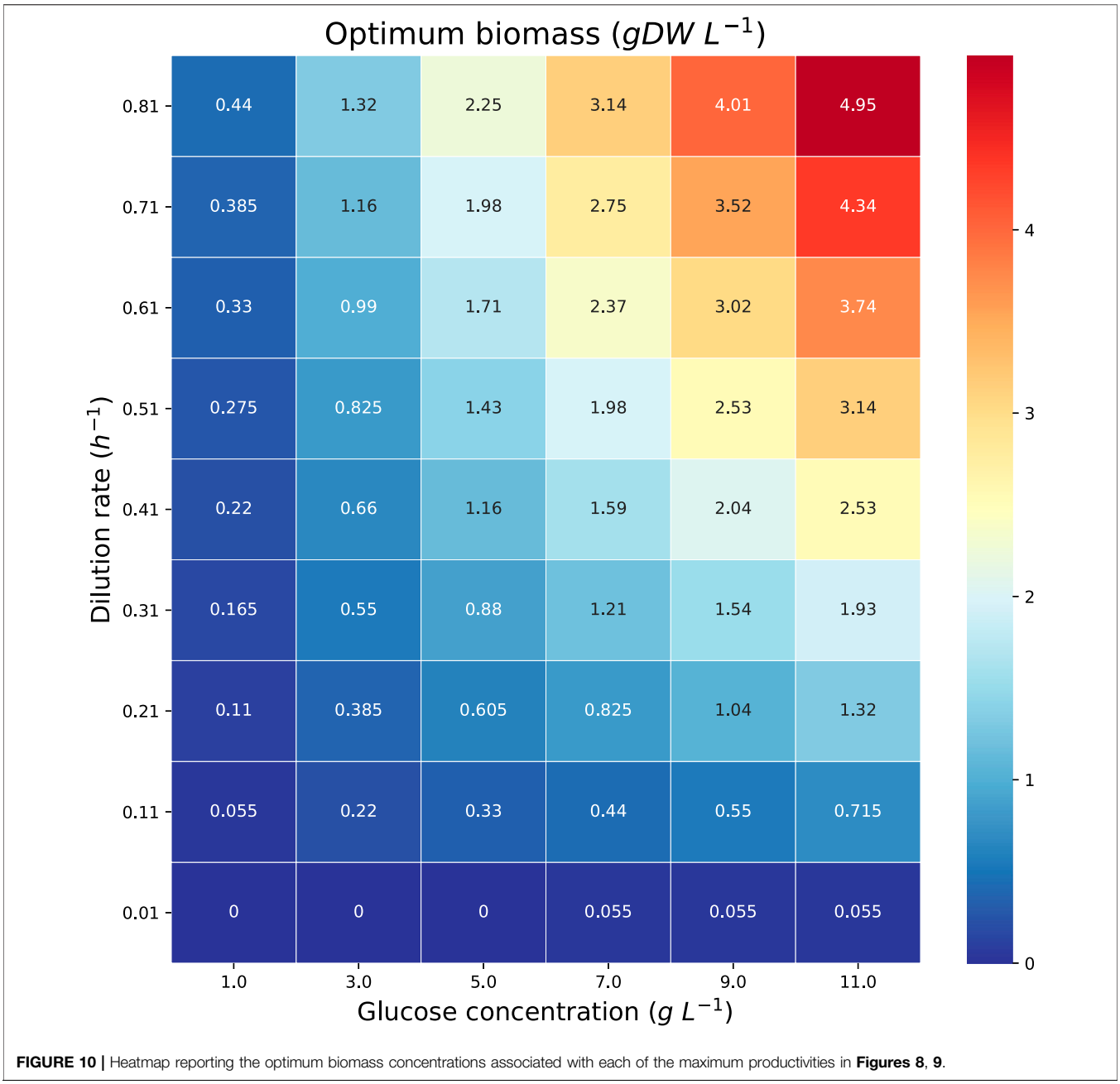


FIGURE 10 | Heatmap reporting the optimum biomass concentrations associated with each of the maximum productivities in **Figures 8**, **9**.

The theoretical maximum volumetric productivity (*VP*) and productivity on substrate (*PS*) of citramalate are reported in the heatmaps in **Figure 8** and **Figure 9**, respectively. The cell densities, or biomass concentrations, for which the productivity is optimized are reported in the heatmap in **Figure 10**. For the explored glucose concentrations and dilution rates, the highest *VP* was reached when the glucose concentration was 11.0 gL^{-1} and the dilution rate was 0.51 h^{-1}, such a productivity is obtained with a biomass of 3.14 $gDW\ L^{-1}$. With respect to *PS*, the highest value was obtained for a glucose concentration of 1.0 gL^{-1} and a dilution rate of 0.51 h^{-1}, such a productivity is obtained with a biomass of 0.275 $gDW\ L^{-1}$. It is important to note that the maximum productivities for all concentrations of glucose are obtained when the dilution rate is 0.51 h^{-1}. As expected, the amount of biomass necessary to maximize the productivity increases as the dilution rate and the glucose concentration increase, **Figure 10**.

The reported *in silico* results were obtained by *fnyzer* (Júlvez and Oliver, 2020a) which transforms the FN in **Figure 5** into a mixed-integer programming problem, and calls the CPLEX solver (IBM, 2010) to compute the numerical values.

4 DISCUSSION

The mass production of commodity chemicals from fossil fuels can entail a serious negative impact in the environment. As a

consequence, alternative approaches are being designed in order to redirect chemical production to more sustainable methods. However, the implementation of these novel approaches at an industrial scale requires optimization before they can replace traditional methods. To achieve this, biotechnology needs to exploit the advantages offered by computational models. Models can provide guidance for the design of experiments, give insights about the underlying mechanisms of the system, perform predictions and rule out infeasible hypotheses. Given the speed at which models can be simulated and optimized, they can save significant amounts of time, effort, and money in the wet lab.

In the last 2 decades, different modeling approaches have been developed, and particular attention has been paid to models of metabolism. In this work, it has been proven that the modeling formalism of FNs can integrate genome-scale constraint-based models, which lack detailed kinetic information, and kinetic models, which account for the concentration of the compounds of the system and are expressed as differential equations. Furthermore, FNs can also accommodate uncertainties inherent to the model, for example, partially unknown parameters.

An FN is represented as the combination of two nets: the event net and the intensity net. The event net models the stoichiometry, whilst the intensity net models the system dynamics. Such a graphical representation produces an overall view of the whole system. The analysis of an FN relies on the solution of a programming problem derived from the FN. If all the reaction rates are linear, i.e., the FN does not have guards, then the resulting programming problem includes only real variables with linear and quadratic constraints that can be solved very efficiently. As a consequence, FNs can handle efficiently genome-scale metabolic networks whose kinetic information is given by flux bounds and linear expressions that define the quantity of metabolites.

In contrast, if the reaction rates are not linear, they need to be approximated by piecewise linear functions, i.e. a guarded FN, which results in programming problems with real and binary variables. The complexity of the algorithms to solve mixed-integer programming problems is exponential in the number of binary variables. In order to obtain a balance between computational burden and accuracy of the model, the number of regions, and hence the number of binary variables, can be modified.

In **Section 3**, it was shown that FNs are a useful tool to predict the behavior of a complex system, such as a continuous culture in a bioreactor. The predictions of the citramalate production for a specific dilution rate and biomass were reliable in comparison to the experimental results obtained in the *in vivo* experiments, especially, the ones in which the glucose concentration in the medium was 10 gL^{-1}.

Although the simulations were highly predictive, the model could be improved further by adding some additional information, such as more components in the culture medium, or constraining the uptake of glucose depending on the glucose import rate, and similarly with the citramalate export rate. Both rates depend on protein transporters that can be saturated. Furthermore, the implementation of omics data could improve the model as well (Sánchez et al., 2017).

Not only was this method useful to reproduce the results of the *in vivo* experiments, but it can also help guide these experiments and optimize the conditions without wasting time and resources. The optimization performed in **Section 3** showed that it is of special interest to explore *in vivo* the conditions that maximized the productivity in the computational simulations.

DATA AVAILABILITY STATEMENT

The Python code generated for this work is available at: https://github.com/jlazaroibanez/opticomch.

AUTHOR CONTRIBUTIONS

JL, JJ, and SO designed the research; JL, GJ, and JJ developed and implemented the methodology; YY ran the fermentations and generated the experimental data reported; MT-A helped establish initial fermentation conditions; GL oversaw fermentation data collection and analysis; JL, JJ, and SO analyzed and interpreted the results; JL, JJ, and SO wrote the manuscript.

FUNDING

This work was supported by the Spanish Ministry of Science and Innovation (ref. DAMOCLES-PID2020-113969RB-I00), by a PhD fellowship from Diputación General de Aragón (DGA) (to JL), by the Biotechnology & Biological Sciences Research Council (UK) grant nos. BB/N02348X/1 (to SO) and BB/N023587/1 (to GL), as part of the IBiotech Program, and by the Industrial Biotechnology Catalyst (Innovate UK, BBSRC, EPSRC) to support the translation, development, and commercialisation of innovative Industrial Biotechnology processes.

ACKNOWLEDGMENTS

We are grateful to Giuseppe Nicosia, Cambridge Systems Biology Centre, and to Andrew Yiakoumetti, University of Nottingham, for helpful discussions and comments. The UCL laboratory thanks James Allen for providing the recombinant *E. coli* strain employed in the fermentation experiments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.855735/full#supplementary-material

Supplementary Data | 1. `Maximum_Productivities.xlsx`: This spreadsheet reports the computed theoretical maximum productivities and their associated cell densities for dilution rates ranging in the interval [0.01, 0.81] h^{-1} and glucose concentrations in the medium ranging in the interval [1, 11] gL^{-1}.

2. `EcolicitFN.xlsx`: This spreadsheet reports the fluxes and concentrations for a system with dilution rate D = 0.25 h^{-1}, glucose concentration in the medium glc = 7.0 gL^{-1} and cell density X = 0.5 $gDWL^{-1}$. For clarity, only positive fluxes are reported and reactions with null flux are omitted.

3. `Regions_and_Runtime.xlsx`: This spreadsheet reports the maximum productivity on substrate (PS), optimum biomass, and run-time for a dilution rate of 0.23 h^{-1}, a glucose concentration of 10 gL^{-1}, and a number of regions in the interval [10, 200]. All the simulation results were obtained on a 4 x Intel i5-7200 CPU, 2.50 GHz running Ubuntu 19.10 with 7.7 GB of RAM.

REFERENCES

Chowdhury, A., Zomorrodi, A. R., and Maranas, C. D. (2014). K-Optforce: Integrating Kinetics with Flux Balance Analysis for Strain Design. *Plos Comput. Biol.* 10, e1003487. doi:10.1371/journal.pcbi.1003487

[Dataset] IBM (2010). IBM ILOG CPLEX Optimizer.

Fernandez-de-Cossio-Diaz, J., Leon, K., and Mulet, R. (2017). Characterizing Steady States of Genome-Scale Metabolic Networks in Continuous Cell Cultures. *Plos Comput. Biol.* 13, e1005835. doi:10.1371/journal.pcbi.1005835

Flassig, R. J., Fachet, M., Höffner, K., Barton, P. I., and Sundmacher, K. (2016). Dynamic Flux Balance Modeling to Increase the Production of High-Value Compounds in green Microalgae. *Biotechnol. Biofuels* 9, 165. doi:10.1186/s13068-016-0556-4

Green, C. A., Yiakoumetti, A., Allen, J. R., Ward, J. M., and Stephens, G. (2018). Ensuring Sustained Genetic Stability and Productivity during Continuous Culture of *E. coli* for the Bio-Production of Citramalate. *New Biotechnol.* 44, 29–30. doi:10.1016/j.nbt.2018.05.128

Hanly, T. J., Tiernan, A. R., and Henson, M. A. (2013). Validation and Optimization of a Yeast Dynamic Flux Balance Model Using a Parallel Bioreactor System. *IFAC Proc. Volumes* 46, 113–118. doi:10.3182/20131216-3-in-2044.00002

Júlvez, J., Dikicioglu, D., and Oliver, S. G. (2018). Handling Variability and Incompleteness of Biological Data by Flexible Nets: a Case Study for Wilson Disease. *npj Syst. Biol. Appl.* 4, 7. doi:10.1038/s41540-017-0044-x

Júlvez, J., and Oliver, S. G. (2020b). A Unifying Modelling Formalism for the Integration of Stoichiometric and Kinetic Models. *J. R. Soc. Interf.* 17, 20200341. doi:10.1098/rsif.2020.0341

Júlvez, J., and Oliver, S. G. (2020a). "fnyzer: A Python Package for the Analysis of Flexible Nets," in *18th International Conference on Computational Methods in Systems Biology* (Konstanz, Germany: Springer), 349–355. vol. 12314 of Lecture Notes in Computer Science. doi:10.1007/978-3-030-60327-4_19

Kim, O. D., Rocha, M., and Maia, P. (2018). A Review of Dynamic Modeling Approaches and Their Application in Computational Strain Optimization for Metabolic Engineering. *Front. Microbiol.* 9, 1690. doi:10.3389/fmicb.2018.01690

Mahadevan, R., Edwards, J. S., and Doyle, F. J. (2002a). Dynamic Flux Balance Analysis of Diauxic Growth in escherichia Coli. *Biophysical J.* 83, 1331–1340. doi:10.1016/S0006-3495(02)73903-9

Mahadevan, R., Edwards, J. S., and Doyle, F. J., III (2002b). Dynamic Flux Balance Analysis of Diauxic Growth in escherichia Coli. *Biophysical J.* 83, 1331–1340. doi:10.1016/s0006-3495(02)73903-9

Malik-Sheriff, R. S., Glont, M., Nguyen, T. V. N., Tiwari, K., Roberts, M. G., Xavier, A., et al. (2020). BioModels-15 Years of Sharing Computational Models in Life Science. *Nucleic Acids Res.* 48, D407–D415. doi:10.1093/nar/gkz1055

Meadows, A. L., Karnik, R., Lam, H., Forestell, S., and Snedecor, B. (2010). Application of Dynamic Flux Balance Analysis to an Industrial escherichia Coli Fermentation. *Metab. Eng.* 12, 150–160. doi:10.1016/j.ymben.2009.07.006

Murata, T. (1989). Petri Nets: Properties, Analysis and Applications. *Proc. IEEE* 77, 541–580. doi:10.1109/5.24143

Orth, J. D., Conrad, T. M., Na, J., Lerman, J. A., Nam, H., Feist, A. M., et al. (2011). A Comprehensive Genome-scale Reconstruction of *Escherichia coli* Metabolism-2011. *Mol. Syst. Biol.* 7, 535. doi:10.1038/msb.2011.65

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What Is Flux Balance Analysis? *Nat. Biotechnol.* 28, 245–248. doi:10.1038/nbt.1614

Reimers, A.-M., and Reimers, A. C. (2016). The Steady-State assumption in Oscillating and Growing Systems. *J. Theor. Biol.* 406, 176–186. doi:10.1016/j.jtbi.2016.06.031

Sánchez, B. J., Zhang, C., Nilsson, A., Lahtvee, P. J., Kerkhoven, E. J., and Nielsen, J. (2017). Improving the Phenotype Predictions of a Yeast Genome-scale Metabolic Model by Incorporating Enzymatic Constraints. *Mol. Syst. Biol.* 13, 935. doi:10.15252/msb.20167411

Sigmarsdóttir, Þ., McGarrity, S., Rolfsson, Ó., Yurkovich, J. T., and Sigurjónsson, Ó. E. (2020). Current Status and Future Prospects of Genome-Scale Metabolic Modeling to Optimize the Use of Mesenchymal Stem Cells in Regenerative Medicine. *Front. Bioeng. Biotechnol.* 8, 239. doi:10.3389/fbioe.2020.00239

Silva, M. (1993). "Introducing Petri Nets," in *Practice of Petri Nets in Manufacturing* (Chapman & Hall, London), 1–62. doi:10.1007/978-94-011-6955-4_1

Stephens, G. M., and Dalton, H. (1987). The Effect of Lipophilic Weak Acids on the Segregational Stability of Tol Plasmids in pseudomonas Putida. *Microbiology* 133, 1891–1899. doi:10.1099/00221287-133-7-1891

Sugiyama, H., Fischer, U., Antonijuan, E., Hoffmann, V. H., Hirao, M., and Hungerbühler, K. (2009). How Do Different Process Options and Evaluation Settings Affect Economic and Environmental Assessments? a Case Study on Methyl Methacrylate (Mma) Production Processes. *Process Saf. Environ. Prot.* 87, 361–370. doi:10.1016/j.psep.2009.08.002

UniProt Consortium (2019). Uniprot: a Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* 47, D506–D515. doi:10.1093/nar/gky1049

Varma, A., and Palsson, B. O. (1994). Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Nat. Biotechnol.* 12, 994–998. doi:10.1038/nbt1094-994

Vishniac, W., and Santer, M. (1957). The Thiobacilli,. *Bacteriol. Rev.* 21, 195–213. doi:10.1128/br.21.3.195-213.1957

Webb, J. P., Arnold, S. A., Baxter, S., Hall, S. J., Eastham, G., and Stephens, G. (2018). Efficient Bio-Production of Citramalate Using an Engineered *Escherichia coli* Strain. *Microbiology* 164, 133–141. doi:10.1099/mic.0.000581

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Lázaro, Jansen, Yang, Torres-Acosta, Lye, Oliver and Júlvez. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Molecular Biosciences

REVIEW
published: 29 April 2022
doi: 10.3389/fmolb.2022.825052

Generalized Structural Kinetic Modeling: A Survey and Guide

Jana C. Massing[1,2,3]* *and Thilo Gross*[1,2,3]

[1]*Helmholtz Institute for Functional Marine Biodiversity at the University of Oldenburg (HIFMB), Oldenburg, Germany,* [2]*Helmholtz Centre for Marine and Polar Research, Alfred-Wegener-Institute, Bremerhaven, Germany,* [3]*Institute for Chemistry and Biology of the Marine Environment (ICBM), Carl-von-Ossietzky University, Oldenburg, Germany*

Many current challenges involve understanding the complex dynamical interplay between the constituents of systems. Typically, the number of such constituents is high, but only limited data sources on them are available. Conventional dynamical models of complex systems are rarely mathematically tractable and their numerical exploration suffers both from computational and data limitations. Here we review generalized modeling, an alternative approach for formulating dynamical models to gain insights into dynamics and bifurcations of uncertain systems. We argue that this approach deals elegantly with the uncertainties that exist in real world data and enables analytical insight or highly efficient numerical investigation. We provide a survey of recent successes of generalized modeling and a guide to the application of this modeling approach in future studies such as complex integrative ecological models.

Keywords: generalized modeling, nonlinear dynamics, biological networks, stability, bifurcation

OPEN ACCESS

Edited by:
Alberto Jesus Martin, Universidad Mayor, Chile

Reviewed by:
Jianwei Shen, North China University of Water Conservancy and Electric Power, China
Mohamed R. Abonazel, Cairo University, Egypt

***Correspondence:**
Jana C. Massing
jana.massing@hifmb.de

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *29 November 2021*
Accepted: *29 March 2022*
Published: *29 April 2022*

Citation:
Massing JC and Gross T (2022) Generalized Structural Kinetic Modeling: A Survey and Guide. Front. Mol. Biosci. 9:825052. doi: 10.3389/fmolb.2022.825052

1 INTRODUCTION

Much of the power of mathematics comes from its ability to describe unknown objects. Consider the number π: almost everybody knows some of its digits, but nobody knows its exact value. Our ignorance of the value, however, does not impinge on its utility in calculations, nor does it prevent us from exploring its properties.

The ability of mathematics to work with unknown objects is not limited to numbers. In modeling we can exploit this ability, by writing models in terms of unspecified functions. This is particularly useful in social or biological models where actual laws realized in nature are only approximately known.

In mathematics, working with unknown functions is the rule rather than the exception. For example, we do it every time we write the definition of a derivative

$$\dot{x} = \frac{\mathrm{d}}{\mathrm{d}t}x(t) = \lim_{\delta\to 0}\frac{x(t+\delta)-x(t)}{\delta}, \tag{1}$$

where x is an arbitrary variable and t is time. In models unspecified functions have been used to a lesser extent, but still the examples of such models are plentiful.

In the present paper we examine the method of Generalized Modeling (GM), which is a formalism by which specific types of insights can be extracted from models containing unspecified functions. Although the approach has been extended to many different contexts, GM has mostly been applied to explore systems of ordinary differential equations (ODEs). For this class of systems, GM can provide insights into the dynamics, identify conditions for the stability of steady states, explore the response of steady states to perturbations, identify parts of the system that are particularly vulnerable to perturbations, etc. Perhaps most importantly, GM generates these insights highly efficiently. It thus enables the exploration of large complex many-variable systems with comparatively little effort.

In the following we explain the basic idea of GM in **Section 2**. After this philosophical introduction, we provide a comprehensive review of previous works that used GM in **Section 3**. This is done partly to point the reader to specific variants and applications that may suit their particular needs, but also partly to illustrate the power of GM, which played a crucial rule in several discoveries leading to high-profile publications. We then provide an introductory example for the GM procedure in **Section 4**. In our experience, seeing this procedure leads to a set of specific questions, which we address in a frequently-asked-question section, **Section 5**. We then discuss the GM approach more extensively in **Section 6**, while paying particular attention to some decisions that need to be made during the modeling process. Different ways in which GM can be analyzed are then discussed in **Section 7**, before a concluding discussion in **Section 8**.

2 BASIC IDEA OF GENERALIZED MODELING

GM is best understood by contrasting it against conventional stability analysis of steady states. The conventional approach can be regarded as a 3-step process.

1) Parameterization: Restrict the model to equations that are specified except for a number of unknown parameters.
2) Steady States: Find the steady states of the ODE.
3) Linearization: Compute the Jacobian matrix, which provides a linearization of the dynamics around the steady state.

Once the Jacobian has been obtained it can be used to explore the stability of steady states, find their bifurcations, gain insights into non-stationary dynamics, etc.

It is interesting to note that the three steps of this basic program above involve very different difficulties: The first is not technically difficult, but it involves "artistic freedom": it is easy to come up with some model but it may require much experience to find the best model for a given phenomenon, and sometimes it is not even clear what constitutes the best model. By contrast, the second step has a clear-cut answer, but we need to find the roots of an equation system, a task that is prohibitively difficult for all but the simplest systems. We are thus often forced to turn to numerics, but even then, no algorithms with guaranteed convergence are known. Finally, the third step only involves differentiation of functions, which is generally easy. In the worst case we can compute the derivatives by finite difference methods.

In terms of the actual technical difficulty step 2, the computation of steady states clearly stands out. In return for braving these difficulties we obtain the number of steady states and their locations. This can be valuable information for some systems, but in many other cases we know the steady states from observation of the system and thus the prediction from the model is often merely used to eliminate some unknown parameters by setting them such that the predicted states from the model match up with their observed counterparts.

Given that the computation of steady states can introduce significant difficulties in the modeling process, but reveals only limited information, it is interesting to ask if we can circumvent this step. For example random matrix models achieve this by directly formulating a model for the Jacobian matrix, rather than deriving the Jacobian matrix from ODEs (Wigner, 1955; May, 1972; Allesina and Tang, 2012). The power of random matrix models is illustrated impressively by Robert May's seminal work (May, 1972), where he formulated a random matrix model for complex food webs. Exploiting the power of the random matrix May was able to prove mathematically that large random food webs are unlikely to be stable. The model thus proved decisively that the large food webs observed in nature must have some special features that lends them their stability. At the same time the abstract random matrix formulation gave researchers very little intuition as to what these features might be. Random matrix models are powerful because they give us direct access to Jacobian matrices in a sufficiently simple form to allow rigorous mathematical and highly efficient numerical exploration. However, as they lack the underlying layer of differential equations, these models tend to be more abstract and hence are often not easily interpretable.

GMs are situated at the halfway point between conventional and random matrix models. They have almost the full power and efficiency of random matrix models while being almost as interpretable as conventional ODE-based models. To understand how this is possible let us consider the three steps of the modeling process again. In GM we do not restrict the processes in the model to specific functional forms, and thus we cannot meaningfully compute the steady states of the model. This means in the GM the steady states are unknown quantities. However even though the steady states are unknown we can still formally linearize the dynamics around them, which yields Jacobian matrices. At first glance the elements of these Jacobian matrices sound intimidating: They are derivatives of unknown functions in unknown steady states. However, these elements can be expressed in terms of a small set of parameters that have a clear and intuitive interpretation in the context of the model.

In GM the three steps of the modeling procedure are thus reordered and slightly modified.

1) Steady states: Consider a class of models that is general enough that steady states must exist in this class. Define symbols to denote the variables in these unknown steady states.
2) Linearization: Formally compute the derivative of the processes with respect to variables to compute the Jacobian.
3) Parameterization: Identify the quantities that appear in the Jacobian as parameters of the model.

The result is a prescription for generating the Jacobian of a steady state as a function of a number of possibly unknown but interpretable parameters. In other words, we directly get the Jacobian matrix in a steady state, which is reminiscent of a random matrix model. However, simultaneously our interpretation of this matrix profits from the underlying layer of differential equations almost as if it were a conventional model (see **Table 1**).

The heart of the GM procedure and the feature that sets GM apart from other models containing unknown functions, is the

TABLE 1 | Comparison of modeling approaches.

Approach	Information Needed	Most difficult step	Insights gained
Conventional model (specified ODEs)	State variables and processes, specific functional forms for processes	Steady state computation or numerical simulation	Simulated time-series, number and stability of steady states, ...
Random matrix model	Statistics of matrix elements	Analytical solution or eigenvalue computation	Stability conditions, nature of bifurcations
Generalized model	State variables and processes	Analytical solution or eigenvalue computation	Stability conditions, nature and location of bifurcations, response to perturbations, identification of interesting parameters and parameter regions

parameterization step, where we use a specific mathematical identity to give meaning to the parameters. This is explained in more detail in **Section 4**.

In the next section we are reviewing some of the past successes of GM. Readers who are eager to see an example of the procedure first may want to skip ahead to **Section 4**.

3 GENERALIZED MODELS IN THE LITERATURE

Since its inception 12 years ago, GM has been applied to a wide range of subjects. In this section we review the areas where GM has made an impact and the ways in which the methodology has been adapted to suit to the various fields.

3.1 Food Web Models

The first GM was a simple predator-prey model proposed by Wolfgang Ebenhöh in 2003. An analysis of this model was eventually published in Gross and Feudel (2004). The predator-prey model was subsequently expanded into a general food-chain model. Analysis of this model revealed minor details in the shape of the functions used in conventional food-chain models can have a strong impact on stability properties (Gross et al., 2004), the same insight was discovered almost simultaneously in a different way in Fussmann and Blasius (2005). A similar GM setting, exploring the so-called paradox of enrichment, was recently studied in Awender et al. (2021).

While GM was initially viewed as a trick that worked in one particular model, the subsequent extensions implemented in Gross et al. (2005) made clear that the approach is generally applicable. This led to an early paper that presented the GM as a general methodology (Gross and Feudel, 2006). This paper concluded by deriving a general food-web model, but did not analyze it in any detail.

Around 2006 GMs were still studied mostly by analytical computation of the bifurcations (Gross, 2004). Although the bifurcation had been computed in food chains of up to ten levels, the structural complexity of food-web topology, still presented a serious obstacle. Instead GM approach was extended to predator-prey systems in space, modeled by partial differential equations (Baurmann et al., 2007) and was applied to study the effect of predator interference (van Voorn et al., 2008) and the dynamics of ecoepidemic models (Stiefs et al., 2009) and to explore the impact of nutrient content on predator-prey systems (Stiefs et al., 2010). The latter paper resolved a controversy that arose because different previous models for nutrient content predicted very different bifurcation diagrams. Analysis of the GM showed that all of these diagrams were projections of the same bigger picture and identified the specific assumptions that explained the differences in the respective projections.

By 2009 work in metabolic models (Steuer et al., 2006) had established numerical procedures for the investigation of GMs. This step provided an efficient method for the exploration of the food-web model formulated in Gross and Feudel (2006). The first application of this model focused on food-web stability. Since May's work, described above, identifying the properties that lend large food webs their stability had been a persistent challenge in ecology. Previous work had made progress by simulating systems of ODEs (Williams and Martinez, 2004; McCann et al., 2005; Brose et al., 2006; Neutel et al., 2002), however numerical limitations meant that only on the order of some thousand

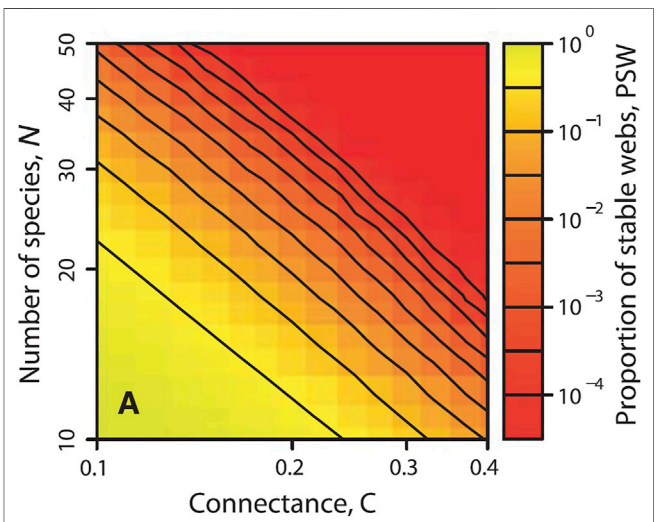


FIGURE 1 | Stability of complex food webs. Color coded is the probability that a steady state in a niche-model (Williams and Martinez, 2000) food web is stable, PSW. The result confirms May's random matrix result that large complex food webs are typically unstable. To make this figure a GM was used to analyze 35 billion different networks. The paper in which it appeared used 10^{11} niche-model food webs. Figure reprinted from Gross et al. (2009).

randomly generated food webs could be considered. By contrast, the higher numerical efficiency of the generalized food-web model allowed to study ca. 100 billion (10^{11}) randomly generated food webs within a month (**Figure 1**). Building on this data several previous insights on food web stability were confirmed, although evidence for the very popular weak link hypothesis (McCann et al., 1998) was only seen in smaller webs, instead a new topological pattern that contributed strongly to stability was identified (Gross et al., 2009). In response to these results there was a sharp increase in the interest in GM and the methodology was adopted by several labs.

Barbara Drossel and coworkers carefully examined the generalized food-web model and its relationship to conventional food-web models. They were thus able to narrow down the ranges for generalized parameters and as a result found that this increased the proportion of stable states that were found in the food webs (Plitzko et al., 2012).

Because generalized ecological models yield tractable Jacobian matrices even for relatively complex systems, they were used as a platform for a number of methodological developments. For example Stiefs et al. (2008) employed the approach to develop a method to visualize bifurcations, Lade and Gross (2012) proposed a new type of warning signal for critical transitions based on GM and Höfener et al. (2011, 2013) used GM to study a delay-coupled network of populations. The latter work led to an algorithms for designing *dynamic motifs*, small subgraphs of a network that exhibit specific dynamical instabilities regardless of the networks that they are embedded in (Do et al., 2012).

The GM approach was further refined in Yeakel et al. (2011), which carefully examined the modeling procedure, and Kuehn et al. (2013) which first supported GM by rigorous mathematical work and then went on to extend the approach to the analysis on non-local dynamics (Kuehn and Gross, 2013).

Another extension is found in the work of Helge Aufderheide, who considered eigenvector localization in GM. He was able to explain why certain food webs have a different structure but the same generalized bifurcation diagram (Aufderheide et al., 2012), a phenomenon first noticed in Gross (2004). In a subsequent work eigenvector methods were used to propose an approach for identifying the species in a food web that are most susceptible to perturbations and those that have the strongest impact on the dynamics of the system (Aufderheide et al., 2013; Doizy et al., 2018).

Aufderheide's approach was subsequently used by Yeakel and coworkers to analyze a 6,000 years time-series of Egyptian food webs (Yeakel et al., 2014). Yeakel had reconstructed an ensemble model of mammalian food webs from depictions in Egyptian art history. This dataset was then fed into the GM, which showed that extinctions from climate change events and the human population growth at the beginning of the 20th century reduced the stability of the web leaving it more and more vulnerable to further perturbations. It was also confirmed that vulnerable species, identified by GM in the initial network, were among the first to go extinct.

The utility of the GM approach for combining complex social and ecological dynamics in a common social-ecological model was highlighted in Lade et al. (2015). Using GM, the authors explored the impact of human behavior on ecological systems (Lade et al., 2013), showing that social dynamics have a strong impact on tipping points. A subsequent review discussed the use of GM in socio-ecological systems more generally (Lade and Niiranen, 2017). Other environmental applications of GM include the analysis of a climate model by Knopf et al. (2006) and stock recruitment by Yeakel and Mangel (2014).

More recently GM was used by different labs to study meta-food webs, a class of models where food webs in different spatial patches are coupled by dispersal (Leibold et al., 2004). After initial works considered a single population model on a spatial network (Tromeur et al., 2016) and food webs on two patches (Gramlich et al., 2016), the dynamics of complex food webs in large spatial networks were studied in Brechtel et al. (2018). In this work methods from algebra were used to show that it is possible to derive master stability functions (Segel and Levin, 1976; Pecora and Carroll, 1998), which then govern the food-web stability in any spatial network.

Anderson and Fahimipour (2021) used GM to study the effects of positive body size scaling of dispersal on the stability of heterogeneous metacommunities. Their results cast doubts on the widely held opinion that the ability of large bodied predators to migrate farther than small bodied species is crucial for stability.

3.2 Models of Metabolism

A second area where GM has been frequently applied is studies of metabolism. This line of work was started by Ralf Steuer in his PhD thesis. The metabolic version of GM is also known as *structural kinetic modeling* after the title of his first publication on the subject (Steuer et al., 2006). In this paper Steuer and coworkers demonstrate that GM can be applied to metabolic systems such as glycolysis in yeast and the photosynthetic Calvin cycle. For the case of glycolysis it was shown that the GM approach could be used to exactly predict one of the parameters in the system, based on stability considerations. These and other findings were later confirmed in Gehrmann et al. (2011), who used GM to analyze a more complex model of glycolysis. Carbonaro and Thorpe (2017) applied structural kinetic modeling to determine metabolic components that are major contributors to network stability in complex metabolic networks associated with glycolysis and pentose phosphate pathway and to predict the impact of perturbations on these components.

In a subsequent paper Steuer et al. (2007) analyzed the TCA cycle in mitochondria (also reviewed in (Steuer, 2007)). To deal with this more complex network they proposed a numerical sampling procedure, explained in detail in the next section. This procedure allowed Steuer to explore the model efficiently. A specific biological question driving this research was why the mitochondria under consideration hardly utilized pyruvate as an energy source. This was resolved when the GM identified pyruvate import into the mitochondrion as one of the main drivers of instability.

The sampling procedure of Steuer greatly increased the scope of GM and turned it into a highly efficient tool for the analysis of large networks. A matlab toolbox for metabolic GMs facilitating this analysis was published by Girbig et al. (2012b). The

methodology was further refined by a careful exploration of various measures to reveal important regulators (Grimbs et al., 2007). They combine GM and machine learning (ML) techniques to identify bifurcations in large systems (Girbig et al., 2012a) and incorporate thermodynamic constraints (Childs et al., 2015). The review by Srinivasan et al. (2015) highlights the potential of this GM + ML approach to scale to whole cell models.

GMs were also used to study the inhibitory feedbacks in the sucrose cycle (Henkel et al., 2011) or in designing drugs. The latter is done by comparing two different metabolic states (e.g. the healthy and non-healthy system) and their responses to perturbation with each other (Murabito et al., 2011; Murabito, 2013). In addition, structural kinetic models are also used to set up more complex models of huge metabolic networks, i.e. hybrid models. Hybrid models describe central processes in high detail, while others are roughly approximated. The central processes can be identified using structural kinetic models (Bulik et al., 2009).

Reznik and Segrè (2010) applied GM in the analysis of further metabolic cycles and showed short cycles to be highly stable. This is particularly true for non-autocatalytic cycles (Reznik et al., 2013b). Extending the methodology to metabolic genetic circuits showed that timescale separation between subsystems has a stabilizing effect (Reznik et al., 2013a). Subsequent works explored the dynamics of common regulatory motifs (Gehrmann and Drossel, 2010; Zumsande and Gross, 2010; Ackermann et al., 2012), systems of interacting compartments (Fürtauer and Nägele, 2016), and applied GM in a metabolic engineering application (Ye et al., 2015). Recently, Frandi et al. (2022) explored the emergence of oscillations in the cell cycle regulatory network of an Alphaproteobacteria via structural kinetic modeling.

3.3 Other Applications and Similar Approaches

In the medicine, GM were used to identify an early warning signal for critical transitions in systemic inflammation (Scheff et al., 2013). Moreover, Zumsande et al. (2011) proposed a GM of bone remodeling, leading to the identification of dynamical instabilities, which explain certain physiological and pathological dynamics of bones.

GMs were also applied to study questions in social dynamics and management. Gross and Feudel (2006) used a model of the Chinese Dynastic cycle as one of its examples, and the dynamics of manufacturing supply networks is analyzed in Ritterskamp et al. (2018); Demirel et al. (2019).

A similar approach to GM is the analysis applied in Kisdi et al. (2013), that uses unspecified evolutionary trade-off curves to identify trade-off functions that lead to stable limit cycles in eco-evolutionary models. Another related method is the general structural sensitivity analysis proposed by Adamson and Morozov (2014b), that considers the infinite-dimensional neighborhood of model functions to determine the sensitivity with regards to the local stability of steady states. Thereby, they provide a method to conduct bifurcation analysis under uncertainty in model functions and to determine probabilities of certain bifurcations (Adamson and Morozov, 2014a). Building on this, they also propose a method for analyzing structural sensitivity by using partially specified models and approximating the projection from the space of valid functions into the generalized bifurcation space via methods of optimal control theory (Adamson et al., 2016). They also address sensitivity analysis of these partially specified models in (Adamson and Morozov, 2020). Their framework shares the use of unspecified functions, the steady-state treatment and the incorporation of the values of unknown functions as parameters in the Jacobian with the GM approach.

4 AN INTRODUCTORY EXAMPLE

Consider a system where a variable, X, changes dynamically in response to gains and losses. We can write the differential equation

$$\dot{X} = G(X) - L(X), \tag{2}$$

where the dot denotes a time derivative, and G and L are unknown functions representing the gain and the loss terms. We have so far not constrained these functions in any way, so the only assumption is that gain and loss are in principle describable by mathematical functions.

In conventional modeling we would now proceed by parameterizing the gain and loss functions, i.e. restricting them to specific functional forms. Thereafter we could compute steady states and then launch into deeper analysis, computing stability, bifurcations etc. GMs build on the insight that most of these deeper analyses do not actually require us to restrict the processes to specific functional forms.

For example the stability of steady states is captured by the so-called Jacobian matrix $\mathbf{J}$, defined as

$$J_{ij} = \frac{\partial}{\partial X_j}\dot{X}_i\bigg|_*, \tag{3}$$

where we used $|_*$ to indicate that the expression is evaluated at the steady state under consideration. For our one-dimensional example system the Jacobian is a 1×1 matrix and its only element is

$$J_{11} = G'(X^*) - L'(X^*), \tag{4}$$

where the dash denotes a partial derivative and X^* is the steady state under consideration. Without further assumptions $G'(X^*)$ is the derivative of an unknown function evaluated at an unknown point. However, we know that terms such as $G'(X^*)$ represent numbers. We can therefore think of $G'(X^*)$ as an unknown parameter of the system. However, defined in this way, the parameter does not have an intuitive interpretation in the context of the application.

GM (in the narrow sense) is a particular way of parameterizing models such that we avoid restricting the processes to specific functional forms while capturing the uncertainty about the system in easily interpretable parameters.

To parameterize the Jacobian in an interpretable way we need to make one more assumption: All variables and process rates

have positive values. In many applications this is very intuitive as variables describe quantities that are naturally non-negative, and process rates have positive values by design (e.g. a gain would not be a gain if it were negative). The special case of one variable or process becoming exactly zero is discussed below.

Let's return to the specific example of **Eq. 2**. The equation describes a class of models in which positive steady states are bound to exist. This is not an additional assumption but merely the effect of working with a broad class of models rather than one particular parameterization. We denote these steady states as X^* and denote the rates of processes in the steady state as $L^* = L(X^*)$ and $G^* = G(X^*)$, respectively. Although we use X^* as a placeholder for every positive steady state in the system, we can formally normalize the equations with respect to X^*,

$$x = \frac{X}{X^*}, \tag{5}$$

such that $X = xX^*$. Likewise, we can define normalized

$$g(x) = \frac{G(xX^*)}{G^*}, \qquad l(x) = \frac{L(xX^*)}{L^*}. \tag{6}$$

Here we followed the GM convention of using lower-case variables for normalized quantities and upper-case variables for unnormalized quantities. We can write a differential for the normalized variable as

$$\dot{x} = \frac{\mathrm{d}}{\mathrm{d}t}\frac{X}{X^*} = \frac{\dot{X}}{X^*} = \frac{G(X) - L(X)}{X^*} = \frac{G^*}{X^*}g(x) - \frac{L^*}{X^*}l(x). \tag{7}$$

By normalizing we have moved from a system in which we did not know the steady state to a system where we know it: In the normalized system the steady state is $x^* = 1$ and in the steady state all processes run at rate 1. The price that we have to pay for this convenience is the appearance of the two factors G^*/X^* and L^*/X^*. Because these factors are unknown scalars we can interpret them as unknown parameters of the system.

Note that G^*/X^* is the per-capita gain per X in the steady state. If X is a biological population we would call it the birth rate, and L^*/X^* would be the per-capita death rate. Even in other systems such per-unit turnovers are generally well interpretable.

Another interesting observation about the two parameters is that they must be equal as $G^* = L^*$ must hold in all steady states. This allows us to define

$$\alpha = \frac{G^*}{X^*} = \frac{L^*}{X^*}. \tag{8}$$

Such parameters are known as *scale parameters* in GM. Our system now becomes

$$\dot{x} = \alpha\big(g(x) - l(x)\big). \tag{9}$$

The identity of the two fractions is a double-edged sword. We essentially used the stationarity property of the steady state to reduce the number of parameters, which generally leads to welcome simplifications. However, if we forget to exploit this simplification we may end up investigating steady states which cannot exist in the real world (say those with $G^* \neq L^*$). This is the one risk in GM that we need to steer clear of. Below we present a simple procedure for larger models that takes care of this point almost automatically.

Having successfully normalized the model we are now ready to launch into the stability analysis of the steady state X^*. For our small example the Jacobian is

$$\mathbf{J} = \big[\alpha\big(g'(1) - l'(1)\big)\big], \tag{10}$$

where the one appears because $x^* = 1$. Since we still haven't constrained the functions g and l we don't know their derivatives, hence they are also unknown parameters of the system. We define

$$g_\mathrm{x} = g'(1) \qquad l_\mathrm{x} = l'(1). \tag{11}$$

To understand the interpretation of these parameters, consider what would happen if, say, the loss were a linear function, $L(X) = aX$, with $a > 0$. In this case the normalization would result in $l(x) = x$, regardless of a, and hence $l_\mathrm{x} = 1$. So every linear relationship results in a parameter value of 1. Furthermore, any power law, $L(X) = aX^p$ results in $l_\mathrm{x} = p$. So a quadratic relationship would be signified by a parameter value of 2, a square root by 1/2, and a reciprocal relationship, e.g. $L(C) = a/X^p$ by a parameter value of $-p$.

Parameters such as g_x and l_x are called elasticities. In the context of GM they are also called *exponent parameters*. One can show that they are the logarithmic derivatives of the original functions. For example

$$l_\mathrm{x} = \left.\frac{\mathrm{d}\ln L}{\mathrm{d}\ln X}\right|_*. \tag{12}$$

We could have saved ourselves some work by using the mathematical identity

$$\left.\frac{\partial L}{\partial X}\right|_* = \frac{L^*}{X^*}\left.\frac{\partial \ln L}{\partial \ln X}\right|_* = \alpha l_\mathrm{x}, \tag{13}$$

however explicit normalization is generally felt to be the saner and safer way to a GM.

Elasticity parameters were originally introduced in economic theory (Reilly, 1940), where they remain in wide use. In biology, elasticities are central parameters studied by metabolic control theory (Fell and Sauro, 1985). Besides their convenient interpretation, elasticities provide a measure of nonlinearity that can be very robustly estimated based on limited and noisy data.

Returning to our example system, we can say that **Eq. 10** captures the dynamics around all steady states in all models of the form of **Eq. 2** by three intuitive parameters: the elasticity of gain and loss, and a turnover rate.

A steady state is stable if all eigenvalues of the Jacobian have negative real parts. In our one-dimensional example the Jacobian is a 1×1 matrix and thus has only one eigenvalue which is identical to the matrix element itself,

$$\lambda = \alpha\big(g_\mathrm{x} - l_\mathrm{x}\big). \tag{14}$$

Hence, a steady state under consideration is stable (i.e. $\lambda < 0$) if

$$g_\mathrm{x} < l_\mathrm{x}. \tag{15}$$

Even in this very simple example the analysis reveals a concrete result. For all steady states in all systems where a single positive variable is subject to gains and losses (systems of the form of **Eq. 2**) the following is true:

- Turnover rate does not directly impact stability. (It may however indirectly impact stability, for example if losses experience stronger nonlinearity under increased turnover.)
- A steady state is stable if the elasticity of loss is greater than the elasticity of gain in the steady state.

To readers with experience in modeling, these results will be hardly surprising. Consider however, that in this small example we have given the GM only very little structural information to work with. We show below that the same procedure can be applied to models of almost arbitrary complexity. If we provide more information, e.g. by specifying complex food-web topologies or metabolic networks, GM can reveal deeper, more detailed insights.

5 FREQUENTLY ASKED QUESTIONS

Our introduction to GM continues in **Section 6**, below. However, in our experience researchers frequently have specific questions after seeing the first introductory example. We therefore seize on this opportunity to answer the most common ones in this frequently asked questions section.

What if there are multiple steady states?

The general model captures the stability of all of them, but because the normalization is done with respect to the steady state, different steady states will be described by different parameter values.

It seems too easy. Does it actually work?

Yes it does, the procedure, as it has been spelled out here has been supported by rigorous mathematical proofs (Kuehn et al., 2013). Perhaps more importantly the papers cited in **Section 2** provide plenty of evidence that valuable information can be gained by GM.

Are there some things you can't do with generalized models? Why doesn't everybody use them?

There are a lot of things that cannot be done with GM, for example you can never compute where a steady state actually is. Also, you cannot simulate a GM, but you can explore it with other analysis methods that are safer, more efficient and often more powerful than simulation. GMs are not meant to replace conventional modeling, instead they are an additional tool by which some information can be gained very cleanly and efficiently. In practice they are used to explore model structures and to narrow down on parameter regions that are feasible, plausible and interesting before exploring in more detail with conventional models.

Is this limited to analysis of steady states?

In principle, no, in practice mostly yes. Christian Kühn has developed a method by which GMs can be used to explore the stability of other attractors (Kuehn and Gross, 2013). While mathematically sound, this extension requires to parameterize the shape of the attractor, which increases the size of the parameter space. Moreover, instead of the comparatively simple stability analysis we have to analyze the model using Floquet theory. In practice an easier alternative is often to change the way the system is modelled. For example, instead of modeling a differential equation system that has a limit cycle we can directly model the Poincaré map, in which the cycle appears as fixed point and can hence be explored by local analysis.

Is this only useful for stability analysis?

Kind of, but not entirely. Stability analysis and its downstream products (bifurcations, robustness, identification of sensitive or influential variables) are currently the best tools in the GM toolbox. But some other analysis can be done as well (see **Section 7**).

Can I have processes that become negative?

No, if your processes become negative, it breaks the normalization. In practice, there is an easy fix to this: Instead of having a process that can run in both directions define two processes that run antagonistically. For example, in a chemical reaction we would treat forward and reverse reactions as two different processes, which may be differently regulated. When modeling ecological dispersal between habitat patches, we model emigration as a separate process from immigration. In most cases this leads to better and more interpretable models.

How about variables or processes becoming zero?

In general this is not a big problem. Suppose you have an ecological model in which some species can go extinct. If you analyze the full GM of the system then the results will apply to steady states where all species coexist. Also, we can verify that the steady states under consideration is a state where all species are present as this information is reflected in the generalized parameters. If we are particularly interested in the case where one of the species goes extinct then we can make another model where that species is absent. We might also be interested in the transition where the extinction occurs, and in general the GM can find it. Consider that the model in which the species is present remains valid as we approach the point of extinction. Validity in this limit is sufficient to detect the transition in which the extinction occurs.

What about conversation laws? Other peculiarities of my system?

Conservation laws present us with similar issues as the stationarity conditions discussed above. Such additional

constraints provide an opportunity to narrow down the parameters space. However, we must ensure that we seize this opportunity to avoid parameterizing systems that violate the conservation law and hence cannot exist in the real world. In the context of GMs conservation laws and also some other knowledge that we might have about the real system can be taken into account in the form of algebraic equation. This is described in the next section.

6 GENERAL MODELING PROCEDURE

We now turn to the procedure by which complex GMs are formulated. For this purpose we follow the example of a simple predator-prey system from Yeakel et al. (2011), but discuss it in greater detail.

6.1 Identification of State Variables

The first step in this process is to identify the state variables that we want to describe. Deciding what state variable should be included is often easy, but can become complicated in models including human behavior.

In GM introducing additional state variables often pays off in terms of interpretability and does not incur a high cost in terms of tractability. Hence, it is generally advisable to include a candidate variable rather than leaving it out. This will typically lead to large but sparse Jacobians that are often preferable to small dense ones.

6.2 Identification of Processes

Once our state variables are in place we have to identify the processes that change them. Each state variable needs at least one gain and one loss process, but there can be multiple of these processes. For example a simple predator-prey model could look like this

$$\dot{X} = S(X) - F(X,Y) - L(X) \tag{16}$$

$$\dot{Y} = G(X,Y) - M(Y), \tag{17}$$

where S describes the reproduction of the prey X, F is the loss of prey due to predation, L is the loss of prey due to other causes than predation, G is the gain by predation of predator Y, and M is the predator's mortality.

At this point we could ask why it is necessary to include for example the L term at all. After all, since S and F are general functions we could easily merge L into F summarizing all the losses, or we could even merge L into S forming an effective growth term (as for example in logistic growth). However, in GM we extract insights mainly from the structure of the model. The more detailed we can specify the structure the more insights we gain. In the example merging S and L into one term would be a bad idea, because the exponent parameter of the merged term is far less interpretable than the two individual exponent parameters of S and L.

Similarly merging F and L would be a bad idea. As this is a predator-prey model we probably want to discuss predation losses separately from other losses. Splitting the processes preserves this ability. Moreover, it allows us further elaboration of the relationship between F and G, shown below.

6.3 Normalization

We now define normalized variables and processes following the same procedure as in the introductory example, i.e. for every variable X we define

$$x = \frac{X}{X^*}, \tag{18}$$

where X^* is an unspecified stationary state, and for every process $P(X)$ we define

$$p(x) = \frac{P(xX^*)}{P^*}, \tag{19}$$

where $P^* = P(X^*)$. Processes of multiple variables can be dealt with analogously. For our example system this leads to the normalized differential equations

$$\dot{x} = \frac{S^*}{X^*}s(x) - \frac{F^*}{X^*}f(x,y) - \frac{L^*}{X^*}l(x) \tag{20}$$

$$\dot{y} = \frac{G^*}{Y^*}g(x,y) - \frac{M^*}{Y^*}m(y). \tag{21}$$

Our next goal is to capture the prefactors in meaningful scale parameters while taking the stationarity condition of the steady state into account. Although there is some freedom in the way we specify our scale parameters, it is often a good idea to use one parameter per variable to denote the total turnover and then define additional parameters as needed to specify how much the individual gains and losses contribute to the turnover.

To introduce parameters in an organized fashion we proceed as follows: We start by considering our normalized differential equations in the steady state, where the time derivative vanishes and all process rates are 1,

$$0 = \frac{S^*}{X^*} - \frac{F^*}{X^*} - \frac{L^*}{X^*} \tag{22}$$

$$0 = \frac{G^*}{Y^*} - \frac{M^*}{Y^*}. \tag{23}$$

These equations state that for each of the variables the sum of the loss terms is identical to the sum of the gain terms,

$$\frac{S^*}{X^*} = \frac{F^*}{X^*} + \frac{L^*}{X^*} \tag{24}$$

$$\frac{G^*}{Y^*} = \frac{M^*}{Y^*} \tag{25}$$

hence we define

$$\alpha_x = \frac{S^*}{X^*} = \frac{F^*}{X^*} + \frac{L^*}{X^*} \tag{26}$$

$$\alpha_y = \frac{G^*}{Y^*} = \frac{M^*}{Y^*}. \tag{27}$$

Where α_x and α_y are now our turnover parameters for the two species. Using these parameters, we can now write our differential equations as

$$\dot{x} = \alpha_x\left(s(x) - \frac{1}{\alpha_x}\frac{F^*}{X^*}f(x,y) - \frac{1}{\alpha_x}\frac{L^*}{X^*}l(x)\right) \tag{28}$$

$$\dot{y} = \alpha_y\left(g(x,y) - m(y)\right). \tag{29}$$

We are almost done here, but we still need to take care of the prefactors in front of the f and l terms. Note that by pulling the turnover rates out of these factors we created some interesting expressions. Let's use ρ to denote the factor in front of the F-term. We can write

$$\rho = \frac{1}{\alpha_x}\frac{F^*}{X^*} = \frac{F^*}{F^* + L^*}, \tag{30}$$

which shows that ρ is the proportion of the prey's loss in the steady state that is due to predation. Likewise, we can define

$$\bar{\rho} = \frac{1}{\alpha_x}\frac{L^*}{X^*} = \frac{L^*}{F^* + L^*}, \tag{31}$$

which is the proportion of the loss in the steady state that occurs due to other sources of mortality. Such scale parameters that describe the branching or merging of flows within the system are called *branching parameters*. When we introduce such parameters, we have to keep in mind that they are not independent as the branching parameters for the gains or losses of a particular variable always add up to one. In our example we can quickly verify that ρ and $\bar{\rho}$ must obey

$$\rho + \bar{\rho} = \frac{F^*}{F^* + L^*} + \frac{L^*}{F^* + L^*} = 1. \tag{32}$$

Keeping this constraint in mind we can now write our model as

$$\dot{x} = \alpha_x\left(s(x) - \rho f(x,y) - \bar{\rho}l(x)\right) \tag{33}$$

$$\dot{y} = \alpha_y\left(g(x,y) - m(y)\right) \tag{34}$$

$$\bar{\rho} = 1 - \rho. \tag{35}$$

In general, it will not be necessary to go through the normalization in such detail as the outcomes always follow the same pattern. We can thus quickly see that for example the equation

$$\dot{Z} = A(X) + B(Y) + C(Z) - Q(Z) - R(Z,Y) - S(Z,Y,X) \tag{36}$$

normalizes to

$$\dot{z} = \alpha\left(\beta_a a(x) + \beta_b b(y) + \beta_c c(z) - \sigma_q q(z) - \sigma_r r(z,y) - \sigma_s s(z,y,x)\right) \tag{37}$$

$$1 = \beta_a + \beta_b + \beta_c \tag{38}$$

$$1 = \sigma_q + \sigma_r + \sigma_s. \tag{39}$$

6.4 Timescale Normalization and Jacobian

Before we calculate the Jacobian, we can always remove one of our turnover parameters by rescaling time. For example, if we measure in terms of multiples of the turnover time of the prey $1/\alpha_x$, both equations are rescaled by this factor. As a result, we can write the system as

$$\dot{x} = s(x) - \rho f(x,y) - \bar{\rho}l(x) \tag{40}$$

$$\dot{y} = \alpha\left(g(x,y) - m(y)\right) \tag{41}$$

$$\bar{\rho} = 1 - \rho, \tag{42}$$

where $\alpha = \alpha_y/\alpha_x$ is the parameter that tells us the relative rate of predator turnover to prey turnover. If the currency of our model is abundance this factor is the prey life expectancy divided by the predator life expectancy. If the currency of the model is biomass, the turnover ratio is the ratio of metabolic rates which is typically governed by allometric scaling laws (Yeakel et al., 2018).

To construct the Jacobian matrix of our predator-prey example we compute the derivatives

$$\left.\frac{\partial \dot{x}}{\partial x}\right|_1 = s_x - \rho f_x - (1-\rho)l_x \tag{43}$$

$$\left.\frac{\partial \dot{x}}{\partial y}\right|_1 = -\rho f_y \tag{44}$$

$$\left.\frac{\partial \dot{y}}{\partial x}\right|_1 = \alpha g_x \tag{45}$$

$$\left.\frac{\partial \dot{y}}{\partial y}\right|_1 = \alpha\left(g_y - m_y\right) \tag{46}$$

where we defined the exponent parameters $s_x, f_x, f_y, l_x, g_x, g_y, m_y$ as needed and substituted $\bar{\rho} = 1 - \rho$. We can now write the Jacobian as

$$\mathbf{J} = \begin{pmatrix} s_x - \rho f_x - (1-\rho)l_x & -\rho f_y \\ \alpha g_x & \alpha\left(g_y - m_y\right) \end{pmatrix}. \tag{47}$$

Compared to Jacobians that we typically find in conventional models this is a relatively simple and neat matrix, nevertheless it captures many insights into the structure of the system. Note in particular that we were not forced to make assumptions on aspects of the system that we are typically uncertain about, such as the exact form of predator-prey kinetics. By contrast many structural features that we can be certain about are represented. For example, net growth is the differences between gains and losses, independent processes add up, prey reproduction is assumed to be independent of the predator, etc.

6.5 Additional Constraints and Auxiliary Variables

In our example system the gain of the predator is still disconnected from the loss of the prey. Surely the functions F and G in our original model are not independent. But they are also not necessarily identical. As a basic approach we could assume that the predator gain is a function of the prey loss, e.g.

$$G(X,Y) = H(F(X,Y)). \tag{48}$$

At this point one may wonder if it is useful at all to write a relationship where an unspecified function G depends on an unspecified function F in an unspecified way H. In GM the answer is generally yes, because all of these functions correspond to well defined elements of our mental model: predation loss F, conversion efficiency H, predation gain G. By representing these elements separately, we make them tangible in the equations, or, in other words, the equations become a better representation of what we have in mind when we consider the system.

Imposing such additional constraints on a GM, typically results in additional constraints on the scale and or exponent parameters. In this example we can quickly verify that there is no impact on the scale parameters: In the steady state our new condition just reads $G^* = H^*$, which does not impose any condition on existing scale parameters that would constrain their values.

To find the implications of the condition for the exponent parameters we normalize the condition using the same procedure that we applied to the differential equations. In this case, we start by defining

$$h(f) = \frac{H(F^* f, Y^* y)}{H^*}. \tag{49}$$

Then we can start from the normalization of G and write

$$g(x,y) = \frac{G(X^* x, Y^* y)}{G^*} = \frac{H(F^* f, Y^* y)}{H^*} = h(f(x,y)). \tag{50}$$

We can then compute the exponent parameters

$$g_x = \frac{\partial}{\partial x} h(f(x,y))\Big|_1 = h_f f_x \tag{51}$$

$$g_y = \frac{\partial}{\partial y} h(f(x,y))\Big|_1 = h_f f_y. \tag{52}$$

These equations now fix two of our exponent parameters, while a new parameter h_f appears that captures the elasticity of predator gain with respect to prey loss. If we are willing to assume a constant conversion efficiency as most models do, H is a linear function and hence $h_f = 1$.

The biomass conversion example presented here is still a very basic case. In other papers the same approach has been used to build significantly more complex relationships into GM. For example, Gross and Feudel (2006) show such auxiliary constraints can be used to build realistic prey-switching behavior into GMs.

For an intermediate illustration, let us think a bit deeper about biomass conversion. Ecological intuition suggests that conversion efficiency should depend on the per-capita consumption of prey by predators. Building this ecological insight into the GM gives us more structure to work with and hence offers potentially deeper insights. So, let's consider an alternative version of H:

$$G(X,Y) = H(F,C), \tag{53}$$

where C is the per-capita consumption

$$C(F,Y) = \frac{F(X,Y)}{Y}. \tag{54}$$

Even this form of the constraint does not constrain our scale parameters further. To find the constraints on the exponent parameters we define normalized forms of the auxiliary variables

$$h(f,c) = \frac{H(F^* f, C^* c)}{H^*} \qquad c(f,y) = \frac{C(F^* f, Y^* y)}{C^*} \tag{55}$$

and then verify

$$g(x,y) = \frac{G(xX^*, yY^*)}{G^*} = \frac{H(fF^*, cC^*)}{H^*} = h(f,c). \tag{56}$$

We can now compute the derivatives, keeping in mind that the auxiliary variables (h, c) are just short hand notations that need to be differentiated using the chain rule

$$g_x = h_f f_x + h_c c_f f_x \tag{57}$$

$$g_y = h_f f_y + h_c\left(c_f f_y + c_y\right). \tag{58}$$

In the solution we interpret the exponent parameters as partial derivatives such that for example c_y denotes only the derivative of c with respect to the second argument, the indirect impact of y on c via f is accounted for in the independent term $c_f f_y$.

Let's try to interpret the parameters that appear here, h_f is the partial derivative of h with respect to x, and since c now appears as an explicit argument of h it means that this is a derivative at constant c. So ecologically speaking this parameter is asking how does the growth of the predator population change if more predators are feeding but the per capita amount stays constant. This is almost certainly a linear relationship so we can assume $h_f = 1$ with much better confidence than before. The parameter h_c corresponds to the question what happens if every predator consumes a greater amount per capita. In this case the efficiency of conversion can go up or down in complex nonlinear ways depending on the ecological situation, so we keep this as a tunable parameter, whose effect can be explored with the GM. Finally, c_f and c_y describe the elasticities of the per-capita consumption. Since we specified C explicitly we can compute

$$c_f = \frac{\partial}{\partial f} c(f,y)\Big|_1 = \frac{\partial}{\partial f} \frac{C(F^* f, Y^* y)}{C(F^*, Y^*)}\Big|_1 = \frac{\partial}{\partial f} \frac{F^* f}{Y^* y} \frac{Y^*}{F^*}\Big|_1 = \frac{\partial}{\partial f} \frac{f}{y}\Big|_1 = 1, \tag{59}$$

which we could have guessed straight away because we defined C to be linear in F. Similarly,

$$c_y = \frac{\partial}{\partial y} \frac{f}{y}\Big|_1 = -1, \tag{60}$$

which is consistent with the inverse relationship we assumed. Summarizing these calculations we can write

$$g_x = (1 + h_c) f_x \tag{61}$$

$$g_y = f_y + h_c\left(f_y - 1\right). \tag{62}$$

We have included this slightly more difficult example in this review because we feel that it illustrates very well that GMs have the ability to incorporate information that we are confident about (e.g. $h_f = 1$) while not forcing us to constrain our options where we do not have such information (e.g. h_c).

We can now substitute the results into the Jacobian matrix, which would remove the now redundant parameters g_x and g_y but insert the elasticity of conversion efficiency with respect to per capita consumption H_c. We could then for example explore how this parameter impacts the stability of the system, or changing it affects the predator-prey ratio (see **Section 7.2**).

6.6 Conservation Laws

Conservation laws can be imposed on a GM in a similar way as the conditions, discussed above. However, there are two different ways in which a conservation law can be used, leading to slightly different notions of stability.

For the purpose of illustration consider that we have a two-dimensional system subject to one conservation law. We could then use the conservation law to reduce the number of variables to one even before normalization. Subsequently the normalization can be carried out normally. Alternatively, we can normalize the system and the conservation law and then interpret the conservation law as a constraint on the generalized parameters.

Both of these approaches are valid, but in the first case we arrive at a Jacobian of size 1×1, whereas in the latter case we arrive at a Jacobian of size 2×2. Both of these Jacobians describe the stability of steady states in the system, but in the former case only perturbations that respect the conservation law are allowed (hence the one-dimensional eigenspace), whereas in the latter also those perturbations are considered that violate the conservation law, leading to a slightly stricter notion of stability.

For most systems the second alternative provides the better notion of stability, unless the system is fundamentally closed and no perturbation from the outside is imaginable. The second alternative is also often simpler to implement as the simpler internal structure more than compensates for the slightly larger size of the Jacobian.

Even if we take the second route, conservation laws will impose some constraints on scale parameters, which is normally harmless, but can become complicated if we have to deal with many such constraints. This happens for example in metabolic models where the number of atoms of different elements are conserved. In such a case some additional machinery is needed to help us manage our scale parameters. A convenient solution is to represent the scale parameters as a linear combination of a set of fundamental flux modes that obey all constraints. This is explained in detail in Steuer et al. (2007).

6.7 Derivative Conditions and Optimality

As a final remark in this section let us briefly mention that we can also impose additional constraints in the form of derivatives. This makes no sense in the context of our ecological example, but, abstractly speaking, we could demand

$$\left.\frac{\partial P(X)}{\partial X}\right|_{*} = 0, \tag{63}$$

where P is some process and X is some state variable. Again, we can normalize

$$0 = \left.\frac{\partial P(X)}{\partial X}\right|_{*} = \left.\frac{\partial P^{*} p(x)}{\partial x}\frac{\partial x}{\partial X}\right|_{*} = \frac{P^{*}}{X^{*}} p_{\mathrm{x}} \tag{64}$$

This illustrates that sometimes scale parameters such as P^{*}/X^{*} can appear in the normalization of additional constraints. They are often easily dealt with as they can typically be replaced by already defined symbols (e.g. α_{x}). In this example the appearance of the scale parameter is of no consequence as the outcome stipulates $p_{\mathrm{x}} = 0$, fixing one of the exponent parameters.

The ability to specify conditions on the derivatives is interesting because it provides us with a way to demand that a variable must be in a local maximum or minimum of some function. This is useful for example if we study biological evolution in an adaptive dynamics model (Allen et al., 2013) and want to force the species to remain in locally evolutionary stable states. The same approach can also be used in governance or cooperation models to demand that agents allocate their resources optimally.

7 ANALYZING GENERALIZED MODELS

The analysis of GMs is fundamentally more constrained than the analysis of conventional models as we lack the ability to simulate the model or compute the steady states. Nevertheless, GMs can be analyzed in a variety of ways.

7.1 Numerical Stability Analysis

The main output of GM are Jacobian matrices. Hence the analysis of GMs is generally based on the analysis of Jacobians. The most direct application of these matrices is stability analysis. Given a specific set of generalized parameters, we can substitute the parameters into the Jacobian and then check the stability of the corresponding steady state by numerically computing the leading eigenvalue.

Leading eigenvalues can be computed highly efficiently using iterative eigensolvers. We can take advantage of this efficiency to explore the parameter space spanned by the generalized parameters by random sampling. For this purpose, we constrain all parameters to plausible ranges, which is possible as the parameters are generally easy to interpret. For example, we may decide to set the parameter f_{x} to cover the whole range from constant to quadratic response, i.e. the same range of possibilities that would be covered by a Holling type-III functional response in conventional models (Holling, 1959).

Once every parameter is thus constrained we can draw an ensemble of random parameter sets and evaluate their stability. We can then get a first impression of the behavior of a system by correlating the individual parameters with a stability. Suppose we have a system with P generalized parameters and we draw M sets of values for these parameters. We can then denote the m'th realization of parameter i as p_i^m where $i \in [1, P]$ and $m \in [1, M]$. Furthermore, we can denote the stability of the steady state described by the parameter set m as s_m.

It is tempting to define s_m as $-\mathrm{Re}(\lambda_0)$, where λ_0 is the leading eigenvalue of the Jacobian. However, this can be misleading in large networks as instabilities can often arise on very different timescales, such that the eigenvalue that is the leading one in most of the parameter space is not the one that causes the instability once stability is lost.

Instead we can define

$$s_m = \begin{cases} 1 & \mathrm{Re}(\lambda_0) < 0 \\ 0 & \text{otherwise} \end{cases} \tag{65}$$

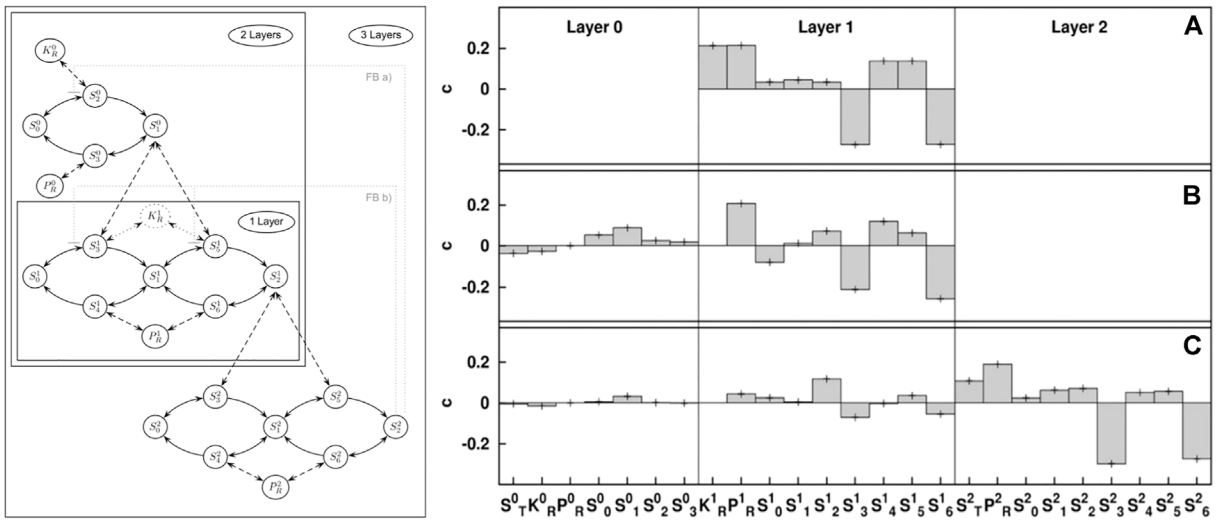


FIGURE 2 | Illustration of stability sampling. The map kinase cascade (left) is a motif that appears in gene regulation. Numerical stability analysis of a GM reveals the impact different generalized parameters have on stability of different versions of the cascade that occur in biology: single layer **(A)**, double layer **(B)**, and triple layer **(C)**. The results reveal stabilizing (positive) and destabilizing (negative) parameters. This numerical result is based on the analysis of 10 million steady states, which reduces the error bars to less than the line width of the plot. Using GM this number of states can be analyzed in seconds on modern computers. Figure adapted from Zumsande and Gross (2010).

so s_m is one if the parameter set m is stable and zero otherwise. Once we know the stability of all parameter sets we can estimate the impact of the individual parameters on stability as

$$c_i = \mathrm{Cov}_m\left(p_i^m, s_m\right). \tag{66}$$

If stability is completely explained by a particular parameter the result will be $c_i = 1$ if the parameter is stabilizing or $c_i = -1$ if it is destabilizing. In general, many parameters will correlate with stability and typical values of important parameters in large networks are around 0.1 (see **Figure 2**).

Let us emphasize that these correlations are not independent of the sampling. Sampling generalized parameters uniformly often results in a sensible sampling of the parameter space. Nevertheless, sampling parameters from wider ranges will result in stronger correlations. Therefore, it is essential to choose the ranges such that they reflect reasonable assumptions about the plausible values of the parameters.

Once we have identified a set of parameters of particular interest we can explore the effect of these parameters on stability in more detail. A common procedure is to vary one parameter x systematically, while all other parameters are randomized. We can then plot the proportion of randomly drawn parameter sets that are stable over the value x. For historical reasons this measure is commonly called the *probability of stable webs*. The same procedure can also be used with two parameters to produce two-dimensional histograms (**Figure 1**).

7.2 Response to Parameter Change

In addition to stability against short-term (pulse) perturbations, we can also ask how a dynamical system responds to a permanent change of parameters, i.e. a press perturbation.

By the implicit function theorem one can show (see Aufderheide et al., 2013) that a sufficiently small press perturbations induces a shift in the steady states described by

$$\boldsymbol{\delta} = -\mathbf{J}^{-1}\boldsymbol{p}. \tag{67}$$

where $\boldsymbol{\delta}$ is the induced shift in the steady state, $\mathbf{J}^{-1}$ is the inverse of the Jacobian matrix, and $\boldsymbol{p}$ is a vector containing the direct impacts of the perturbations on the individual equations.

If we apply this equation to a GM then the vector $\boldsymbol{\delta}$ will contain the impact on the steady state in the normalized variables, e.g. an element of 0.05 would correspond to a 5% increase. Similarly, $\boldsymbol{p}$ contains the direct impact on the differential equations in units of normalized turnover.

For example let us again consider the predator-prey system from **Section 6**. We can now ask, how switching on an additional loss term for the prey, say, from harvesting, impacts the steady state. We assume that a small fraction ϵ of the total turnover of the prey is harvested, hence

$$\boldsymbol{p} = \begin{pmatrix} \epsilon \\ 0 \end{pmatrix}. \tag{68}$$

Substituting this relationship and the Jacobian into **Eq. 68** we arrive at

$$\boldsymbol{\delta} = -\frac{1}{\det \mathbf{J}} \begin{pmatrix} \alpha\left(g_y - m_y\right) & \rho f_y \\ -\alpha g_x & s_x - \rho f_x - (1-\rho) l_x \end{pmatrix} \begin{pmatrix} -\epsilon \\ 0 \end{pmatrix}, \tag{69}$$

where

$$\det \mathbf{J} = \alpha\left(\left(s_x - \rho f_x - (1-\rho) l_x\right)\left(g_y - m_y\right) + g_x \rho f_y\right), \tag{70}$$

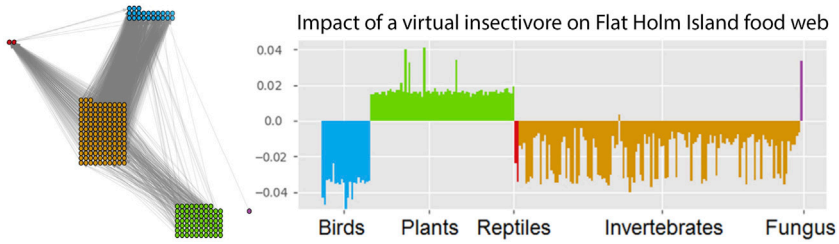


FIGURE 3 | Impact of a (virtual) insectivore on the Flat Holm island food web. The impact is assessed with a GM of the empirically observed food web (left). This network contains 227 species (23 birds, 2 reptiles, 133 invertebrates, 68 plants and 1 fungus). GM impact analysis predicts which species would profit and which species would suffer from a typical insectivore invasion (right). This reveals a general pattern of responses but can also be used to identify indicator species for the detection of bioinvasions. Figure adapted from Doizy et al. (2018).

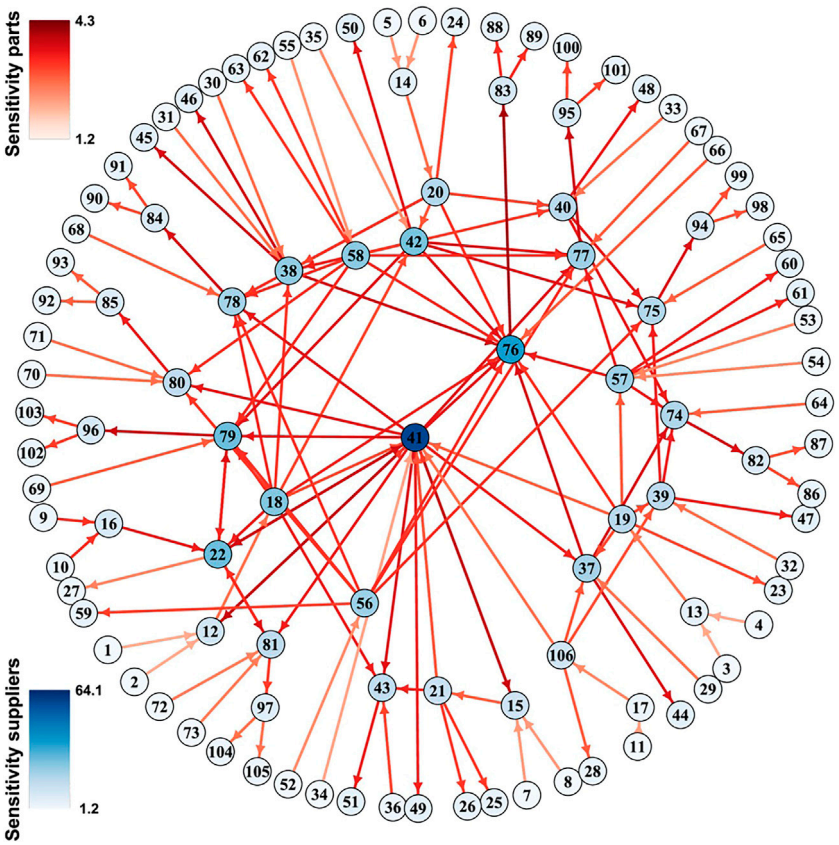


FIGURE 4 | Sensitivity in a luxury goods supply network. GM was used to study the sensitivity (colors) of firms (nodes) and products (links) in a luxury goods supply network. Figure adapted from Demirel et al. (2019).

We can now examine the two components of the resulting impact on the steady state

$$\delta_{\text{prey}} = \frac{\alpha\left(g_{\text{y}} - m_{\text{y}}\right)}{\det \mathbf{J}}\epsilon, \qquad \delta_{\text{pred}} = -\frac{\alpha g_{\text{x}}}{\det \mathbf{J}}\epsilon. \tag{71}$$

If there is no strong interference or social interaction between predators we can assume that predation is linear in predator abundance, $g_y = 1$. Moreover, in the absence of diseases and overcrowding the mortality of the predator can be assumed to be linear $m_{\text{y}} = 1$. We can now see that in this case

$$\delta_{\text{prey}} = \frac{\alpha\left(g_{\text{y}} - m_{\text{y}}\right)}{\det \mathbf{J}}\epsilon = 0 \tag{72}$$

So the prey is not impacted by a small amount of harvesting at all. This happens because the prey population is still controlled by the predator and any additional losses of the

prey are in the long run compensated by reduced predation. Even though the prey abundance does not change, we can see from δ_{pred} that the predator population is impacted by the harvesting of its prey, and at low harvesting rates responds with a proportional loss. An example for GM impact analysis in a larger network is shown in **Figure 3**.

In some applications we are not interested in the impact of a specific perturbation, but in the response to general perturbations. For this purpose Aufderheide et al. (2013) propose two measures

$$\mathrm{Se_i} = \log\left(-\sum_n \frac{|v_i|^{(n)}}{\lambda_n}\right), \qquad \mathrm{In_i} = \log\left(-\sum_n \frac{|w_i|^{(n)}}{\lambda_n}\right), \quad (73)$$

where λ_n is the nth eigenvalue of **J**, and $\boldsymbol{v_n}$ and $\boldsymbol{w_n}$ are the corresponding left and right eigenvectors. For example, $w_1|^{(2)}$ is the first element in the right eigenvector of **J** corresponding to the second eigenvalue.

Among the two measures Se_i indicates how sensitive variable i will react to typical perturbations of the network, whereas Im_i indicates how strongly a perturbation of node i will propagate to typical nodes in the network. This former is illustrated in a manufacturing network in **Figure 4**. If desired dynamical importance of a variable can be defined as the product of these measures.

7.3 Mathematical Stability Analysis and Bifurcation Theory

Because GM avoids the computation of steady states, the elements of the matrix are often simple expressions, which are convenient for pen and paper math. Hence GM is sometimes used to create Jacobians for methodological studies that need example Jacobians with certain properties. A recent example is Barter et al. (2021) which proposed a method for inferring causality (in the form of the Jacobian matrix) from correlations and used GMs to create suitable examples. Other examples for this use of GMs include Höfener et al. (2011), Höfener et al. (2013); Do et al. (2012).

The most common mathematical use of the Jacobian is bifurcation analysis, the study of the transitions between dynamical regimes in the system. In a system of differential equations local bifurcations of stationary states occur when a change in parameters causes at least one eigenvalue of the Jacobian to change sign (Guckenheimer and Holmes, 1983; Kuznetsov, 2004). Crossing the threshold where such a bifurcation occurs typically leads to qualitative transition in the system.

There are only two generic ways in which bifurcations appear in GMs. Either a single real eigenvalue becomes zero or a complex conjugate eigenvalue pair crosses the imaginary axis (saddle-node bifurcation and its variants). The former scenario corresponds to bifurcations in which the number of steady states changes, whereas the latter typically marks the onset of oscillations (Hopf bifurcation) (Guckenheimer et al., 1997).

The saddle-node-type bifurcations can be easily computed in GMs of any size. Here the task is to find the combinations of generalized parameters that lead to a zero eigenvalue of the Jacobian. Because the determinant of a matrix is the product of its eigenvalues (Gantmacher, 2000) the Jacobian has a zero eigenvalue if and only if its determinant is zero. In contrast to the eigenvalues of a matrix which can only be computed for very small systems the determinant can be computed straight forwardly for systems of any size.

For the Hopf bifurcation, we need to locate combinations of the generalized parameters that lead to a complex conjugate pair of purely imaginary eigenvalues. To find such pairs a determinant based method was proposed originally by Guckenheimer et al. (1997) and rediscovered independently in Gross and Feudel (2004). Using this method Hopf bifurcations can in principle be located in systems of any size, however the resulting equations become too complicated to be useful if the system has more than ca. 10 variables.

Once we have located the bifurcations in a system, they can be visualized in bifurcation diagrams. In particular, several analyses of GMs have used three-parameter bifurcation diagrams, which can be created by the method described in Stiefs et al. (2008) (see **Figure 5**). Each point in the volume spanned by these diagrams corresponds to a particular steady state. All steady states located in the same volume share qualitatively similar local dynamical properties, whereas qualitative transitions take place at the surfaces, which mark bifurcation points.

A *de facto* convention in GM is to orient three-parameter diagrams such that the steady states that are stable are located in the top-most volume of parameter space, whereas the steady states in all other volumes are typically unstable.

Three parameter bifurcation diagrams allow the researcher to quickly get an overview of the interaction of up to three relevant parameters. Moreover, they allow to quickly identify parameter regions where different bifurcations meet and intersect. This is particularly interesting because the intersections can reveal other dynamical features that are otherwise harder to discover. For example, it is known that a region of chaotic dynamics must exist close to the intersection of two Hopf bifurcation surfaces (Kuznetsov, 2004) (**Figure 6**). This approach was used for example in Zumsande et al. (2011) to identify a region of chaotic dynamics in the MAP Kinase cascade, a common regulatory motif in cell biology.

7.4 Return to Conventional Models

GMs are especially helpful in systems where large uncertainties exist. In these systems they can greatly speed up the initial exploration identifying interesting parameter regions and phenomena, based on limited information. However, as we progress beyond the initial exploration of the system we typically gain additional insights and/or data that we want to reflect in the model. Some of these additional insights can be used to restrict generalized parameter ranges. Others may lead to deeper understanding that results in changes to the model.

One nice feature of GMs is that they can be iteratively expanded to reflect new insights into the system. For example, in our initial exploration we may model a given process simply by

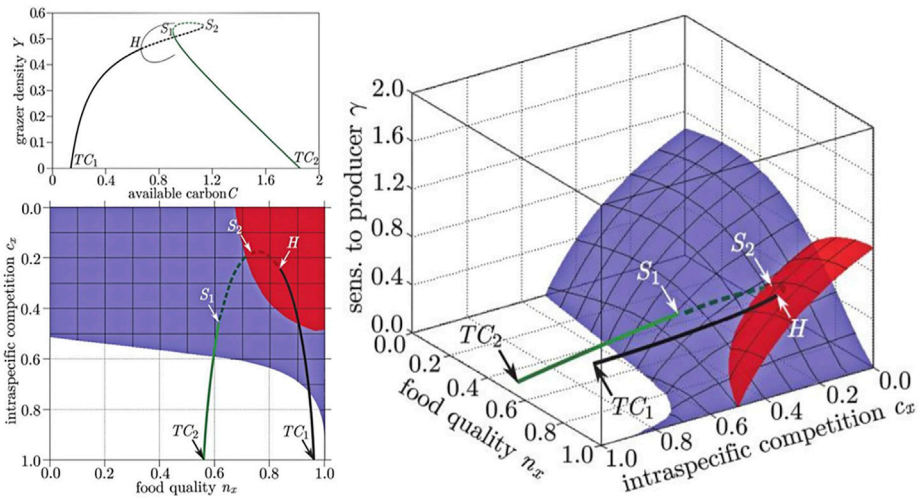


FIGURE 5 | Comparison of bifurcation diagrams in conventional and generalized studies of food quality in a producer-grazer system. In the conventional model (top left) the steady state in which grazers survive emerges from a transcritical bifurcation (TC_1), as a conventional parameter is increased the steady state loses stability in a Hopf (H) bifurcation, where a limit cycle (thin line) is created. The steady state subsequently reacquires stability by going through two saddle-node bifurcations (S_1, S_2), before vanishing in another transcritical bifurcation (TC_2). The stability of all steady states is captured by the GM (shown in two projections: right, bottom left). In the GM the bifurcation points form surfaces (red: Hopf, blue saddle node) and the states visited by the conventional example model corresponds to a trajectory through the volume spanned by the GM. The surface on which the transcritical bifurcations occur is the identical to the front left face of space shown in the right diagram and is omitted to avoid occlusion. Figure adapted from Stiefs et al. (2010).

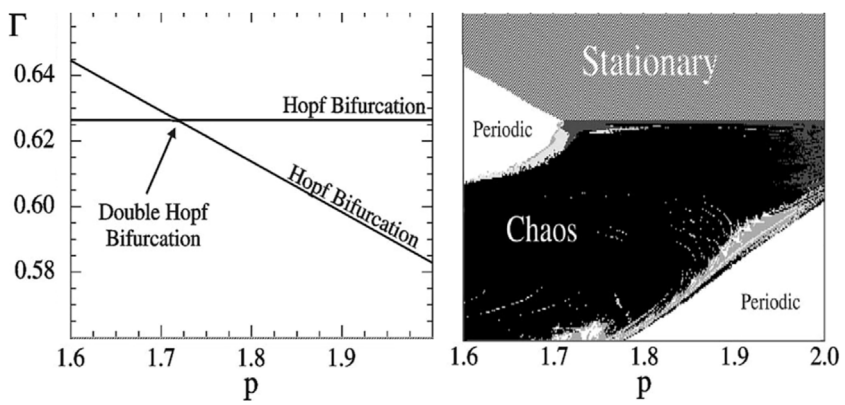


FIGURE 6 | Chaos in a four-trophic food chain. A GM reveals that the steady states are stable in the topmost volume of the parameter space shown in the two-parameter bifurcation diagram **(left)**. When parameters are changed, steady states can lose their stability by crossing either of two Hopf bifurcations (lines). At the intersection of these lines a codimension-2 double-Hopf bifurcation point is located. Observing this bifurcation in the GM points to a chaotic parameter region nearby. Indeed, this region can be discovered by numerical calculation of Lyapunov exponents in a conventional model **(right)**. In the region where the GM predicts stability of the steady state the conventional model is stationary (dashed), in the unstable regions we observe oscillations (dark grey), quasi-periodicity (light and medium grey) and chaos (black). Figure adapted from Gross et al. (2005).

a single unknown function $F(X, Y)$ and the resulting Jacobian will contain parameters such as f_x and f_y. Once we understand this process in more detail we may want to replace this completely unconstrained function by a formulation that uses some newly gained insights. For example we may realize that F is the product of some independent factors $F(X, Y) = A(X)Y$. We can now normalize this additional equation and use it to simply replace the old parameters in our Jacobian by new ones (in this example, $f_x = a_x$, $f_y = 1$). In this way additional detail can be added to models iteratively without redoing the complete normalization procedure. This is discussed in more detail in Yeakel et al. (2011).

As we understand our model better we may eventually want to restrict more and more functions to specific functional forms. This can lead to hybrid models where some equations are completely specified, whereas others still contain generalized terms. In this we need to solve manually for the steady states of fully specified equations, the resulting stationarity conditions then typically act as additional constraints on the generalized

parameters. An example of such an hybrid model is the laser system discussed in Gross and Feudel (2006).

Quite commonly, GM will identify parameter regions of particular interest. To explore the dynamics in these regions it is often desirable to run some numerical simulations. This means that we need a way to construct conventional models that are consistent with a given set of generalized parameters. In general, there will be many models that match the desired parameter set and many different ways to find them. However, the easiest and fastest procedure is to replace the general functions in the model by specific functions that obey the normalization condition $F(1) = \alpha$, where α is the desired turnover rate. This guarantees that the specific model that we are constructing still has a steady state at $X^\star = 1$ which saves us the work of computing the steady state.

For illustration consider again our introductory example

$$\dot{X} = G(X) - L(X), \tag{74}$$

we know already that the Jacobian after timescale normalization is

$$\mathbf{J} = \alpha(g_x - l_x). \tag{75}$$

Let's say we are interested in a steady state that is characterized by $\alpha = 1$, $g_x = 1/2$ and $l_x = 2$. The challenge is now to find specific functions $G(X)$ and $L(X)$ such that there is a steady state that matches these parameter values. One class of functions that meet our condition $F(1) = 1$ are the power laws $F(X) = X^p$. Computing the exponent parameter corresponding to such a power law, yields p. Thus, $G(X) = X^{1/2}$ and $L(X) = X^2$ meet the stationarity condition and match the desired parameters. Hence one possible example model is

$$\dot{X} = X^{1/2} - X^2. \tag{76}$$

This is already a solution, but let's say the first term $X^{1/2}$ would be unrealistic in the context of our application. Instead we want a term of the form

$$G(X) = \frac{AX}{X + K}. \tag{77}$$

To make this work, we first enforce the normalization conditions $G(1) = 1$, by setting $A = 1 + K$. Then we choose K such that

$$p = \frac{1}{2} = \frac{\partial}{\partial X} \frac{(1 + K)X}{K + X}\bigg|_1 = \frac{K}{(K + 1)} \tag{78}$$

which required $K = 1$. Hence, also

$$\dot{X} = \frac{2X}{1 + X} - X^2 \tag{79}$$

is a specific model that is consistent with the desired parameter values. The same approach can be used to construct specific realizations of complex GMs. Therefore this method can serve as a constructive proof that each point in the generalized parameter space corresponds to a realizable steady state in a plausible conventional model (Kuehn et al., 2013).

A small caveat regarding the procedure above is that the specific construction results in degeneracy in certain bifurcations. For simulation studies this is normally not a concern, but may cause peculiar results in bifurcation analyses.

8 SUMMARY AND DISCUSSION

In the present paper we summarized the state of the art in generalized modeling. In the past generalized modeling has been used in more than 50 publications in diverse areas of Science, Engineering, Mathematics and Medicine (see references cited above). As a result, the method has seen increasing adoption by labs around the world.

Although generalized modeling is mathematically straight forward, its philosophy differs in important ways from conventional modeling. In our experience, these differences mean that young researchers with limited experience in dynamics find it easier to adopt generalized modeling than seasoned modelers, for whom it takes greater cognitive effort to switch to a different mental framework.

Conventional modeling is very much grounded in the belief in an ultimate truth: At least within the model, the variables are governed by precise and exact rules and equations that given some initial conditions permit only one possible outcome. This is even true for stochastic models, which postulate micro-scale randomness, but work with precisely defined laws on the level of distributions.

By contrast, generalized models, acknowledge that we have a limited view of reality and may hence be unable to perceive the exact laws that are at work in the system. Instead of postulating one definite reality, generalized models work with the whole infinite ensemble of possible realities that are consistent with the available structural knowledge. They explore dynamical implications within this ensemble, allowing the researcher to further narrow down the set of possible worlds.

As we have shown, some questions can be answered very efficiently by analyzing the whole ensemble of possible worlds captured by the generalized model. These questions include the analysis of dynamical stability and bifurcations of steady states, prediction of responses to different types of perturbations, and identification of important parameters and parameter regions.

In summary generalized modeling offers a highly efficient approach to extract types of insights from limited information. This efficiency of generalized modeling is not limited to numerical efficiency, but also allows mathematical solutions in systems of intermediate complexity, and perhaps most importantly saves the researcher time. Formulating a generalized model involves considerably less work than a comparable conventional model. It avoids extensive research and considerations which may be necessary in a conventional model to fix rate constants and parameterize kinetic laws.

Once a researcher is familiar with the general procedure and, more importantly, has adapted to its philosophy, generalized models can typically be formulated, analyzed and adapted within just a few hours. We hope that this review will help many new researchers to discover this exciting and entertaining modeling approach.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication. The authors worked together on all aspects of this paper.

FUNDING

This work was supported by the Ministry of Science and Culture of Lower Saxony (HIFMB Project) and the Volkswagen Foundation (ZN3285).

REFERENCES

Ackermann, E., Weiel, E. M., Pfaff, T., and Drossel, B. (2012). Boolean versus Continuous Dynamics in Modules with Two Feedback Loops. *Eur. Phys. J. E* 35, 1–13. doi:10.1140/epje/i2012-12107-9

Adamson, M. W., and Morozov, A. Y. (2020). Identifying the Sources of Structural Sensitivity in Partially Specified Biological Models. *Sci. Rep.* 10, 16926. doi:10.1038/s41598-020-73710-z

Adamson, M. W., and Morozov, A. Y. (2014a). Bifurcation Analysis of Models with Uncertain Function Specification: How Should We Proceed? *Bull. Math. Biol.* 76, 1218–1240. doi:10.1007/s11538-014-9951-9

Adamson, M. W., and Morozov, A. Y. (2014b). Defining and Detecting Structural Sensitivity in Biological Models: Developing a New Framework. *J. Math. Biol.* 69, 1815–1848. doi:10.1007/s00285-014-0753-3

Adamson, M. W., Morozov, A. Y., and Kuzenkov, O. A. (2016). Quantifying Uncertainty in Partially Specified Biological Models: How Can Optimal Control Theory Help Us? *Proc. R. Soc. A.* 472, 20150627. doi:10.1098/rspa.2015.0627

Allen, B., Nowak, M. A., and Dieckmann, U. (2013). Adaptive Dynamics with Interaction Structure. *The Am. Naturalist* 181, E139–E163. doi:10.1086/670192

Allesina, S., and Tang, S. (2012). Stability Criteria for Complex Ecosystems. *Nature* 483, 205–208. doi:10.1038/nature10832

Anderson, K. E., and Fahimipour, A. K. (2021). Body Size Dependent Dispersal Influences Stability in Heterogeneous Metacommunities. *Sci. Rep.* 11, 17410. doi:10.1038/s41598-021-96629-5

Aufderheide, H., Rudolf, L., Gross, T., and Lafferty, K. D. (2013). How to Predict Community Responses to Perturbations in the Face of Imperfect Knowledge and Network Complexity. *Proc. Biol. Sci.* 280, 20132355. doi:10.1098/rspb.2013.2355

Aufderheide, H., Rudolf, L., and Gross, T. (2012). Mesoscale Symmetries Explain Dynamical Equivalence of Food Webs. *New J. Phys.* 14, 105014. doi:10.1088/1367-2630/14/10/105014

Awender, S., Wackerbauer, R., and Breed, G. A. (2021). Stability of Generalized Ecological-Network Models. *Chaos* 31, 023106. doi:10.1063/5.0029934

Barter, E., Brechtel, A., Drossel, B., and Gross, T. (2021). A Closed Form for Jacobian Reconstruction from Time Series and its Application as an Early Warning Signal in Network Dynamics. *Proc. R. Soc. A.* 477, 20200742. doi:10.1098/rspa.2020.0742

Baurmann, M., Gross, T., and Feudel, U. (2007). Instabilities in Spatially Extended Predator-Prey Systems: Spatio-Temporal Patterns in the Neighborhood of Turing-Hopf Bifurcations. *J. Theor. Biol.* 245, 220–229. doi:10.1016/j.jtbi.2006.09.036

Brechtel, A., Gramlich, P., Ritterskamp, D., Drossel, B., and Gross, T. (2018). Master Stability Functions Reveal Diffusion-Driven Pattern Formation in Networks. *Phys. Rev. E* 97, 032307. doi:10.1103/PhysRevE.97.032307

Brose, U., Williams, R. J., and Martinez, N. D. (2006). Allometric Scaling Enhances Stability in Complex Food Webs. *Ecol. Lett.* 9, 1228–1236. doi:10.1111/j.1461-0248.2006.00978.x

Bulik, S., Grimbs, S., Huthmacher, C., Selbig, J., and Holzhütter, H. G. (2009). Kinetic Hybrid Models Composed of Mechanistic and Simplified Enzymatic Rate Laws - a Promising Method for Speeding up the Kinetic Modelling of Complex Metabolic Networks. *FEBS J.* 276, 410–424. doi:10.1111/j.1742-4658.2008.06784.x

Carbonaro, N. J., and Thorpe, I. F. (2017). Using Structural Kinetic Modeling to Identify Key Determinants of Stability in Reaction Networks. *J. Phys. Chem. A.* 121, 4982–4992. doi:10.1021/acs.jpca.7b01852

Childs, D., Grimbs, S., and Selbig, J. (2015). Refined Elasticity Sampling for Monte Carlo-Based Identification of Stabilizing Network Patterns. *Bioinformatics* 31, i214–i220. doi:10.1093/bioinformatics/btv243

Demirel, G., MacCarthy, B. L., Ritterskamp, D., Champneys, A. R., and Gross, T. (2019). Identifying Dynamical Instabilities in Supply Networks Using Generalized Modeling. *Jrnl Ops Manage.* 65, 136–159. doi:10.1002/joom.1005

Do, A.-L., Höfener, J., and Gross, T. (2012). Engineering Mesoscale Structures with Distinct Dynamical Implications. *New J. Phys.* 14, 115022. doi:10.1088/1367-2630/14/11/115022

Doizy, A., Barter, E., Memmott, J., Varnham, K., and Gross, T. (2018). Impact of Cyber-Invasive Species on a Large Ecological Network. *Sci. Rep.* 8, 13245. doi:10.1038/s41598-018-31423-4

Fell, D. A., and Sauro, H. M. (1985). Metabolic Control and its Analysis. Additional Relationships Between Elasticities and Control Coefficients. *Eur. J. Biochem.* 148, 555–561. doi:10.1111/j.1432-1033.1985.tb08876.x

Frandi, A., Pini, F., Beroual, W., Bianchetti, A., Chiodi, A., Mascolo, E., et al. (2022). "Toward a Comparative Systems Biology of the Alphaproteobacterial Cell Cycle," in *Cell Cycle Regulation and Development in Alphaproteobacteria* (Cham: Springer), 1–27. doi:10.1007/978-3-030-90621-4_1

Fürtauer, L., and Nägele, T. (2016). Approximating the Stabilization of Cellular Metabolism by Compartmentalization. *Theor. Biosci.* 135, 73–87. doi:10.1007/s12064-016-0225-y

Fussmann, G. F., and Blasius, B. (2005). Community Response to Enrichment Is Highly Sensitive to Model Structure. *Biol. Lett.* 1, 9–12. doi:10.1098/rsbl.2004.0246

Gantmacher, F. R. (2000). *Matrix Theory*. Philadelphia: American Mathematical Society.

Gehrmann, E., and Drossel, B. (2010). Boolean versus Continuous Dynamics on Simple Two-Gene Modules. *Phys. Rev. E* 82, 046120. doi:10.1103/PhysRevE.82.046120

Gehrmann, E., Gläßer, C., Jin, Y., Sendhoff, B., Drossel, B., and Hamacher, K. (2011). Robustness of Glycolysis in Yeast to Internal and External Noise. *Phys. Rev. E* 84, 021913. doi:10.1103/PhysRevE.84.021913

Girbig, D., Grimbs, S., and Selbig, J. (2012a). Systematic Analysis of Stability Patterns in Plant Primary Metabolism. *PLoS ONE* 7, e34686–12. doi:10.1371/journal.pone.0034686

Girbig, D., Selbig, J., and Grimbs, S. (2012b). A Matlab Toolbox for Structural Kinetic Modeling. *Bioinformatics* 28, 2546–2547. doi:10.1093/bioinformatics/bts473

Gramlich, P., Plitzko, S. J., Rudolf, L., Drossel, B., and Gross, T. (2016). The Influence of Dispersal on a Predator-Prey System with Two Habitats. *J. Theor. Biol.* 398, 150–161. doi:10.1016/j.jtbi.2016.03.015

Grimbs, S., Selbig, J., Bulik, S., Holzhütter, H. G., and Steuer, R. (2007). The Stability and Robustness of Metabolic States: Identifying Stabilizing Sites in Metabolic Networks. *Mol. Syst. Biol.* 3, 146. doi:10.1038/msb4100186

Gross, T., and Feudel, U. (2006). Generalized Models as a Universal Approach to the Analysis of Nonlinear Dynamical Systems. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.* 73, 016205. doi:10.1103/PhysRevE.73.016205

Gross, T., Ebenhöh, W., and Feudel, U. (2004). Enrichment and Foodchain Stability: The Impact of Different Forms of Predator-Prey Interaction. *J. Theor. Biol.* 227, 349–358. doi:10.1016/j.jtbi.2003.09.020

Gross, T., Ebenhöh, W., and Feudel, U. (2005). Long Food Chains Are in General Chaotic. *Oikos* 109, 135–144. doi:10.1111/j.0030-1299.2005.13573.x

Gross, T., and Feudel, U. (2004). Analytical Search for Bifurcation Surfaces in Parameter Space. *Physica D: Nonlinear Phenomena* 195, 292–302. doi:10.1016/j.physd.2004.03.019

Gross, T. (2004). *Population Dynamics: General Results from Local Analysis*. Tönning: Der Andere Verlag.

Gross, T., Rudolf, L., Levin, S. A., and Dieckmann, U. (2009). Generalized Models Reveal Stabilizing Factors in Food Webs. *Science* 325, 747–750. doi:10.1126/science.1173536

Guckenheimer, J., and Holmes, P. (1983). *Nonlinear Oscillations, Dynamical Systems and Bifurcation of Vector Fields*. Heidelberg: Springer-Verlag.

Guckenheimer, J., Myers, M., and Sturmfels, B. (1997). Computing Hopf Bifurcations I. *SIAM J. Numer. Anal.* 34, 1–21. doi:10.1137/S0036142993253461

Henkel, S., Nägele, T., Hörmiller, I., Sauter, T., Sawodny, O., Ederer, M., et al. (2011). A Systems Biology Approach to Analyse Leaf Carbohydrate Metabolism in Arabidopsis Thaliana. *J. Bioinform. Sys. Biol.* 2011, 2. doi:10.1186/1687-4153-2011-2

Höfener, J. M., Sethia, G. C., and Gross, T. (2013). Amplitude Death in Networks of Delay-Coupled Delay Oscillators. *Phil. Trans. R. Soc. A.* 371, 20120462. doi:10.1098/rsta.2012.0462

Höfener, J. M., Sethia, G. C., and Gross, T. (2011). Stability of Networks of Delay-Coupled Delay Oscillators. *Epl* 95, 40002. doi:10.1209/0295-5075/95/40002

Holling, C. S. (1959). Some Characteristics of Simple Types of Predation and Parasitism. *Can. Entomol.* 91, 385–398. doi:10.4039/Ent91385-7

Kisdi, E., Geritz, S. A. H., and Boldin, B. (2013). Evolution of Pathogen Virulence Under Selective Predation: A Construction Method to Find Eco-Evolutionary Cycles. *J. Theor. Biol.* 339, 140–150. doi:10.1016/j.jtbi.2013.05.023

Knopf, B., Flechsig, M., and Zickfeld, K. (2006). Multi-Parameter Uncertainty Analysis of a Bifurcation Point. *Nonlin. Process. Geophys.* 13, 531–540. doi:10.5194/npg-13-531-2006

Kuehn, C., Gross, T., and Gross, T. (2013). Nonlocal Generalized Models of Predator-Prey Systems. *Discrete Continuous Dynamical Syst. - Ser. B* 18, 693–720. doi:10.3934/dcdsb.2013.18.693

Kuehn, C., Siegmund, S., and Gross, T. (2013). Dynamical Analysis of Evolution Equations in Generalized Models. *IMA J. Appl. Maths.* 78, 1051–1077. doi:10.1093/imamat/hxs008

Kuznetsov, Y. A. (2004). *Elements of Applied Bifurcation Theory*. New York: Springer-Verlag.

Lade, S. J., and Gross, T. (2012). Early Warning Signals for Critical Transitions: A Generalized Modeling Approach. *Plos Comput. Biol.* 8, e1002360–6. doi:10.1371/journal.pcbi.1002360

Lade, S. J., and Niiranen, S. (2017). Generalized Modeling of Empirical Social-Ecological Systems. *Nat. Resource Model.* 30, e12129. doi:10.1111/nrm.12129

Lade, S. J., Niiranen, S., and Schlüter, M. (2015). Generalized Modeling of Empirical Social-Ecological Systems. *ArXiv e-prints 1503.02846*.

Lade, S. J., Tavoni, A., Levin, S. A., and Schlüter, M. (2013). Regime Shifts in a Social-Ecological System. *Theor. Ecol.* 6, 359–372. doi:10.1007/s12080-013-0187-3

Leibold, M. A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J. M., Hoopes, M. F., et al. (2004). The Metacommunity Concept: a Framework for Multi-Scale Community Ecology. *Ecol. Lett.* 7, 601–613. doi:10.1111/j.1461-0248.2004.00608.x

May, R. M. (1972). Will a Large Complex System Be Stable? *Nature* 238, 413–414. doi:10.1038/238413a0

McCann, K., Hastings, A., and Huxel, G. R. (1998). Weak Trophic Interactions and the Balance of Nature. *Nature* 395, 794–798. doi:10.1038/27427

McCann, K. S., Rasmussen, J. B., and Umbanhowar, J. (2005). The Dynamics of Spatially Coupled Food Webs. *Ecol. Lett.* 8, 513–523. doi:10.1111/j.1461-0248.2005.00742.x

Murabito, E., Smallbone, K., Swinton, J., Westerhoff, H. V., and Steuer, R. (2011). A Probabilistic Approach to Identify Putative Drug Targets in Biochemical Networks. *J. R. Soc. Interf.* 8, 880–895. doi:10.1098/rsif.2010.0540

Murabito, E. (2013). Targeting Breast Cancer Metabolism: A Metabolic Control Analysis Approach. *Curr. Synth. Sys. Biol.* 1, 104. doi:10.4172/2332-0737.1000104

Neutel, A.-M., Heesterbeek, J. A. P., and de Ruiter, P. C. (2002). Stability in Real Food Webs: Weak Links in Long Loops. *Science* 296, 1120–1123. doi:10.1126/science.1068326

Pecora, L. M., and Carroll, T. L. (1998). Master Stability Functions for Synchronized Coupled Systems. *Phys. Rev. Lett.* 80, 2109–2112. doi:10.1103/PhysRevLett.80.2109

Plitzko, S. J., Drossel, B., and Guill, C. (2012). Complexity-stability Relations in Generalized Food-Web Models with Realistic Parameters. *J. Theor. Biol.* 306, 7–14. doi:10.1016/j.jtbi.2012.04.008

Reilly, E. E. (1940). The Use of the Elasticity Concept in Economic Theory: With Special Reference to Some Economic Effects of a Commodity Tax. *Can. J. Econ. Polit. Sci.* 6, 39. doi:10.2307/137053

Reznik, E., Kaper, T. J., and Segrè, D. (2013a). The Dynamics of Hybrid Metabolic-Genetic Oscillators. *Chaos* 23, 013132. doi:10.1063/1.4793573

Reznik, E., and Segrè, D. (2010). On the Stability of Metabolic Cycles. *J. Theor. Biol.* 266, 536–549. doi:10.1016/j.jtbi.2010.07.023

Reznik, E., Watson, A., and Chaudhary, O. (2013b). The Stubborn Roots of Metabolic Cycles. *J. R. Soc. Interf.* 10, 20130087. doi:10.1098/rsif.2013.0087

Ritterskamp, D., Demirel, G., MacCarthy, B. L., Rudolf, L., Champneys, A. R., and Gross, T. (2018). Revealing Instabilities in a Generalized Triadic Supply Network: A Bifurcation Analysis. *Chaos* 28, 073103. doi:10.1063/1.5026746

Scheff, J. D., Calvano, S. E., and Androulakis, I. P. (2013). Predicting Critical Transitions in a Model of Systemic Inflammation. *J. Theor. Biol.* 338, 9–15. doi:10.1016/j.jtbi.2013.08.011

Segel, L. A., and Levin, S. A. (1976). "Application of Nonlinear Stability Theory to the Study of the Effects of Diffusion on Predator-Prey Interactions," in AIP Conference Proceedings (American Institute of Physics), 123–152.

Srinivasan, S., Cluett, W. R., and Mahadevan, R. (2015). Constructing Kinetic Models of Metabolism at Genome-Scales: A Review. *Biotechnol. J.* 10, 1345–1359. doi:10.1002/biot.201400522

Steuer, R. (2007). Computational Approaches to the Topology, Stability and Dynamics of Metabolic Networks. *Phytochemistry* 68, 2139–2151. doi:10.1016/j.phytochem.2007.04.041

Steuer, R., Gross, T., Selbig, J., and Blasius, B. (2006). Structural Kinetic Modeling of Metabolic Networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 11868–11873. doi:10.1073/pnas.0600013103

Steuer, R., Nesi, A. N., Fernie, A. R., Gross, T., Blasius, B., and Selbig, J. (2007). From Structure to Dynamics of Metabolic Pathways: Application to the Plant Mitochondrial Tca Cycle. *Bioinformatics* 23, 1378–1385. doi:10.1093/bioinformatics/btm065

Stiefs, D., Gross, T., Steuer, R., and Feudel, U. (2008). Computation and Visualization of Bifurcation Surfaces. *Int. J. Bifurcation Chaos* 18, 2191–2206. doi:10.1142/s0218127408021658

Stiefs, D., van Voorn, G. A. K., Kooi, B. W., Feudel, U., and Gross, T. (2010). Food Quality in Producer-Grazer Models: A Generalized Analysis. *Am. Naturalist* 176, 367–380. doi:10.1086/655429

Stiefs, D., Venturino, E., and Feudel, U. (2009). Evidence of Chaos in Eco-Epidemic Models. *Math. Biosciences Eng.* 6, 855–871. doi:10.3934/mbe.2009.6.855

Tromeur, E., Rudolf, L., and Gross, T. (2016). Impact of Dispersal on the Stability of Metapopulations. *J. Theor. Biol.* 392, 1–11. doi:10.1016/j.jtbi.2015.11.029

van Voorn, G. A., Stiefs, D., Gross, T., Kooi, B. W., Feudel, U., and Kooijman, S. A. (2008). Stabilization Due to Predator Interference: Comparison of Different Analysis Approaches. *Math. Biosci. Eng.* 5, 567–583. doi:10.3934/mbe.2008.5.567

Wigner, E. P. (1955). Characteristic Vectors of Bordered Matrices with Infinite Dimensions. *Ann. Maths.* 62, 548. doi:10.2307/1970079

Williams, R. J., and Martinez, N. D. (2000). Simple Rules Yield Complex Food Webs. *Nature* 404, 180–183. doi:10.1038/35004572

Williams, R. J., and Martinez, N. D. (2004). Stabilization of Chaotic and Non-Permanent Food-Web Dynamics. *The Eur. Phys. J. B - Condensed Matter* 38, 297–303. doi:10.1140/epjb/e2004-00122-1

Ye, J., Xu, H., Huang, X., Ke, C., and Feng, E. (2015). Dynamics Analysis and Prediction of Genetic Regulation in Glycerol Metabolic Network via Structural Kinetic Modelling. *Math. Probl. Eng.* 2015, 1–9. doi:10.1155/2015/673120

Yeakel, J. D., Kempes, C. P., and Redner, S. (2018). Dynamics of Starvation and Recovery Predict Extinction Risk and Both Damuth's Law and Cope's Rule. *Nat. Commun.* 9, 657. doi:10.1038/s41467-018-02822-y

Yeakel, J. D., and Mangel, M. (2014). A Generalized Perturbation Approach for Exploring Stock Recruitment Relationships. *Theor. Ecol.* 8, 1–13. doi:10.1007/s12080-014-0230-z

Yeakel, J. D., Pires, M. M., Rudolf, L., Dominy, N. J., Koch, P. L., Guimarães, P. R., et al. (2014). Collapse of an Ecological Network in Ancient Egypt. *Proc. Natl. Acad. Sci. U.S.A.* 111, 14472–14477. doi:10.1073/pnas.1408471111

Yeakel, J. D., Stiefs, D., Novak, M., and Gross, T. (2011). Generalized Modeling of Ecological Population Dynamics. *Theor. Ecol.* 4, 179–194. doi:10.1007/s12080-011-0112-6

Zumsande, M., and Gross, T. (2010). Bifurcations and Chaos in the MAPK Signaling Cascade. *J. Theor. Biol.* 265, 481–491. doi:10.1016/j.jtbi.2010.04.025

Zumsande, M., Stiefs, D., Siegmund, S., and Gross, T. (2011). General Analysis of Mathematical Models for Bone Remodeling. *Bone* 48, 910–917. doi:10.1016/j.bone.2010.12.010

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Massing and Gross. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 04 May 2022
doi: 10.3389/fmolb.2022.874669

Check for updates

Discrete Biochemical Systems Theory

Eberhard O. Voit and Daniel V. Olivença*

The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA, United States

Almost every biomedical systems analysis requires early decisions regarding the choice of the most suitable representations to be used. De facto the most prevalent choice is a system of ordinary differential equations (ODEs). This framework is very popular because it is flexible and fairly easy to use. It is also supported by an enormous array of stand-alone programs for analysis, including many distinct numerical solvers that are implemented in the main programming languages. Having selected ODEs, the modeler must then choose a mathematical format for the equations. This selection is not trivial as nearly unlimited options exist and there is seldom objective guidance. The typical choices include ad hoc representations, default models like mass-action or Lotka-Volterra equations, and generic approximations. Within the realm of approximations, linear models are typically successful for analyses of engineered systems, but they are not as appropriate for biomedical phenomena, which often display nonlinear features such as saturation, threshold effects or limit cycle oscillations, and possibly even chaos. Power-law approximations are simple but overcome these limitations. They are the key ingredient of Biochemical Systems Theory (BST), which uses ODEs exclusively containing power-law representations for all processes within a model. BST models cover a vast repertoire of nonlinear responses and, at the same time, have structural properties that are advantageous for a wide range of analyses. Nonetheless, as all ODE models, the BST approach has limitations. In particular, it is not always straightforward to account for genuine discreteness, time delays, and stochastic processes. As a new option, we therefore propose here an alternative to BST in the form of discrete Biochemical Systems Theory (dBST). dBST models have the same generality and practicality as their BST-ODE counterparts, but they are readily implemented even in situations where ODEs struggle. As a case study, we illustrate dBST applied to the dynamics of the aryl hydrocarbon receptor (AhR), a signal transduction system that simultaneously involves time delays and stochasticity.

Keywords: canonical model, delay, discrete event, generalized mass action system, power-law approximation, system, stochastic event, aryl hydrocarbon receptor

OPEN ACCESS

Edited by:
Daniel Garrido, Pontificia Universidad Católica de Chile, Chile

Reviewed by:
Paola Lecca, Free University of Bozen-Bolzano, Italy
Marija Cvijovic, Chalmers University of Technology, Sweden

****Correspondence:***
Eberhard O. Voit
eberhard.voit@bme.gatech.edu

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *12 February 2022*
Accepted: *28 March 2022*
Published: *04 May 2022*

Citation:
Voit EO and Olivença DV (2022) Discrete Biochemical Systems Theory. Front. Mol. Biosci. 9:874669. doi: 10.3389/fmolb.2022.874669

INTRODUCTION

Arguably the greatest challenge of systems modeling in the biomedical sciences is the choice of optimal process representations. Often the true magnitude of this challenge is ignored and the modeler either constructs an *ad hoc* model or chooses a default, such as a Lotka-Volterra system for describing the interactions among competing populations (Volterra, 1926; Lotka, 1956; May, 1973) or a mass action formulation or some variation of the Michaelis-Menten rate law for enzyme

catalyzed processes (Michaelis and Menten, 1913; Voit et al., 2015). These default representations may be further extended or refined with the inclusion of environmental variables in a population model (Stein et al., 2013; Dam et al., 2020) or the inclusion of modulating effects, such as the regulation of a biochemical reaction through competitive or allosteric inhibition (Cornish-Bowden, 2012). Because there are no iron-clad rules for choosing a model, researchers often arrive at rather different formulations even for the same phenomenon. An illustrative example is the phosphofructokinase reaction in glycolysis, for which numerous rate functions of drastically different complexity have been proposed (Voit, 2017a). The choice of optimal representations becomes even more challenging at the intersections of typical biological domains, such as the combination of genetics, metabolism, and organismal physiology, because the default models of the various subdisciplines are different, thereby creating the need of multiscale models that operate at different temporal, spatial and organizational scales.

One could argue that biological processes must obey the laws of physics and that, therefore, optimal—or at least adequate—representations are prescribed. While this is true in a fundamental sense, most biological processes are so convoluted that exact physical representations of all contributing aspects become infeasible (Voit, 2008). As an example, consider the generation of two daughter cells from a bacterial mother cell. At a high level, one bacterium becomes two, two become four, and so on, and it is easy to formulate an exponential function that describes the progression well. However, if it is necessary to account for more details, for instance, in order to understand a mutant with aberrant behavior, it becomes clear that the cell division process is immensely complicated (Schafer, 1998; Carlton et al., 2020). It is multifaceted and involves so many different aspects at the molecular level that it is hardly possible to formulate the governing processes, proceeding in time and space, with elementary functions that are directly derived from the first principles of physics.

A second aspect of the challenge of biomedical systems modeling is the fact that it is usually difficult to capture the dynamics of a molecular or cellular component directly. Even the simple Michaelis-Menten rate law of enzyme kinetics does not prescribe the changing concentration of a substrate or product as the reaction progresses, but expresses the speed of the reaction as a function of the substrate concentration. By contrast, it is often feasible to characterize all influences that lead to an increase or decrease in a system component over time (Voit, 2020). Indeed, the literature contains uncounted articles about "the effect of . . . on . . .," which explicitly or implicitly describe how a target variable changes in response to some input. Thus, this view focuses on the change in a component, rather than the state of this component, and this change is driven by the totality of all contributing factors. A natural mathematical formulation of this situation is a system of ordinary differential equations (ODEs) which, after all, equate the instantaneous change in a variable to all processes affecting this variable. Consequently, the biomathematical literature contains an enormous body of work using ODEs to analyze biological systems [for introductory texts, see (Keshet, 2005; Klipp et al., 2016; Voit, 2017b)]. Even so, it must be kept in mind that ODEs are approximations of natural processes, which are often genuinely discrete (see **Supplementary Data S1, S2**).

While ODEs have become the standard modeling default, the conundrum of determining the best possible model structure persists. Two generic solutions are 1) the use of *ad hoc* representations that are often chosen simply for convenience and match the natural processes sufficiently well and 2) suitable, unbiased approximations. Among the latter, linear systems are most straightforward but are often at odds with the genuine nonlinearities of biomedical systems. A prominent alternative is Biochemical Systems Theory (BST) (Savageau, 1976; Voit, 2000; Torres and Voit, 2002; Voit, 2013), which uses power-law representations for all processes, thereby creating highly structured nonlinear models in immutable, predefined formats (**Supplementary Data S1**).

Independent of what representations are chosen to design ODE models, the ODE format in itself faces a number of challenges. Of particular prominence among these are time delays and stochastic effects (**Supplementary Data S2**). Sometimes, these can be addressed with sophisticated numerical ODE solvers, but the formulation and implementation can quickly become convoluted and often requires intimate knowledge of the inner workings of these solvers.

An illustrative example for the crucial role of delays is a situation that arose when we analyzed the dynamics of anemia during malaria, a disease that is caused by *Plasmodium* parasites that invade red blood cells and eventually cause them to burst. Red blood cells furthermore disappear in large numbers due to a so-called bystander effect, in which many non-infected red blood cells perish for reasons that are not well understood. One difficult challenge that arose during our modeling attempts was the fact that red blood cells naturally have a narrowly determined life span with rather small variation; in humans, it is about 115 days ± 15% (Franco, 2012). The modeling challenge becomes apparent in the assessment of how many cells are expected to disappear at a given time point during the infection (Fonseca and Voit, 2015; Fonseca et al., 2016). Some disappear due to the infection or the bystander effect, but many are removed by the spleen because they have reached the end of their natural life. To account for the latter aspect, one needs to know the age of each cell at any given point in time. However, ODEs do not recount the ages of individual cells. Thus, the disappearance of cells from the blood stream must be based on averages, which are adequate under steady-state conditions, but not for dynamic changes caused by the growing parasite population. Even the use of delay differential equations (DDEs) is inconvenient in this case, whereas a discrete, recursive modeling approach is straightforward (Fonseca and Voit, 2015; Fonseca et al., 2016). Other pertinent examples of delays and stochasticity are presented in **Supplementary Data S2**.

This article proposes an alternative to BST models that facilitates the modeling of genuine discreteness, delays, small numbers of components, stochastic events and combinations of these complicating factors. This alternative consists of a discrete,

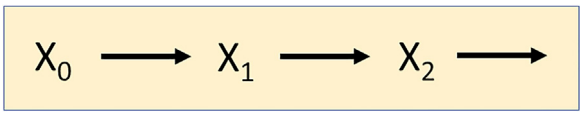


FIGURE 1 | Simple linear pathway with constant input X_0.

recursive version of BST, here dubbed "dBST," which is straightforwardly constructed and implemented.

RESULTS

For the practicing computational modeler in the biosciences, a partial solution to the drawbacks of ODEs can be the use of systems of discrete-time, recursive equations, where the changes in variables are represented on the basis of power-law functions, as is the case in BST. This replacement of ODEs with recursive equations raises the immediate question whether any genuine features of ODE models are lost. The answer can be approached in two ways. First, it is rather evident that the recursive equations converge to the ODEs if the step size decreases to 0 in the limit. In fact, computer algorithms for solving ODEs use small discrete step sizes. Second, one may test whether representative nonlinear phenomena that are typically represented with ODEs, such as saturation, limit cycles, and deterministic chaos, can also be represented through recursive equations with a reasonable step size. **Supplementary Data S3** discusses mathematical similarities between BST and dBST systems. Here, we focus on the response repertoire of dBST systems and their features. We also present a case study illustrating the *de novo* design of a dBST system that simultaneously accounts for both, delays and stochasticity.

Response Repertoire of Discrete Biochemical Systems Theory Models

Simple Introductory Example of a Linear Pathway

The 2-variable BST system

$$\begin{aligned}\dot{X}_1 &= 2X_0 - rX_1^{0.8}\\ \dot{X}_2 &= rX_1^{0.8} - 2.5X_2^{0.5}\end{aligned} \qquad (1)$$

$$\begin{aligned}X_0 &= 1\\ X_1(t_0) &= 1.18\\ X_2(t_0) &= 0.64\end{aligned}$$

represents a simple linear pathway with constant input $X_0 = 1$ and rate $r = 1.75$ for the conversion of X_1 into X_2. The pathway is shown in **Figure 1**:

As an illustration, the system is initiated very close to its steady state (1.181653, 0.64). At t = 5, the input X_0 is persistently increased by 20%. Solving the equations shows that the system responds to the changed input by approaching a new steady state (1.484, 0.922) (**Figure 2**).

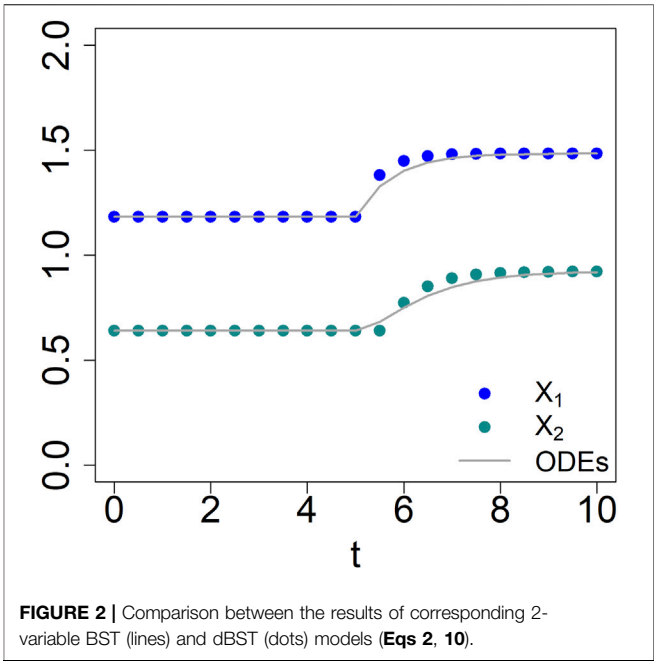


FIGURE 2 | Comparison between the results of corresponding 2-variable BST (lines) and dBST (dots) models (**Eqs 2**, **10**).

The corresponding dBST system in standard notation (**Supplementary Data S3**) is

$$\begin{aligned}X_{1,q\cdot\vartheta} &= X_{1,(q-1)\cdot\vartheta} + \vartheta\left[2X_0 - rX_{1,(q-1)\cdot\vartheta}^{0.8}\right]\\ X_{2,q\cdot\vartheta} &= X_{2,(q-1)\cdot\vartheta} + \vartheta\left[rX_{1,(q-1)\cdot\vartheta}^{0.8} - 2.5X_{2,(q-1)\cdot\vartheta}^{0.5}\right]\end{aligned} \qquad (2a)$$

To simplify this notation for easier reading, we rename $\tilde{X}_i = X_{i,q\cdot\vartheta}$ and $X_i = X_{i,(q-1)\cdot\vartheta}$, for $i = 1, 2$, which simplifies the appearance of **Eq. 2a** to

$$\begin{aligned}\tilde{X}_1 &= X_1 + \vartheta\left[2X_0 - rX_1^{0.8}\right]\\ \tilde{X}_2 &= X_2 + \vartheta\left[rX_1^{0.8} - 2.5X_2^{0.5}\right]\end{aligned} \qquad (2b)$$

Choosing as step size $\vartheta = 0.5$ reveals output that is quite similar to that of the BST system, although one notes that the responses, especially of X_2, are slightly different immediately following the switch in input (t = 5). Importantly, both formulations exhibit essentially the same dynamics and approach exactly the same steady state (**Figure 2**). Other step sizes yield similar results.

Limit Cycles

Limit cycles are representations of oscillations that are stable in a sense that, when perturbed by external influences, return to the original frequency and amplitude. Limit cycles are ubiquitous in biology (Keshet, 2005).

It has been proposed that many disease patterns can be seen mathematically as shifts from physiological to pathological limit cycles (Claude, 1995).

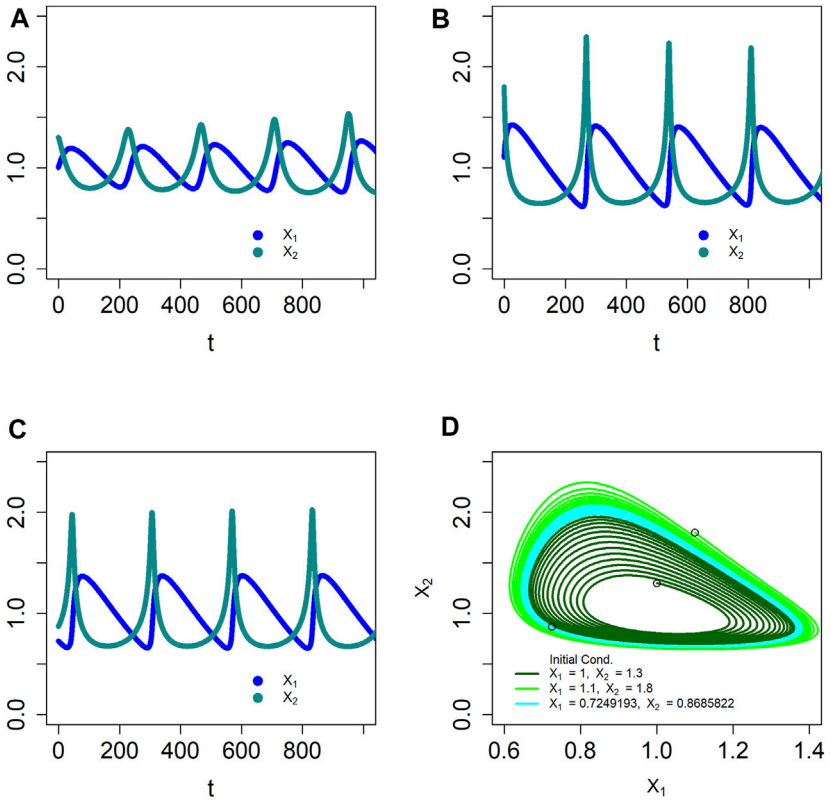


FIGURE 3 | The dBST system in **Eq. 4** models a stable limit cycle, as confirmed by simulations starting inside **(A)**, outside **(B)** and essentially on the limit cycle **(C)**. Panel **(D)** displays a typical phase-plane plot with superimposed oscillations spiraling out (dark green) or in (light green) toward the limit cycle, as well as starting very close to the limit cycle itself (cyan); the initial locations are indicated by circles. The trajectories appear to be smooth because the step size is rather small. The ODE model produces essentially the same solutions, even though the maximal amplitudes are slightly different.

Like BST systems, dBST systems can capture the dynamics of stable limit cycles (Lewis and Voit, 1991; Yin and Voit, 2008). An example is the stable oscillator

$$\begin{aligned}\dot{X}_1 &= 0.011\,\left[X_1^2X_2^3 - X_1X_2\right]\\ \dot{X}_2 &= 0.01\left[X_1^{-1}X_2^4 - X_1^3X_2^5\right]\end{aligned} \tag{3}$$

which in standard dBST format reads

$$\begin{aligned}X_{1,q\cdot\vartheta} &= X_{1,(q-1)\cdot\vartheta} + \vartheta\cdot 0.011\cdot\left[X^2_{1,(q-1)\cdot\vartheta}\,X^3_{2,(q-1)\cdot\vartheta}\right.\\ &\quad\left. - X_{1,(q-1)\cdot\vartheta}\,X_{2,(q-1)\cdot\vartheta}\right]\\ X_{2,q\cdot\vartheta} &= X_{2,(q-1)\cdot\vartheta} + \vartheta\cdot 0.01\cdot\left[X^{-1}_{1,(q-1)\cdot\vartheta}X^4_{2,(q-1)\cdot\vartheta}\right.\\ &\quad\left. - X^3_{1,(q-1)\cdot\vartheta}X^5_{2,(q-1)\cdot\vartheta}\right]\end{aligned} \tag{4a}$$

To the human eye, this format may look rather unwieldy, but it is easily implemented into computer code. Furthermore, using the simplified notation introduced in **Eq. 2**, we obtain

$$\begin{aligned}\tilde{X}_1 &= X_1 + \vartheta\cdot 0.011\cdot\left[X_1^2X_2^3 - X_1X_2\right]\\ \tilde{X}_2 &= X_2 + \vartheta\cdot 0.01\cdot\left[X_1^{-1}X_2^4 - X_1^3X_2^5\right]\end{aligned} \tag{4b}$$

Solving the system with step size $\vartheta = 0.1$ confirms that the system indeed has a stable limit cycle. Namely, initial conditions inside the limit cycle, like $(X_{1,0}, X_{2,0}) = (1, 1.3)$, lead to increasing oscillations, while conditions outside, such as $(X_{1,0}, X_{2,0}) = (1.1, 1.8)$, generate damped oscillations; starting essentially on the limit cycle, e.g., $(X_{1,0}, X_{2,0}) = (0.7249193, 0.8685822)$ demonstrates constant amplitudes. In the phase plane, the corresponding plots are outward and inward spirals, as well as the stable orbit (**Figure 3**). If we use the step size $\vartheta = 1$, the system still displays a limit cycle of similar shape, but with larger amplitudes (not shown). Much larger step sizes eventually become too coarse and destroy the features of the limit cycle.

Deterministic Chaos

Discrete BST systems are also rich enough to permit deterministic chaos. While a rigorous proof is difficult, an example is a discrete system gleaned from the well-known Lorenz oscillator (Lorenz, 1963), which mathematician and meteorologist Edward Lorenz developed as a simplified representation of atmospheric convection, which had been modeled previously as a fluid layer for which the temperatures at the top and the bottom were kept constant

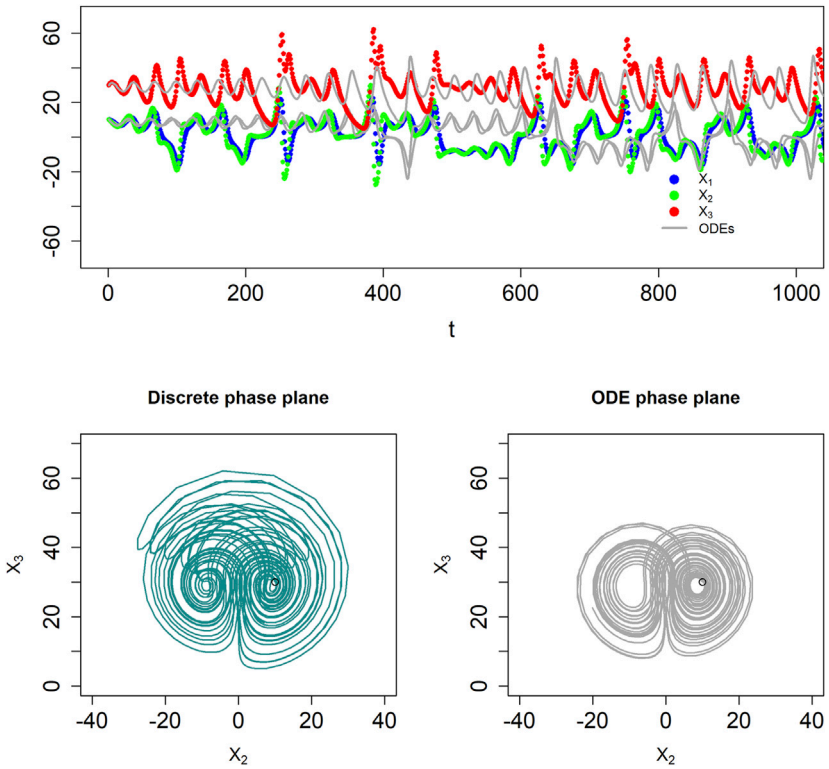


FIGURE 4 | The dBST system in **Eq. 6** captures deterministic chaos, similar to the ODE system proposed by Lorenz. The top panel shows results in the time domain, in comparison to the corresponding ODE system (thin grey lines). The BST and dBST systems diverge quickly (top panel), which is a genuine feature of chaotic systems. The bottom panel displays phase-plane plots of the two models, showing the discrete nature of the system in the form of connected straight lines. The initial locations are indicated with circles. Note that the maximal amplitudes of the dBST system are larger.

at different values (Saltzman, 1962). The ODE format of this system reads:

$$\begin{aligned} \dot{X}_1 &= 0.2 \cdot X_2 - 0.2 \cdot X_1 \\ \dot{X}_2 &= 0.6 \cdot X_1 - 0.02 \cdot X_2 - 0.02 \cdot X_1 \cdot X_3 \\ \dot{X}_3 &= 0.02 \cdot X_1 \cdot X_2 - 0.05 \cdot X_3 \end{aligned} \tag{5}$$

The reformulation into recursive equations is straightforward, and the dBST in simplified notation (see **Eq. 2**) reads

$$\begin{aligned} \tilde{X}_1 &= X_1 + \vartheta \cdot 0.2 \cdot [X_2 - X_1] \\ \tilde{X}_2 &= X_2 + \vartheta \cdot [0.6 \cdot X_1 - 0.02 \cdot X_2 - 0.02 \cdot X_1 \cdot X_3] \\ \tilde{X}_3 &= X_3 + \vartheta \cdot [0.02 \cdot X_1 \cdot X_2 - 0.05 \cdot X_3] \end{aligned} \tag{6}$$

Solving the system with step size $\vartheta = 1$ and initial conditions (10, 10, 30) reveals dynamics similar to that of the chaotic Lorenz equations in ODE format, both in the time domain and in phase plane (**Figure 4**), although the numerical details of the results are different, which is not surprising, as chaotic systems are extremely sensitive to all numerical settings, such as parameter values, initial conditions, and the step size for solving the ODEs or the discrete equations. One also notes that the maximum amplitudes of the BST system are somewhat higher than in the ODE model.

Typical Simulations in dBST That Are More Intuitive Than in ODE Models

The simulations described in this section are straightforwardly implemented in dBST, and while it is possible to implement some of them in ODEs, such an implementation is sometimes cumbersome or difficult to intuit. Indeed, one may have to be creative if some of these issues are to be included into ODE solutions and understand the inner workings of the numerical solution algorithms. For example, the widely used Runge-Kutta method averages the slope for each solution step, and additional statements, such as if-conditions, can influence this average or cannot be taken into consideration, depending on how the solver was coded. Furthermore, using numerical solvers with variable step size requires care so that the choice of the optimal step size is not affected.

In the following, we focus on different types of stochastic events and the dependence of the system dynamics on thresholds for dependent variables. Details regarding delays are discussed in **Supplementary Data S2** and in the later Case Study *Aryl Hydrocarbon Receptor Signal Transduction*.

Stochastic Variations in Rates

Returning to the introductory at the beginning of the *Results* section, it is easy in dBST to replace the constant rate r of the process converting X_1 into X_2 with a rate that stochastically varies

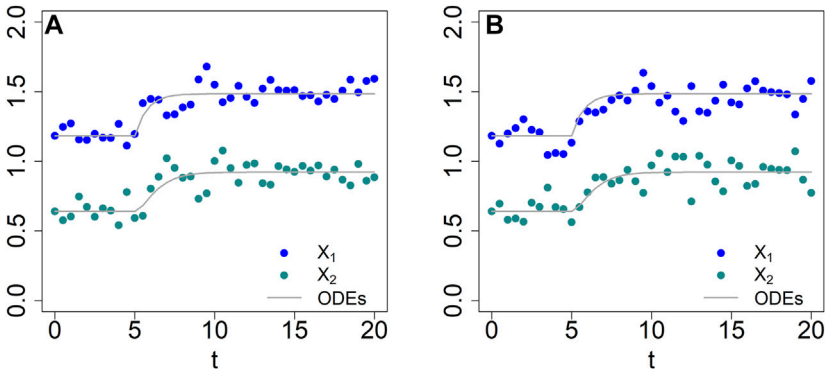


FIGURE 5 | Comparison of results from the deterministic and two instances **(A and B)** of stochastic models ($\vartheta = 0.5$) of the simple pathway in **Figure 2**. The two dBST simulations were obtained with a stochastically varying rate r in **Eq. 1** (dots), starting from different seeds, while the solid lines are the results of **Eq. 2** with constant rate r.

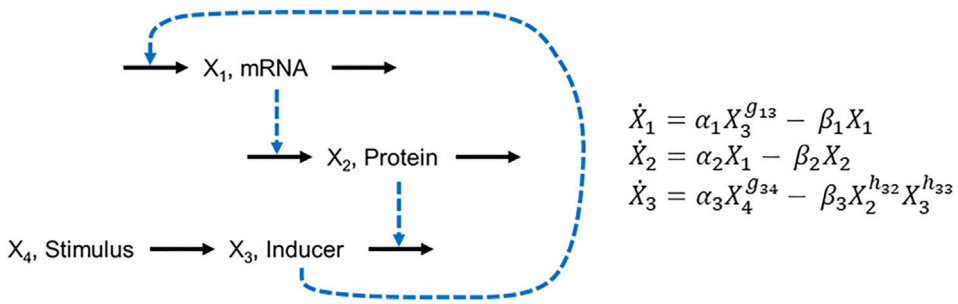


FIGURE 6 | Diagram of the lac operon and corresponding equations, as proposed by Savageau (2001).

within a range of, say, ± 10% of the nominal value in the example. For this illustration, we randomly sampled a rate from this range at every iteration. Two solutions are shown in **Figure 5**. Variations on this theme are also readily implemented. For instance, it is possible to sample a new rate less frequently than at every step.

Events Where System Variables Affect External Events

Suppose a signaling system responds stochastically to an environmental trigger, which is a ubiquitous situation in biological systems, especially if the concept of an "environment" includes the biophysical surroundings of cells. We develop this example in several steps, because some cases are easily addressed with ODEs, whereas others are not. In the simplest case of a stochastic environmental input it is possible to include if-statements into a numerical solver, such as the deSolve R package (Soetaert et al., 2010), which was designed to solve various initial-value problems, differential algebraic equations and partial differential equations. However, if one or more of the system variables influence the random variable, the situation is much more complicated, as the random variable must be adressed inside the solver, which is difficult for a ODE solver but straightforward in the discrete case.

Suppose at first that the environmental trigger is present or absent for stochastically long time periods that begin at random time points and whose magnitude affects the response of the signaling system. As a specific example, consider the lac operon of the bacterium *E. coli*, where external lactose triggers changes in gene expression (Lewis, 2005). Savageau (2001) proposed a model of the system in the form of the diagram in **Figure 6** and represented it with S-system equations. In this model, X_1 is the concentration of mRNA of the lac operon, X_2 is the concentration of the enzyme β-galactosidase, which catalyzes the conversion of lactose into galactose and glucose, and X_3 and X_4 are the intracellular and extracellular concentrations of lactose, respectively. X_4 is considered an independent variable and therefore that does not require its own differential equation.

Savageau discussed the kinetic orders (g and h parameters) but did not provide specific parameter values for them or for the rate constants (α and β parameters). We use this example for a series of demonstrations, specifying the parameter values as shown in **Eq. 7**.

$$\begin{aligned} \dot{X}_1 &= 0.2X_3^2 - 0.1X_1 \\ \dot{X}_2 &= 0.5X_1 - X_2 \\ \dot{X}_3 &= 0.1X_4 - 0.1X_2X_3 \end{aligned} \tag{7}$$

For the first demonstration, suppose that the stimulus, external lactose (X_4), is available in irregular time periods and concentrations that vary randomly within reasonable ranges. If

TABLE 1 | A priori known schedule of switches in stimulus in the lac operon model.

Time	0	10	60	110	160	210	260	310
Stimulus	0.1	1	0.5	0.1	1.2	0.4	1	0.1

the switch points and magnitudes of X_4 are known beforehand, the simulation of the ODE system is straightforward. For instance, if the system starts at its steady state (0.43, 0.22, 0.46) with $X_4 = 0.1$ and the stimulus changes according to **Table 1**, one directly obtains the output in **Figure 7A**.

If the timing and magnitudes are stochastic, a simulation with a standard ODE solver can be cumbersome as one needs to know how to evaluate functions of time inside the algorithm. Nonetheless, this situation can still be addressed, for instance, with deSolve or in Matlab. This case is again straightforwardly implemented in a dBST system. One set of results is shown in **Figure 7B** for event times sampled from an exponential distribution with a rate of 1/60. The magnitude of the signal at every event was sampled from a normal distribution with mean $\mu = 0.5$ and standard deviation $\sigma = 0.25$.

As the next phase of the example, we analyze the situation discussed by in Savageau (2001), where the format of the first ODE in **Figure 6** depends on the current value of X_3. Specifically, the author defined

$$\dot{X}_1 = \begin{cases} \alpha_{1L} - \beta_1 X_1 & if\ X_3 < X_{3L} \\ \alpha_1 X_3^{g_{13}} - \beta_1 X_1 & if\ X_{3L} \le X_3 \le X_{3H} \\ \alpha_{1H} - \beta_1 X_1 & if\ X_3 > X_{3H} \end{cases} \quad (8)$$

where X_{3L} and X_{3H} are threshold values and the corresponding "low" and "high" rate constants α_{3L} and α_{3H} are different. While it is possible to address this task by embedding if-conditions into an ODE solver, these situations of thresholds are much more easily called up in dBST: The If-statements are directly implemented in the recursive step for variable X_1. As a demonstration, suppose again that the external trigger changes in unpredictable patterns, which we assume to be random in terms of timing and magnitude, as before, and that the thresholds are in effect. Two typical results are shown in **Figures 7C,D**.

As the most complicated variation of the example, let us now suppose that both the magnitude and frequency of the stochastic events depend on the state of the system. For instance, the amount of external lactose X_4 to be imported into the cell could stochastically depend on the current mRNA prevalence X_1 and also the internal lactose concentration X_3, which together could have an effect on the characteristics of the import transporters. This situation cannot easily be addressed with an ODE solver, if at all, as one can no longer first calculate the current level of X_4 and then present it to the solver as a time-dependent function. Instead, such a situation mandates that the stochastic variables be evaluated inside the solver, a process that can interfere with the ODE solution. By contrast, the discrete version is easily implemented. An example is presented in **Figure 7E**.

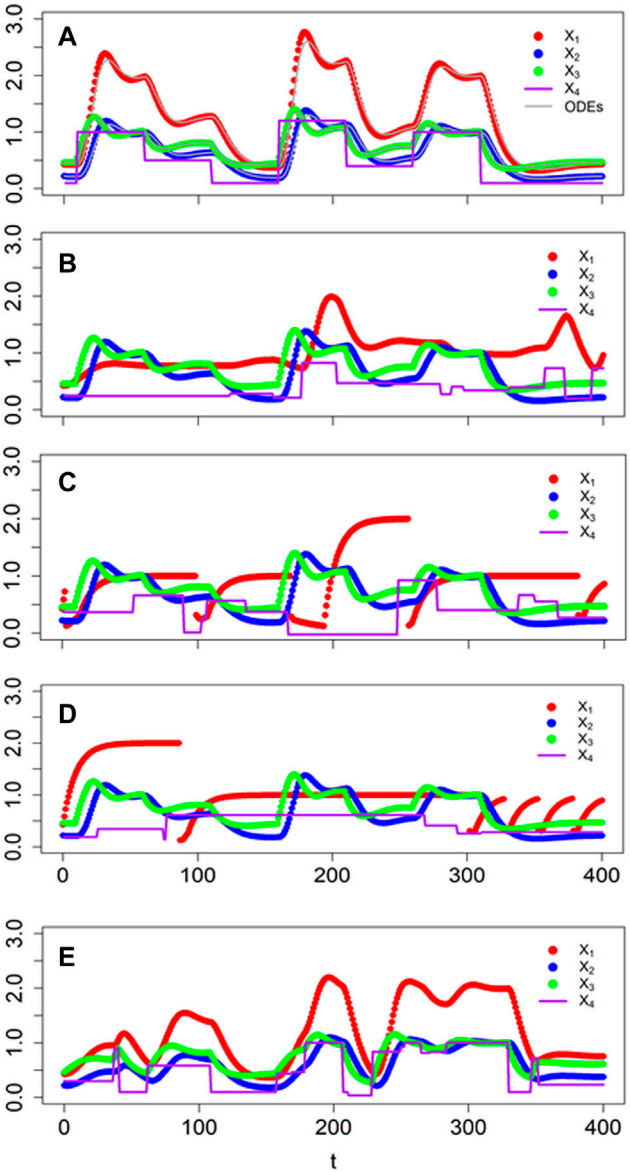


FIGURE 7 | Comparison of responses of BST (faint grey lines) and dBST models (dots) to perturbations in external lactose concentraion (solid lavendar line). **(A)** Timing and magnitudes are known *a priori* (**Table 1**). **(B)** Signaling events occur in a stochastic manner and signal magnitudes are random within a given range (see Text for details). **(C,D)** Two sets of responses, where switches in the equation of X_1 depend not only on the stochastic input X_4, but also on the value of X_3. **(E)** The value of X_3 is used to generate a success probability for a Bernoulli random variable. Specifically, if the Bernoulli process returns 1, the new value for X_4 is given by a truncated normal with mean 0.5 and a standard deviation equal to the value of X_1.

Parameter Estimation

Much has been written about the estimation of parameter values of ODE systems from time series data [*e.g.,* see reviews (Gennemark and Wedelin, 2007; Chou and Voit, 2009; Gennemark and Wedelin, 2009; Gábor and Banga, 2015)]. In order to make typical gradient methods and evolutionary

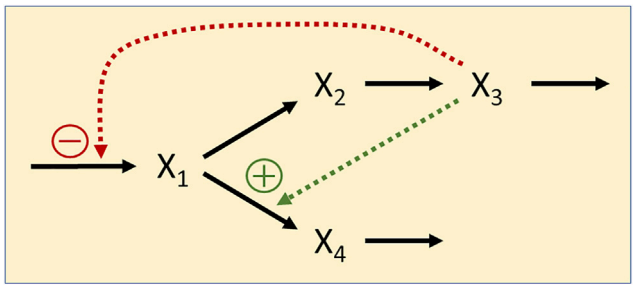


FIGURE 8 | Simple branched feedback with dual regulation by X_3.

algorithms more efficient, it has been suggested to smooth the time series data, use points from the smoothed trends, and estimate slopes of the time trends corresponding to the chosen points (Varah, 1982; Voit and Savageau, 1982; Voit and Almeida, 2004). Substituting these quantities in the ODE system converts the task of estimating parameter values for ODEs into an estimation from algebraic equations. Thus, for estimation purposes, the ODE for each variable X_i,

$$\dot{X}_i = F_i(X_1, \ldots, X_n) \quad i = 1, \ldots, n \tag{9}$$

is converted into a system of K algebraic equations of the format

$$\begin{aligned} S_i(t_1) &= F_i(X_1(t_1), \ldots, X_n(t_1)) \\ S_i(t_2) &= F_i(X_1(t_2), \ldots, X_n(t_2)) \\ S_i(t_K) &= F_i(X_1(t_K), \ldots, X_n(t_K)) \end{aligned} \tag{10}$$

Here, each equation corresponds to one chosen time point. The X values are either the raw data or the corresponding data from the smoothed time trend, while S indicates the corresponding slope of the time course.

This method of using estimated values and slopes tends to be computationally much more efficient than parameter inferences directly from ODEs, for instance with a gradient method or an evolutionary algorithm (Voit and Almeida, 2004). One drawback is that the estimation of slopes exacerbates noise in the data (Knowles and Renka, 2014; Voit, 2017b). To some degree, this problem is alleviated by smoothing the data appropriately.

For the discrete system, no slopes need to be estimated as the difference to be used instead, $X_{i,q\cdot\vartheta} - X_{i,(q-1)\cdot\vartheta}$, is directly obtained from the data. Thus, given measurements for all X_i at different time points, and possibly a smoothing step, the estimation of the parameters of a dBST system is straightforward.

As an example, consider the branched pathway in **Figure 8**, for which we pretend to have experimental measurements that had been smoothed, for instance, with a spline. For the illustration, we actually created synthetic "data" from a GMA (BST) model in ODE format and did not worry about noise, in order to assess most clearly to what degree BST and dBST models correspond and reflect the synthetic data. Analogously to the BST model, the format of the dBST equations is dictated directly by the flow structure and regulation of the pathway. In the simplified notation of **Eq. 2b**, the dBST equations take the form

$$\begin{aligned} \tilde{X}_1 &= X_1 + \vartheta \cdot \left[a_1 \cdot X_3^{g_1} - b_1 \cdot X_1^{h_{11}} - c_1 \cdot X_1^{h_{12}} \cdot X_3^{h_{13}}\right] \\ \tilde{X}_2 &= X_2 + \vartheta \cdot \left[b_1 \cdot X_1^{h_{11}} - b_2 \cdot X_2^{h_2}\right] \\ \tilde{X}_3 &= X_3 + \vartheta \cdot \left[b_2 \cdot X_2^{h_2} - b_3 \cdot X_3^{h_3}\right] \\ \tilde{X}_4 &= X_4 + \vartheta \cdot \left[c_1 \cdot X_1^{h_{12}} \cdot X_3^{h_{13}} - c_2 \cdot X_4^{h_4}\right] \end{aligned} \tag{11}$$

and we suppose that the values of the parameters are unknown. We analyze datasets with different densities of observation time points. For each parameter optimization, we use the optim function in R (R Core Team, 2018), which is based on the Nelder-Mead method (Nelder and Mead, 1965). Multiple sequential optimizations were performed for each example and the process was stopped when the difference of consecutive errors was less than 10^{-3}.

For the first illustration, we suppose that data had been obtained in intervals of τ =1, that is, for t = 0, 1, 2, ..., 60, and define ϑ = 1. The estimation result, shown in **Figure 9A**, captures the data well. The associated residual error, divided by the total number of data in the four time courses (4 n) is SSE/4 n = 0.1007525/(4 * 61) = 4.13×10^{-4}. Fitting the same data, but with step size ϑ = 0.5 (results not shown), the fitting error is roughly halved, with SSE/4 n = 0.0443176/(4 * 61) = 1.81×10^{-4}.

As a second illustration, we assume that the data are much sparser (τ =3), that is, with measurements obtained at time points t = 0, 3, 6, 9, . . . , 60; we again define ϑ = 1. The result is shown in **Figure 9B**. The model still fits the data well, with a residual error, divided by the total number 4 n of data, of SSE/4 n = 0.03697914/(4*21) = 4.40×10^{-4}, which is slightly, but not substantially higher than for the denser dataset. The estimated parameter values are slightly different from those obtained for the denser dataset, which is not surprising. However, it is interesting that the sparsity of the data hardly seems to affect the estimation.

Quasi as a baseline for comparison, we also fit the synthetic data with ODE equations in GMA format. They also recapture the data well (**Figure 9C**), even though the estimated parameter values are not identical to those used to create the data (**Table 2**), indicating some numerical redundancy among the parameters. The residual error, divided by the number of data is SSE/4 n = 0.03966285/(4*61) = 1.62×10^{-4}, which is again in the same range as for the discrete model. The parameter values are slightly different from those estimated with the dBST model. This result is to be expected because the meaning of each multiplicative parameter is, strictly speaking, not identical for BST and dBST models, as the former represent instantaneous rates and the latter stepwise changes.

Case Study: Aryl-Hydrocarbon Receptor Signal Transduction

The aryl-hydrocarbon receptor (AhR) is a highly conserved sensor for specific cues during development and normal physiology (Stockinger et al., 2014; Brinkmann et al., 2019; Zhu et al., 2019), as well as for external, xenobiotic compounds (Stevens et al., 2009; Simon et al., 2015) or danger signals derived from the invasion of parasites, which are mediated through compounds like the tryptophan-derivative kynurenine

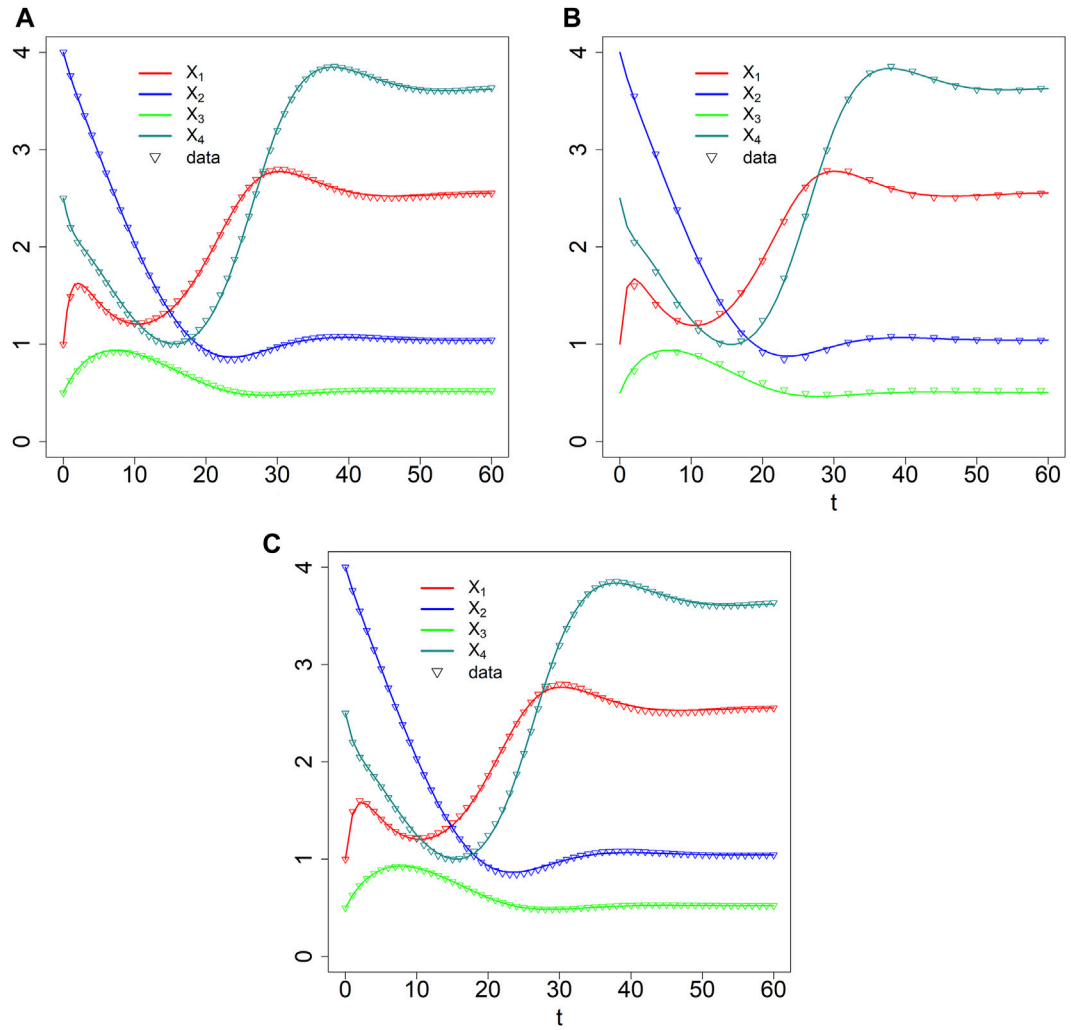


FIGURE 9 | Data fits with dBST [τ = 1 **(A)** and τ = 3 **(B)**] and BST **(C)** models.

(Julliard et al., 2014; Gupta et al., 2022). In response to such signals, the AhR signal transduction system triggers the upregulation of a host of genes, most prominently those coding for cytochrome P450 enzymes that metabolize toxicants.

The generic functionality of the AhR-system is depicted in **Figure 10**. Once a ligand (L) binds to AhR, the activated AhR forms a complex with the AhR nuclear translocator ARNT. This complex translocates to the nucleus, where it serves as a transcription factor that binds to the xenobiotic response element XRE—or a non-canonical XRE analog—within the promotor regions of numerous inducible target genes (Huang and Elferink, 2012). The AhR repressor AhRR competes with AhR for ARNT (Evans et al., 2008). Intriguingly, the gene coding for AhRR is itself under the control of the AhR-ARNT transcription factor, thereby creating a negative feedback loop that eventually stops the expression of AhR-ARNT controlled genes (Zudaire et al., 2008). As one might expect, reality is more complicated, for instance, due to compounds like the hypoxia inducible factor-1α (HIF1α) that compete with AhR and AhRR for ARNT (Spence et al., 1970) and to several cofactors modulating the process (Simon et al., 2015), but the AhR-ARNT-AhRR system by itself contains enough interesting complexity for the present illustration.

One issue in setting up an ODE model is the substantial time delay between transcription factor binding, the actual availability of AhRR, and the resulting repression of target gene expression (Koussounadis et al., 2015). In yeast and mouse, this type of delay was found to be at the order of 3–6 h (Fournier et al., 2010) (Liu et al., 2016). A delay of this magnitude is crucial in the AhR system, as it noticeably delays the inhibitory effect of AHRR on target gene expression.

A second issue is the fact that transcription and translation are known to be stochastic processes (Raj and van Oudenaarden, 2008). In fact, at least in some cases, activation of a promotor causes the production of proteins to occur in short bursts and yields variable protein amounts that occur at random time intervals (McAdams and

TABLE 2 | Parameter estimates obtained for different settings of the dBST model in **Eq. 11** and the corresponding BST model.

Parameter	Value for "data" generation	dBST (τ =1) ϑ = 1	dBST (τ =1) ϑ = 0.5	dBST (τ = 3) ϑ = 1	BST
X_0	1	0.981337	0.826518	0.951492	1.553729
a_1	0.25	0.263397	0.279577	0.263219	0.15743
b_1	0.08	0.166199	0.132337	0.159418	0.140825
b_2	0.2	0.276071	0.245437	0.265376	0.249848
b_3	0.3	0.395566	0.355906	0.386093	0.351592
c_1	0.15	0.091586	0.091028	0.089742	0.097641
c_2	0.25	0.169922	0.172505	0.166313	0.173117
d_1	0.1	4.775786	10.00496	8.298336	−0.02349
g_1	−2	−1.71411	−1.94822	−1.76486	−2.03513
h_{11}	1	0.55773	0.677036	0.56065	0.625422
h_{12}	2	1.736943	2.072465	1.794026	2.459446
h_{13}	0.5	−0.29347	0.072883	−0.24301	0.619297
h_2	0.4	0.327098	0.355548	0.345685	0.358772
h_3	0.6	0.498803	0.531523	0.525627	0.504745
h_4	0.8	0.935483	0.968566	0.950606	1.038234
SSQ/n		4.13×10^{-4}	1.81×10^{-4}	4.40×10^{-4}	1.62×10^{-4}

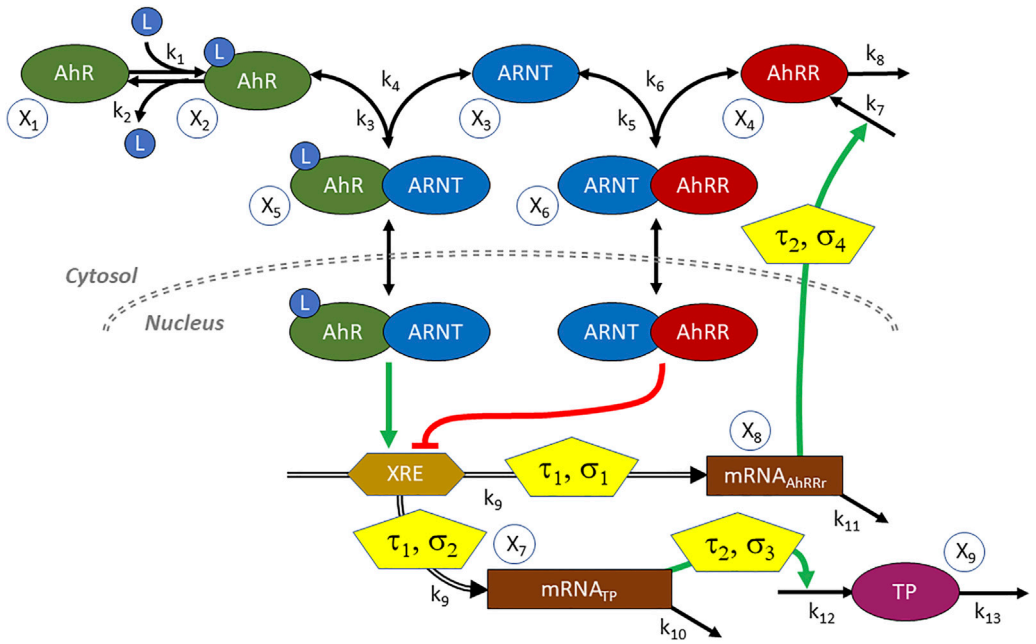


FIGURE 10 | Diagram of the AhR signal transduction system. AhR is activated by a ligand L and binds to the nuclear translocator ARNT. The complex serves as a transcription factor of genes whose promoter regions contain the xenobiotic response element XRE. These genes code for a variety of target proteins (TP) including, notably, the AhR repressor AhRR. AhRR competes with AhR for ARNT, and the complex inhibits gene expression. Transcription and translation incur delays (τ_1, τ_2) and are stochastic in nature (σ_1, ..., σ_4). L is also considered to be stochastic.

Arkin, 1997). Delays and stochasticity of course are not mutually exclusive but occur at the same time (Gedeon and Bokes, 2012). A model of this stochasticity for the case of AhR, using a discrete model based on the Gillespie algorithm, was presented by Simon et al. (2015). However, it did not explicitly account for the time delays between AhR binding and target protein expression and the role of AhRR as repressor.

Taken delays and stochasticity into account, we can formulate a dBST model that allows us to test the effects of delay and stochasticity. In mass-action and power-law format, and with simplified notation (see **Eq. 2b**),such a model has the following format:

$$
\begin{aligned}
\tilde{X}_1 &= X_1 + \vartheta \cdot [\, k_2 \cdot X_2 - k_1 \cdot L \cdot X_1] \\
\tilde{X}_2 &= X_2 + \vartheta \cdot [k_1 \cdot L \cdot X_1 - k_2 \cdot X_2 - k_3 \cdot X_2 \cdot X_3 + k_4 \cdot X_5\,] \\
\tilde{X}_3 &= X_3 + \vartheta \cdot [-k_3 \cdot X_2 \cdot X_3 + k_{41} \cdot X_5 - k_5 \cdot X_3 \cdot X_4 + k_6 \cdot X_6] \\
\tilde{X}_4 &= X_4 + \vartheta \cdot [-k_5 \cdot X_3 \cdot X_4 + k_{61} \cdot X_6 + k_7 \cdot X_8 - k_8 \cdot X_4\,] \\
\tilde{X}_5 &= X_5 + \vartheta \cdot [\, k_3 \cdot X_2 \cdot X_3 - k_4 \cdot X_5] \\
\tilde{X}_6 &= X_6 + \vartheta \cdot [\, k_5 \cdot X_3 \cdot X_4 - k_6 \cdot X_6\,] \\
\tilde{X}_7 &= X_7 + \vartheta \cdot [\, k_9 \cdot X_5 \cdot (X_6 + 1)^g - k_{10} \cdot X_7] \\
\tilde{X}_8 &= X_8 + \vartheta \cdot [\, k_9 \cdot X_5 \cdot (X_6 + 1)^g - k_{11} \cdot X_8] \\
\tilde{X}_9 &= X_9 + \vartheta \cdot [\, k_{12} \cdot X_7 - k_{13} \cdot X_9]
\end{aligned}
\qquad (12)
$$

The variable names are defined in **Figure 10**. For this illustration, we choose reasonable rate constants as shown in **Table 3** and set the inhibition parameter as $g = -4$; to avoid numerical issues for $X_6 = 0$, we define the inhibition as $(X_6 + 1)^g$.

TABLE 3 | Parameter values for the AhR signal transduction system.

k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}	k_{11}	k_{12}	k_{13}	g	τ_1	τ_2
1	2	1	1	1	2	2	2	10	10	10	10	6	–4	3	4

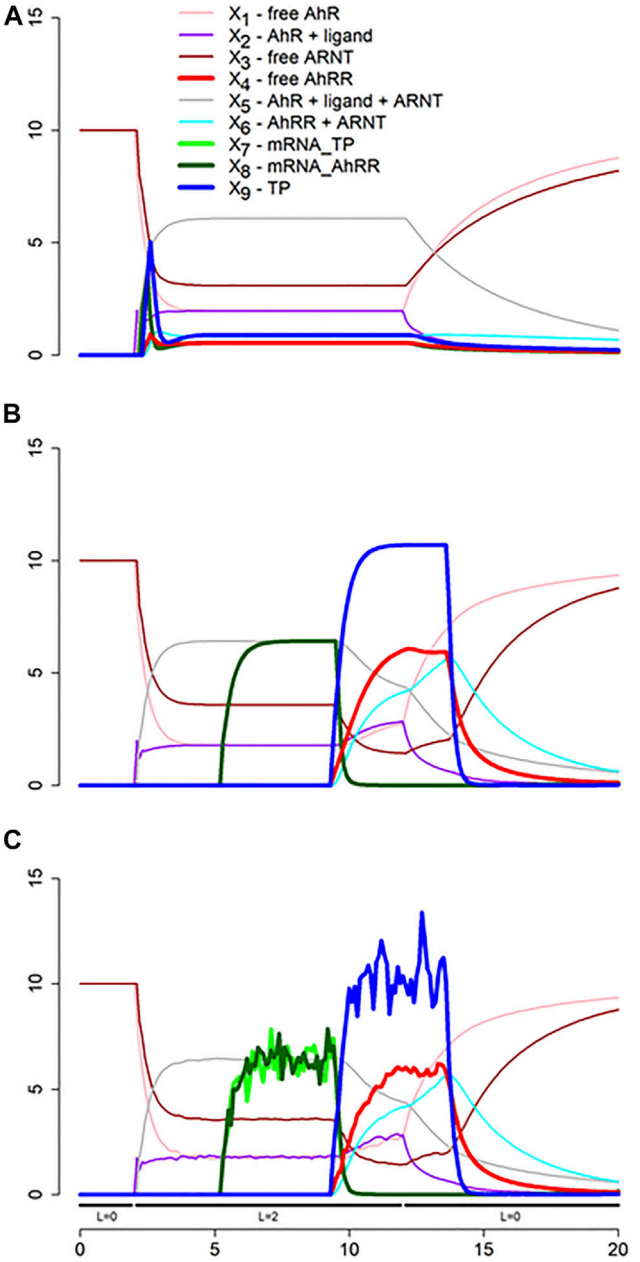


FIGURE 11 | Simulation results of three scenarios. **(A)** Time delays and stochasticity are simply ignored. **(B)** Delays for transcription and for translation and activation of protein are taken into account. Note that the dynamics of X_7 and X_8 is the same. **(C)** Ligand availability, transcription and translation are considered stochastic. See Text for further details.

The simulations start at the steady state in the absence of ligand ($L = 0$), which is $(X_{10}, ..., X_{90}) = (10, 0, 10, 0, 0, 0, 0, 0, 0)$. The shaded terms are subject to delay, stochasticity, or both (see **Figure 10**). Time is roughly in hours. The step size was taken as $\vartheta = 0.1$.

We show the result of three scenarios. In the first, the time delays and any stochasticity are simply ignored (**Figure 11A**). The second scenario accounts for two types of delays (**Figure 11B**), one for transcription and one for translation and activation of protein, and the third incorporates both, delays and noise (**Figure 11C**). For simplicity, we assume the same delay for the transcription ($\tau_1 = 3$ h) and translation and activation ($\tau_2 = 4$ h) of AhRR and a representative target protein (TP), even though these delays are in reality protein-specific (Koussounadis et al., 2015). Also for simplicity, we assume the same stochastic structure for σ_1 and σ_2 and for σ_3 and σ_4 (see **Figure 10**). Specifically, these stochastic events are modeled with values from the normal distribution N (1, 0.1), which are multiplied to the affected fluxes. We also added stochasticity to the ligand availability; it did not have much effect but shows up in the dynamics of X_1, ..., X_4.

Both simulations start at the steady state without ligand (L = 0). At time t = 2, the ligand concentration is set to 2, and at t = 12 it is returned to 0. The result of the first scenario simulation (**Figure 11A**) reveals that the production of target protein (TP) very briefly peaks, but that not much TP is produced, due to the immediate onset of inhibition by AhRR. By constrast, accounting for time delays yields a dramatically different dynamics (**Figure 11B**): Critically, the time delays permit transcription and translation to occur unabatedly until the repression sets in. Specifically, after 3 h, mRNA becomes available, and after an additional 4 h, proteins emerge, including AhRR, which quickly binds to ARNT and begins repressing transcription, resulting soon after in decreased protein production. If the ligand is available beyond time t = 12, the production of protein oscillates, and as soon as the ligand is no longer present, the system returns to the original steady state (not shown). The results for an ODE model without delays and stochasticity are essentially the same as in **Figure 11A**, and a delay differential equation model, ignoring stochasticity, produces more or less the same results as in **Figure 11B**. The combination of delays and stochasticity is difficult to capture with differential equations, but it is easily implemented in dBST **Figure 11C**.

DISCUSSION

Modeling approaches utilizing the framework of Biochemical Systems Theory (BST) have proven powerful in biomedical systems analysis

for over seven decades (Savageau, 1969a; Savageau, 1969b; Savageau, 1970). Our goal in the present article was a demonstration that discrete BST ("dBST") models are noteworthy alternatives to ODE-based BST systems and that they can shed light on complex biomedical phenomena in a similar manner. Discrete dBST models are arguably also more intuitive to newcomers coming from the field of biology, for whom differential equations are often an obscure and dreaded domain of insider mathematics.

The proposal of using dBST is most certainly not a call for abandoning systems of ODEs in biomedical models. ODEs have proven immensely beneficial in all of science, and biomedical applications are no exception. Nonetheless, there are situations that are difficult to align with the concept of instantaneous change. Examples include genuinely discrete events, delays, and stochastic phenomena affecting the phenomenon under study. For instance, we showed elsewhere, in the context of red blood cell death during malarial anemia, that the precise dynamics of blood infections is very difficult to capture with ODEs, but straightforward to implement in a discrete-recursive model (Fonseca and Voit, 2015; Fonseca et al., 2016). Similarly, we demonstrate here and in the **Supplementary Data S2** that delays and internal or external stochastic influences affecting a dynamical system are often more easily incorporated into discrete rather than differential equations.

Many of the advantages of BST as a tool for model selection and analysis translate directly into its discrete analog, dBST. Whereas it is generally difficult to choose the most appropriate mathematical formats for representing ill-characterized phenomena *a priori*, BST and dBST offer guidance at the very beginning of the modeling process, where it is most urgently needed. At the very least, the use of power-law functions offers a viable, unbiased starting point. The power-law format used in BST and dBST is no panacea, but it is a local approximation of mathematically guaranteed quality that typically has a wider range of validity than linear formulations and, embedded into ODEs, is provenly rich enough to permit the inclusion of any differentiable nonlinearities (Voit and Savageau, 1986; Savageau and Voit, 1987).

The use of dBST instead of BST does not create practical design or implementation problems *per se*, and paradigmatic nonlinearities, such as limit cycles and chaos, can be captured in dBST, as we demonstrated here. If the goal of a dBST model is to mimic a corresponding ODE system as closely as possible, a small step size may have to be chosen. For instance, in the example of limit cycles, a larger step size retained the basic structure and shape of the limit cycle system, but the numerical features were clearly affected. However, the typical task in practical applications is not to create an analog of an ODE system but to convert observed data, together with contextual information, into a computable structure. This inference process is actually simpler in dBST than BST, as most biomedical phenomena are naturally discrete and the determination of optimal parameter values does not require the estimation of slopes.

We demonstrated the ease of designing a dBST model with several small examples and with a moderately complex signal transduction system that triggers changes in gene expression following an exposure to specific toxicants or internal ligands. This phenomenon is difficult to capture with an ODE model because it is critically affected by substantial time delays, which are comingled with the well-known stochastic nature of gene transcription and translation. Our analysis makes it evident that these aspects must not be ignored lest erroneous results are obtained. It also shows how straightforward it is to incorporate these aspects into a dBST model.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/LBSA-VoitLab/dBST.

AUTHOR CONTRIBUTIONS

Conceived the study: EV; executed all computational analyses: EV and DO; wrote and edited the article: EV and DO.

FUNDING

This work was supported by the NIEHS: NIH-2P30ES019776-05 (PI: Carmen Marsit); Salary support for EV and DO; Georgia Research Alliance (PI: EV) (general research support; no grant number). The funding agencies are not responsible for the content of this article.

ACKNOWLEDGMENTS

The authors greatly appreciate constructive feedback from Jacob Davis and Carla Kumbale.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.874669/full#supplementary-material

REFERENCES

Brinkmann, V., Ale-Agha, N., Haendeler, J., and Ventura, N. (2019). The Aryl Hydrocarbon Receptor (AhR) in the Aging Process: Another Puzzling Role for This Highly Conserved Transcription Factor. *Front. Physiol.* 10, 1561. doi:10.3389/fphys.2019.01561

Carlton, J. G., Jones, H., and Eggert, U. S. (2020). Membrane and Organelle Dynamics during Cell Division. *Nat. Rev. Mol. Cel Biol.* 21, 151–166. doi:10.1038/s41580-019-0208-1

Chou, I.-C., and Voit, E. O. (2009). Recent Developments in Parameter Estimation and Structure Identification of Biochemical and Genomic Systems. *Math. biosciences* 219, 57–83. doi:10.1016/j.mbs.2009.03.002

Claude, D. (1995). Shift of a Limit Cycle in Biology: From Pathological to Physiological Homeostasia*. *Chaos* 5, 162–166. doi:10.1063/1.166099

Cornish-Bowden, A. (2012). *Fundamentals of Enzyme Kinetics*. Chichester, UK: J. Wiley & Sons, 510.

Dam, P., Rodriguez-R, L. M., Luo, C., Hatt, J., Tsementzi, D., Konstantinidis, K. T., et al. (2020). Model-based Comparisons of the Abundance Dynamics of Bacterial Communities in Two Lakes. *Sci. Rep.* 10, 2423. doi:10.1038/s41598-020-58769-y

Evans, B. R., Karchner, S. I., Allan, L. L., Pollenz, R. S., Tanguay, R. L., Jenny, M. J., et al. (2008). Repression of Aryl Hydrocarbon Receptor (AHR) Signaling by AHR Repressor: Role of DNA Binding and Competition for AHR Nuclear Translocator. *Mol. Pharmacol.* 73, 387–398. doi:10.1124/mol.107.040204

Fonseca, L. L., Alezi, H. S., Barnwell, J. W., Galinski, M. R., and Voit, E. O. (2016). Quantifying the Removal of Red Blood Cells in *Macaca mulatta* during a Plasmodium Coatneyi Infection. *Malar. J.* 15, 410. doi:10.1186/s12936-016-1465-5

Fonseca, L. L., and Voit, E. O. (2015). Comparison of Mathematical Frameworks for Modeling Erythropoiesis in the Context of Malaria Infection. *Math. biosciences* 270, 224–236. doi:10.1016/j.mbs.2015.08.020

Fournier, M. L., Paulson, A., Pavelka, N., Mosley, A. L., Gaudenz, K., Bradford, W. D., et al. (2010). Delayed Correlation of mRNA and Protein Expression in Rapamycin-Treated Cells and a Role for Ggc1 in Cellular Sensitivity to Rapamycin. *Mol. Cell Proteomics* 9, 271–284. doi:10.1074/mcp.m900415-mcp200

Franco, R. S. (2012). Measurement of Red Cell Lifespan and Aging. *Transfus. Med. Hemother* 39, 302–307. doi:10.1159/000342232

Gábor, A., and Banga, J. R. (2015). Robust and Efficient Parameter Estimation in Dynamic Models of Biological Systems. *BMC Syst. Biol.* 9, 74. doi:10.1186/s12918-015-0219-2

Gedeon, T., and Bokes, P. (2012). Delayed Protein Synthesis Reduces the Correlation between mRNA and Protein Fluctuations. *Biophysical J.* 103, 377–385. doi:10.1016/j.bpj.2012.06.025

Gennemark, P., and Wedelin, D. (2009). Benchmarks for Identification of Ordinary Differential Equations from Time Series Data. *Bioinformatics* 25, 780–786. doi:10.1093/bioinformatics/btp050

Gennemark, P., and Wedelin, D. (2007). Efficient Algorithms for Ordinary Differential Equation Model Identification of Biological Systems. *IET Syst. Biol.* 1, 120–129. doi:10.1049/iet-syb:20050098

Gupta, A., Galinski, M. R., and Voit, E. O. (2022). Dynamic Control Balancing Cell Proliferation and Inflammation Is Crucial for an Effective Immune Response to Malaria. *Front. Mol. Biosci.* 8, 800721. doi:10.3389/fmolb.2021.800721

Huang, G., and Elferink, C. J. (2012). A Novel Nonconsensus Xenobiotic Response Element Capable of Mediating Aryl Hydrocarbon Receptor-dependent Gene Expression. *Mol. Pharmacol.* 81, 338–347. doi:10.1124/mol.111.075952

Julliard, W., Fechner, J. H., and Mezrich, J. D. (2014). The Aryl Hydrocarbon Receptor Meets Immunology: Friend or Foe? A Little of Both. *Front. Immunol.* 5, 458. doi:10.3389/fimmu.2014.00458

Keshet, L. (2005). *Mathematical Models in Biology*. Philadelphia: Society of Industrial and Applied Mathematics.

Klipp, E., Liebermeister, W., Wierling, C., Kowald, A., Lehrach, H., and Herwig, R. (2016). *Systems Biology: A Textbook*. Weinheim, Germany: Wiley VCH Verlag.

Knowles, I., and Renka, R. J. (2014). Methods for Numerical Differentiation of Noisy Data. *Electr. J. Diff. Equations* 21, 35–246.

Koussounadis, A., Langdon, S. P., Um, I. H., Harrison, D. J., and Smith, V. A. (2015). Relationship between Differentially Expressed mRNA and mRNA-protein Correlations in a Xenograft Model System. *Sci. Rep.* 5, 10775. doi:10.1038/srep10775

Lewis, D. C. (1991). "A Qualitative Analysis of S-Systems: Hopf Bifurcations," in *Canonical Nonlinear Modeling: S-System Approach to Understanding Complexity*. Editor E. O. Voit (New York, NY: Van Nostrand Reinhold), 304–344.

Lewis, M. (2005). The Lac Repressor. *Comptes Rendus Biologies* 328, 521–548. doi:10.1016/j.crvi.2005.04.004

Liu, Y., Beyer, A., and Aebersold, R. (2016). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* 165, 535–550. doi:10.1016/j.cell.2016.03.014

Lorenz, E. N. (1963). Deterministic Nonperiodic Flow. *J. Atmos. Sci.* 20, 130–141. doi:10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2

Lotka, A. J. (1956). *Elements of Mathematical Biology*. New York: Dover Publications.

May, R. M. (1973). *Stability and Complexity in Model Ecosystems*. New Jersey, United States: Princeton University Press.

McAdams, H. H., and Arkin, A. (1997). Stochastic Mechanisms in Gene Expression. *Proc. Natl. Acad. Sci. U.S.A.* 94, 814–819. doi:10.1073/pnas.94.3.814

Michaelis, L., and Menten, M. L. (1913). Die Kinetik der Invertinwirkung. *Biochemische Z.* 49, 333–369.

Nelder, J. A., and Mead, R. (1965). A Simplex Method for Function Minimization. *Comput. J.* 7, 308–313. doi:10.1093/comjnl/7.4.308

R Core Team, R. (2018). *A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raj, A., and van Oudenaarden, A. (2008). Nature, Nurture, or Chance: Stochastic Gene Expression and its Consequences. *Cell* 135, 216–226. doi:10.1016/j.cell.2008.09.050

Saltzman, B. (1962). Finite Amplitude Free Convection as an Initial Value Problem-I. *J. Atmos. Sci.* 19, 329–341. doi:10.1175/1520-0469(1962)019<0329:fafcaa>2.0.co;2

Savageau, M. A. (1970). Biochemical Systems Analysis. 3. Dynamic Solutions Using a Power-Law Approximation. *J. Theor. Biol.* 26, 215–226. doi:10.1016/s0022-5193(70)80013-3

Savageau, M. A. (1976). *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*. Reading, MA: Addison-Wesley Pub. Co. Advanced Book Program. (reprinted 2009).

Savageau, M. A. (1969). Biochemical Systems Analysis. II. The Steady-State Solutions for an N-Pool System Using a Power-Law Approximation. *J. Theor. Biol.* 25, 370–379. doi:10.1016/s0022-5193(69)80027-5

Savageau, M. A. (1969). Biochemical Systems Analysis. I. Some Mathematical Properties of the Rate Law for the Component Enzymatic Reactions. *J. Theor. Biol.* 25, 365–369. doi:10.1016/s0022-5193(69)80026-3

Savageau, M. A. (2001). Design Principles for Elementary Gene Circuits: Elements, Methods, and Examples. *Chaos* 11, 142–159. doi:10.1063/1.1349892

Savageau, M. A., and Voit, E. O. (1987). Recasting Nonlinear Differential Equations as S-Systems: A Canonical Nonlinear Form. *Math. Biosciences* 87, 83–115. doi:10.1016/0025-5564(87)90035-6

Schafer, K. A. (1998). The Cell Cycle: a Review. *Vet. Pathol.* 35, 461–478. doi:10.1177/030098589803500601

Simon, T. W., Budinsky, R. A., and Rowlands, J. C. (2015). A Model for Aryl Hydrocarbon Receptor-Activated Gene Expression Shows Potency and Efficacy Changes and Predicts Squelching Due to Competition for Transcription Co-activators. *PLoS One* 10, e0127952. doi:10.1371/journal.pone.0127952

Soetaert, K., Petzoldt, T., and Setzer, R. W. (2010). Solving Differential Equations in R: Package deSolve. *J. Stat. Softw.* 33, 1–25. doi:10.18637/jss.v033.i09

Spence, K. F., Jr., Decker, S., and Sell, K. W. (1970). Bursting Atlantal Fracture Associated with Rupture of the Transverse Ligament. *J. Bone Jt. Surg.* 52, 543–549. doi:10.2106/00004623-197052030-00013

Stein, R. R., Bucci, V., Toussaint, N. C., Buffie, C. G., Rätsch, G., Pamer, E. G., et al. (2013). Ecological Modeling from Time-Series Inference: Insight into Dynamics and Stability of Intestinal Microbiota. *Plos Comput. Biol.* 9, e1003388. doi:10.1371/journal.pcbi.1003388

Stevens, E. A., Mezrich, J. D., and Bradfield, C. A. (2009). The Aryl Hydrocarbon Receptor: a Perspective on Potential Roles in the Immune System. *Immunology* 127, 299–311. doi:10.1111/j.1365-2567.2009.03054.x

Stockinger, B., Meglio, P. D., Gialitakis, M., and Duarte, J. H. (2014). The Aryl Hydrocarbon Receptor: Multitasking in the Immune System. *Annu. Rev. Immunol.* 32, 403–432. doi:10.1146/annurev-immunol-032713-120245

Torres, N. V., and Voit, E. O. (2002). *Pathway Analysis and Optimization in Metabolic Engineering*. Cambridge, U.K.: Cambridge University Press.

Varah, J. M. (1982). A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations. *SIAM J. Sci. Stat. Comput.* 3, 28–46. doi:10.1137/0903003

Voit, E. O. (2017). *A First Course in Systems Biology*. 2nd Ed. New York, NY: Garland Science.

Voit, E. O. (2000). *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*. Cambridge, UK: Cambridge University Press.

Voit, E. O., Martens, H. A., and Omholt, S. W. (2015). 150 Years of the Mass Action Law. *Plos Comput. Biol.* 11, e1004012. doi:10.1371/journal.pcbi.1004012

Voit, E. O., and Savageau, M. A. (1982). Power-law Approach to Modeling Biological Systems; III. Methods of Analysis. *J. Ferment. Technol.* 60, 223–241.

Voit, E. O. (2020). *Systems Biology*. Oxford, U.K.: Oxford University Press.

Voit, E. O. (2017). The Best Models of Metabolism. *Wiley Interdiscip. Rev. Syst. Biol. Med.* 9, e1391. doi:10.1002/wsbm.1391

Voit, E. O., and Almeida, J. (2004). Decoupling Dynamical Systems for Pathway Identification from Metabolic Profiles. *Bioinformatics* 20, 1670–1681. doi:10.1093/bioinformatics/bth140

Voit, E. O. (2013). Biochemical Systems Theory: A Review. *ISRN Biomathematics* 2013, 1–53. doi:10.1155/2013/897658

Voit, E. O. (2008). Modelling Metabolic Networks Using Power-Laws and S-Systems. *Essays Biochem.* 45, 29–40. doi:10.1042/bse0450029

Voit, E. O., and Savageau, M. A. (1986). Equivalence between S-Systems and Volterra Systems. *Math. Biosciences* 78, 47–55. doi:10.1016/0025-5564(86)90030-1

Volterra, V. (1926). Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem. R. Accad. Dei Lincei.* 2, 31–113.

Yin, W., and Voit, E. O. (2008). Construction and Customization of Stable Oscillation Models in Biology. *J. Biol. Syst.* 16, 463–478. doi:10.1142/s0218339008002502

Zhu, K., Meng, Q., Zhang, Z., Yi, T., He, Y., Zheng, J., et al. (2019). Aryl Hydrocarbon Receptor Pathway: Role, Regulation and Intervention in Atherosclerosis Therapy (Review). *Mol. Med. Rep.* 20, 4763–4773. doi:10.3892/mmr.2019.10748

Zudaire, E., Cuesta, N., Murty, V., Woodson, K., Adams, L., Gonzalez, N., et al. (2008). The Aryl Hydrocarbon Receptor Repressor Is a Putative Tumor Suppressor Gene in Multiple Human Cancers. *J. Clin. Invest.* 118, 640–650. doi:10.1172/JCI30024

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Voit and Olivença. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 16 May 2022
doi: 10.3389/fmolb.2022.863470

Predicting Metabolic Adaptation Under Dynamic Substrate Conditions Using a Resource-Dependent Kinetic Model: A Case Study Using *Saccharomyces cerevisiae*

K. J. A. Verhagen[1], *S. A. Eerden*[1], *B. J. Sikkema*[1] *and S. A. Wahl*[1,2]*

[1]*Department of Biotechnology, Delft University of Technology, Delft, Netherlands,* [2]*Lehrstuhl für Bioverfahrenstechnik, FAU Erlangen-Nürnberg, Erlangen, Germany*

Exposed to changes in their environment, microorganisms will adapt their phenotype, including metabolism, to ensure survival. To understand the adaptation principles, resource allocation-based approaches were successfully applied to predict an optimal proteome allocation under (quasi) steady-state conditions. Nevertheless, for a general, dynamic environment, enzyme kinetics will have to be taken into account which was not included in the linear resource allocation models. To this end, a resource-dependent kinetic model was developed and applied to the model organism *Saccharomyces cerevisiae* by combining published kinetic models and calibrating the model parameters to published proteomics and fluxomics datasets. Using this approach, we were able to predict specific proteomes at different dilution rates under chemostat conditions. Interestingly, the approach suggests that the occurrence of aerobic fermentation (Crabtree effect) in *S. cerevisiae* is not caused by space limitation in the total proteome but rather an effect of constraints on the mitochondria. When exposing the approach to repetitive, dynamic substrate conditions, the proteome space was allocated differently. Less space was predicted to be available for non-essential enzymes (reserve space). This could indicate that the perceived "overcapacity" present in experimentally measured proteomes may very likely serve a purpose in increasing the robustness of a cell to dynamic conditions, especially an increase of proteome space for the growth reaction as well as of the trehalose cycle that was shown to be essential in providing robustness upon stronger substrate perturbations. The model predictions of proteome adaptation to dynamic conditions were additionally evaluated against respective experimentally measured proteomes, which highlighted the model's ability to accurately predict major proteome adaptation trends. This proof of principle for the approach can be extended to production organisms and applied for both understanding metabolic adaptation and improving industrial process design.

Keywords: proteome adaptation, kinetic modeling, *Saccharomyces cerevisiae*, Crabtree effect, resource allocation, dynamic conditions, feast/famine

OPEN ACCESS

Edited by:
Alberto Jesus Martin, Universidad Mayor, Chile

Reviewed by:
Yoshihiro Toya, Osaka University, Japan
Jose Utrilla, National Autonomous University of Mexico, Mexico

***Correspondence:**
S. A. Wahl
aljoscha.wahl@fau.de

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *27 January 2022*
Accepted: *29 March 2022*
Published: *16 May 2022*

Citation:
Verhagen KJ, Eerden SA, Sikkema BJ and Wahl SA (2022) Predicting Metabolic Adaptation Under Dynamic Substrate Conditions Using a Resource-Dependent Kinetic Model: A Case Study Using Saccharomyces cerevisiae. Front. Mol. Biosci. 9:863470. doi: 10.3389/fmolb.2022.863470

INTRODUCTION

The ability of microorganisms to adapt to changing extracellular environmental conditions is essential for their survival and leads to metabolic robustness and competitive fitness (Gerosa and Sauer, 2011; Chubukov et al., 2014). Depending on the environmental conditions, different metabolic functions and/or flux distributions are needed that require a different proteome composition (Litsios et al., 2018). The proteome adaption is triggered by not yet fully unraveled protein signaling cascades and further mechanisms (Zhao et al., 2016). An intuitive example of this adaption is described for *Saccharomyces cerevisiae* (*S. cerevisiae*) when shifting from growth under minimal to rich medium conditions; cells grown under rich nutrient conditions require a significantly smaller proteome fraction for biosynthesis genes (de Godoy et al., 2008; Nagaraj et al., 2012; Liebermeister et al., 2014) than cells grown in the mineral medium, in which amino acids and other biomass precursors are not present but have to be synthesized from glucose.

On the other hand, next to optimization of proteome resources, cells do maintain metabolic fitness and/or robustness (Basan, 2018), especially under substrate limiting conditions cells seem to invest in proteins that may not be required yet, for example, to quickly utilize alternative substrates without delays in growth (Dekel and Alon, 2005). However, any additional increase in protein abundance also results in higher costs due to occupation of ribosomes, resource consumption, and potentially additional protein misfolding. Different hypotheses have been formulated and respective models were developed to understand the optimization and trade-offs.

Constraint-based modeling approaches are essential to analyze putative properties of metabolic networks. The well-established and frequently used method for the analysis of (large genome-scale) metabolic networks is flux balance analysis (FBA) (Varma and Palsson, 1994; Orth et al., 2010). This method calculates feasible solutions under steady-state conditions, depending on a defined objective function (biomass or ATP maximization) (Schuetz et al., 2007). However, this method cannot be applied to dynamic cultivation conditions and does not consider gene regulation or protein expression. To overcome these limitations, dynamic flux balance analysis (dFBA) was developed to maximize biomass growth over time, with changing extracellular conditions (Mahadevan et al., 2002). To include the synthesis costs of proteins and ribosomes, resource balance analysis (RBA) was developed, allowing for the prediction of the optimal allocation of intracellular resources for steady-state growth (Goelzer et al., 2015). Looking at cellular behavior in terms of resource allocation has also been used to explain overflow metabolism (Basan et al., 2015; Nilsson and Nielsen, 2016). In this paradoxical phenomenon, cells use catabolic pathways with low ATP yields per substrate such as alcoholic fermentation when growing at high growth rates, even when a high-yield pathway such as respiration is available. Following the current hypothesis, the answer is that these fermentative pathways are much cheaper in terms of proteome space cost, meaning that the ATP production rate per protein mass is larger (Nilsson and Nielsen, 2016).

Combining approaches from both dFBA and RBA leads to conditional FBA (cFBA) (Rügen et al., 2015; Reimers et al., 2017), which combined both temporal changes in the extracellular environment with constraints on intracellular resource allocation. These powerful tools are able to reproduce and predict metabolic phenotypes beyond steady-state conditions and extend our understanding of microbial physiology. Nevertheless, short-term dynamics require yet another mechanism: kinetics instead of a quasi-steady state of the intracellular metabolites to capture the rapid intracellular changes of metabolites as well as kinetic regulation.

Experimentally, *S. cerevisiae* cultures have a different metabolic response to substrate perturbations depending on the cultivation condition, especially cells cultured under repetitive dynamic substrate conditions, the so-called "feast/famine" regime showed a different response compared to cultures grown under steady-state limitations (Suarez-Mendez et al., 2014). Ethanol production after a substrate pulse was observed for cultures originating from a chemostat (Wu et al., 2006), while no ethanol was observed for cells under a repetitive excess/limitation regime (Suarez-Mendez, 2015). Furthermore, the intracellular response to substrate excess has significantly different properties: while the ATP concentration dropped after a pulse originating from a chemostat culture (Wu et al., 2006), a rise was observed for a feast/famine culture. Moreover, the biomass yield of a feast/famine culture was lower than that of a chemostat culture. Last, chemostat-grown cells showed short- and long-term accumulation of glycolytic intermediates after a substrate pulse, while this was not observed for feast/famine cultures. Storage synthesis and degradation leads to "wasting" of ATP (futile cycle) which was shown to rescue cellular metabolism, that is, balance pathway capacities in case of sudden perturbations (van Heerden et al., 2014).

The observed differential metabolic response implies an adaptation during the prior dynamic growth condition. Similar differences in adaptations have been observed earlier, for example, the lag phase before exponential growth (Brejning and Jespersen, 2002; Jõers and Tenson, 2016), upon a change in the substrate (Chu and Barnes, 2016), and in the period just after switching to a different dilution rate in a chemostat (Abulesz and Lyberatos, 1989).

There are three levels of metabolic regulation commonly assumed to be dominant (Wegner et al., 2015): 1) allosteric regulation, in which enzyme activity is modified by non-covalent binding with other molecules. The response time of this type of regulation is almost instant (Pincus et al., 2017), and it is often used for local fine-tuning in metabolism, and thus it is unlikely to cause this adaptation effect. 2) Post-translational modifications (PTMs), in which enzyme activity is altered by the addition of covalent attachments. The timescale of this response is a matter of seconds to minutes (Karim et al., 2014), and it is often part of short-term responses to stress situations (e.g., sudden changes in the environment). 3) Translational regulation, which influences the composition of the proteome. This regulation has a response time of hours

(Cohen et al., 2008), which is in the same order of magnitude as the generation time, and thus the choices made at this level are important for the long-term strategy. It is also considered the most expensive regulatory level: degradation and synthesis of proteins requires significant amounts of ATP.

Recent studies have shown that the amount of protein in a cell is limited due to macromolecular crowding, the kinetics of protein synthesis, and degradation (Vazquez et al., 2008; Molenaar et al., 2009). When all the proteome space is occupied, increasing the concentration of one protein is only possible at the cost of another (Pareto Frontier) (Mori et al., 2019).

We were curious to study the impact of short-term vs. long-term adaptations to substrate perturbations encountered in natural and laboratory environments. Therefore, we developed a resource-dependent kinetic model and exposed this to different dynamic environments to evaluate the impact of the allocation of proteins in the cellular proteome on the metabolic fitness of a yeast cell under short-term extracellular substrate dynamics.

MATERIALS AND METHODS

Proteome-Dependent Kinetic Yeast Model

The proteome-dependent kinetic yeast model is based on a system of ordinary differential equations (ODEs) that describe the mass balances of all intra- and extra-cellular metabolites. This system of ODEs is solved with the ode15s function in MATLAB 2020b, for which the absolute tolerance is set to 1e-4, and all variables are constrained to be higher than zero with the "nonnegative" setting. A detailed description of the final proteome-dependent kinetic yeast model used is given in **Supplementary Material S1**.

To predict which proteome composition is the most competitive for defined environmental conditions, a Monte Carlo approach is used. The metabolic behavior of 1,000 random proteomes, generated around a seed proteome, is compared based on an objective function. Under steady-state conditions, the minimization of the residual substrate concentration was used as an objective function. Under dynamic conditions, the minimization of a time-weighted average substrate concentration was used, to promote fast consumption of the available substrate, therefore selecting for competitive proteomes:

$$\frac{\int_0^{t_{cycle}} c_s \cdot t \, dt}{\int_0^{t_{cycle}} t \, dt}.$$

Subsequently, it is determined whether the solution is balanced. If the objective function is optimized and the solution is balanced, the objective function and the seed proteome are updated. In the next iteration, the proteome is then generated around this new seed proteome, with a maximum deviation of 25% per sector.

Proteome Allocation to Model Sectors

All proteins from experimental datasets are sorted in the same nine protein sectors that are used in the model, to allow for a direct comparison of the experimental proteomes and the optimized proteomes. The proteins are categorized per sector based on either the protein name or the description in the KEGG database (Goffeau et al., 1996; Kanehisa et al., 2016) (**Supplementary Material S7**). The whole dataset is sorted with the MATLAB 2020b functions "strcmp" and "contains," which are used to search the dataset for specific names or keywords to group the proteins.

Parameter Optimization

The proteome cost parameters are estimated by optimization with the MATLAB 2020b function fmincon. For all parameter optimizations, a multi-start approach is used. This approach minimizes the risk of reaching a local minimum in the solution space by starting the optimization from different initial guesses. The tolerance of the function is set to 1e-12 for all optimizations. For the estimation of the k_{cat} parameters, the difference between the experimental and simulated fluxes is minimized. Additional weight in the objective function was applied for the growth rate, as k_{cat} parameters have to be rejected if the maximum growth rate is not reached.

Overcapacity Simulations

The amount of overcapacity in the yeast proteome is determined by introducing a 10th protein sector. This new protein sector does not have a function for the cells, and hence, only takes up space in the proteome. Therefore, the fraction of the proteome that can be allocated into the extra sector without altering the metabolic fluxes is defined as overcapacity. The overcapacity is estimated for each sector separately, to minimize the changes in each step. The sectors are sorted in a decreasing order and then optimized for overcapacity in this order. The amount of overcapacity in each sector is determined in a step-wise approach. Per iteration, one percent of the specific protein sector is removed and allocated into the extra sector. Subsequently, the fluxes of the adapted proteome are compared to the reference fluxes, and only if the change in the fluxes remains within the boundaries, the seed proteome is updated. This new seed proteome is then used for the next iteration, in which the sector size is again decreased by 1%. By decreasing the sector size by 1% of the current size, the step size is reduced with each iteration. If the flux profile deviates more than the threshold value, the adapted proteome allocation is rejected. The fluxes are evaluated based on the following criterion: the average value of the uptake and growth fluxes should not deviate more than 1% from the reference flux, to ensure that the same substrate uptake and growth rates are achieved.

RESULTS

Construction of a Proteome-Dependent Kinetic Model

We wanted to construct a proteome-dependent kinetic model, which was small, but still able to reproduce the main phenotypes observed for *S. cerevisiae*. Furthermore, it should be calibrated with available experimental data. We constructed the model based on the kinetic model of yeast glycolysis (Teusink et al.,

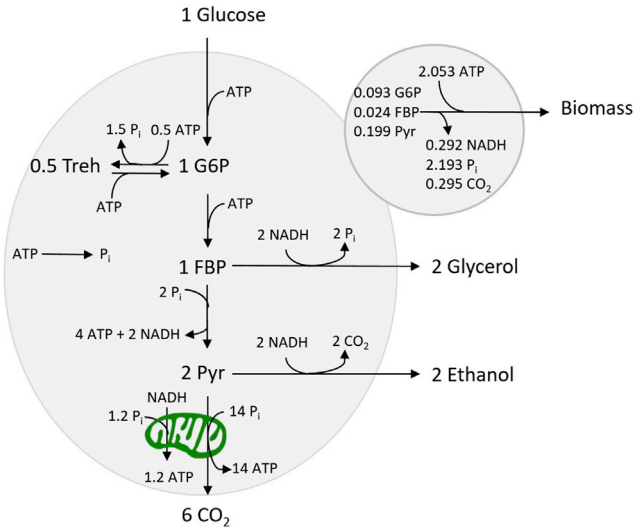


FIGURE 1 | Map showing the metabolic network used in this model.

2000) which we extended with reactions for the trehalose cycle and respiration pathway as well as a growth reaction (see **Figure 1**). Each (lumped) reaction has been associated with a proteome fraction resulting in a proteome-dependent kinetic model of yeast central carbon metabolism and growth.

The Embden–Meyerhof glycolytic pathway has been implemented as three lumped reactions (uptake, upper, and lower glycolysis) with three intermediates: G6P, FBP, and pyruvate. The stoichiometry of the growth reaction was based on Suarez-Mendez et al. (2016). The NADPH requirement was assumed to be met by using the pentose phosphate pathway, which in sum (together with PGI) converts one G6P to six CO_2 and 12 NADPH. The required NADPH flux was balanced by a respective consumption of G6P. The ATP demand for growth has been derived from Della-Bianca et al. (2014) taking into account that the demand was expressed as catabolized glucose amounts. Furthermore, the trehalose cycle was included as two lumped reactions, based on an existing kinetic model of the trehalose cycle (Smallbone et al., 2011) (see **Supplementary Material S1** for details).

Due to a lack of kinetic models of yeast TCA cycle and oxidative phosphorylation, the two respiratory reactions (from cytosolic NADH and pyruvate, v_{NDE}, and v_{TCA}, respectively) have been implemented using general Michaelis–Menten kinetics. However, the two reactions are interdependent—both connect to the electron transport chain—and consequently, the rate is determined by the same proteome fraction. A maximum value for the rate of the two reactions combined is defined, reflecting the capacity in the electron transport chain, limited by the provided proteome sector size (see **Supplementary Material S1**).

The biomass reaction contains many complex reactions, and the kinetics of the full process currently cannot be derived from basic principles. Therefore, a holistic approach based on experimental observations was chosen, i.e., the growth rate has been found to correlate with the energy charge (Boer et al., 2010). Here, the growth rate is described by a sigmoid function that is the most sensitive within the range of an energy charge between 0.7 and 0.9 as observed for growing cells (Boer et al., 2010).

Calibration of Model Parameters Using Available Experimental Data

The specific activity for the defined pathways has a major impact on model predictions. To obtain realistic values, the specific enzyme activities (k_{cat}) were estimated from experimental omics datasets. In the proposed model, the $k_{cat,i}$ for each reaction i is defined as the maximum reaction rate per fraction of proteome (mol/$Cmol_X$/h), where 100% proteome reflects 500 mg protein per g_X (Ertugay and Hamamci, 1997). Hence, the maximum rate of the reaction i ($V_{max,i}$) with a given sector fraction φ_i is:

$$V_{max,i} = \varphi_i \cdot k_{cat,i}.$$

From this, $\vec{c}$ enzymatic rate V_i is calculated by multiplying the $V_{max,i}$ with the function $f_i(\vec{c})$ describing the effects on the enzymatic rate due to substrate and product concentrations as well as effects by allosteric activators and inhibitors (see **Supplementary Material S1** for specification of $f_i(\vec{c})$ for each reaction):

$$V_i = V_{max,i} \cdot f_i\left(\vec{c}\right).$$

The k_{cat} parameters have been estimated by combining the proteome and fluxome measurements under batch conditions. The proteome fractions were taken from de Godoy et al. (2008) using *S. cerevisiae* grown under batch conditions with a defined glucose minimal medium and aligned according to the protein classification in the KEGG database. Specifically, grouping all proteins with the KEGG BRITE label "Genetic Information Processing" and all proteins with the "Metabolism" label that were not classified as "Central Carbon Metabolism" or "Energy Metabolism" being assigned to the "growth protein sector," assuming that their size is growth rate-dependent in the minimal medium. Furthermore, for the calculations, it was assumed that the whole proteome sector of cells grown under excess substrate at the maximal growth rate was used.

The corresponding flux distribution, i.e., under batch conditions was obtained from Heyland et al. (2009) with the exception of fluxes for the trehalose cycle—these were obtained from the feast/famine experiments conducted by Suarez-Mendez et al. (2017). For both trehalose synthesis and degradation, the maximum value of the flux reached in one feast/famine cycle was used, which was $5.10 \cdot 10^{-3}$ mol/$Cmol_X$/h for trehalose synthesis and $4.09 \cdot 10^{-3}$ mol/$Cmol_X$/h for the degradation of trehalose. The k_{cat} value for maintenance was set to 0.0155 mol/$Cmol_X$/h, which is the maintenance requirement measured at near-zero growth rates (Vos et al., 2016).

To obtain the k_{cat} parameters, parameter optimization was performed, estimating the parameters which produced the smallest deviation between the simulated and experimental fluxes (Heyland et al., 2009), using the batch proteome composition taken from de Godoy et al. (2008) (see

TABLE 1 | Comparison of the predicted fluxes of a chemostat experiment at a dilution rate of 0.4 h^{-1} with the experimental flux distribution of Heyland et al. (2009). Upt, uptake; UGlc, upper glycolysis; LGlc, lower glycolysis; Ferm, fermentation; Esnk, electron sink/glycerol pathway; Resp, respiration; Trsn, trehalose synthesis; Trdg, trehalose degradation; Grwt, growth.

	Upt	Uglc	LGlc	Ferm	Esnk	Resp	Trsn	Trdg	Grwt
Predicted flux (mol/$Cmol_x$/h)	0.504	0.4669	0.8251	0.7391	0.0898	0.0719	0.0072	0.0072	0.4
Experimental flux (mol/$Cmol_x$/h)	0.4753	0.4373	0.8745	0.7272	0.0428	0.0808	0.0051	0.0041	0.4
Deviation	+6%	+7%	-6%	+2%	+110%	-11%	+41%	+76%	0%

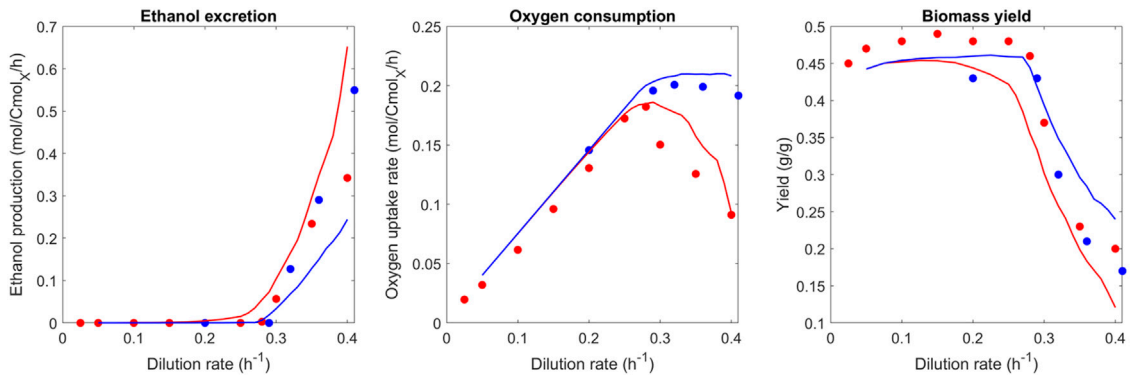


FIGURE 2 | Comparison of predicted and observed phenotypic rates (ethanol excretion, oxygen uptake rate, and biomass yield) at different dilution rates. Blue represents the best proteome out of 1,000 randomly generated proteomes; red represents the best proteome out of 100 randomly generated proteomes (limited evolution with adaptation from the batch proteome). For the experimental data similarly–red represents data from Van Hoek et al. (1998) (seven generations at steady-state starting from batch), and blue represents a respiration-adapted culture (Rieger et al., 1983).

Supplementary Material S2 for details). Using this approach, the proteome-dependent kinetic model was able to largely reproduce the experimental flux distribution (**Table 1**), and this k_{cat} calibration was used in all further calculations.

Prediction of the Steady-State Growth Phenotype Under Carbon-Limited Steady-State Conditions

S. cerevisiae is a Crabtree-positive yeast, and thus fermentation is observed next to oxidative phosphorylation at substrate uptake rates above an observed "critical" rate (Barford and Hall, 1979). The ability of the model to reproduce the Crabtree effect is assessed by optimizing proteomes for dilution rates in the range from 0.05 h^{-1} to 0.4 h^{-1}. The proteome optimization was started at the dilution rate of 0.4 h^{-1} using the experimental batch proteome as a starting value. The most competitive proteome out of 1,000 randomly generated proteome allocations was selected using minimization of the residual substrate concentration as an objective function. Subsequently, this procedure was repeated for the next lower dilution rate. The optimal proteome allocation of the previous dilution rate was used as a starting value. To validate the model, the predicted fluxes and metabolite concentrations were compared with a flux and metabolome dataset (Suarez-Mendez et al., 2016) at different dilution rates under chemostat conditions. This comparison of predicted and measured fluxes and metabolite concentrations can be found in **Supplementary Material S3** and in **Supplementary Figures S2, S3**, respectively.

The experimental data for ethanol production and oxygen consumption in **Figure 2** show that the ethanol production starts at a dilution rate of 0.28 h^{-1} (Rieger et al., 1983; Van Hoek et al., 1998). Above this critical dilution rate, the oxygen consumption rate decreases, while ethanol production keeps increasing. Ethanol production is first predicted by the model for a dilution rate of 0.25 h^{-1}, which is a lower rate than the experimental data. Furthermore, there is no decrease in the oxygen consumption rate above a dilution rate of 0.28 h^{-1} for the optimized proteomes, which was observed in experimental studies (Van Hoek et al., 1998). From the model, this can be explained by the proteome-specific ATP production "cost": Respiration has a high yield compared to fermentation (**Table 2.**). Hence, reducing the size of the respiration proteome sector will not be predicted by the model as it is not beneficial. The predicted plateau originates from a constraint that was introduced manually (12% of the proteome for respiration) to reflect the maximum oxygen consumption rate measured by Rieger et al. (1983) after long-term evolution. The continuous increase in the ethanol production rate can then be explained by the increasing need for ATP with an increasing growth rate while respiration is at its maximum.

This result conflicts with the discussed dataset of Van Hoek et al. (1998) as well as the model predictions of Nilsson and Nielsen (2016), which was partly based on this experimental dataset. This mismatch and conclusions will be discussed in more detail later. Notably, there is also experimental evidence from previous studies that the predicted plateau is reasonable. It was

TABLE 2 | Comparison of the proteome-specific ATP yield for fermentation and respiration obtained by Nilsson and Nielsen(2016) and this study. Values of this study were derived from simulations performed at a growth rate of 0.4 h^{-1}.

	Fermentation ($mol_{ATP}/g_{prot}/h$)	Respiration ($mol_{ATP}/g_{prot}/h$)
Nilsson and Nielsen (2016)	0.40	0.21
This study	0.18	0.20

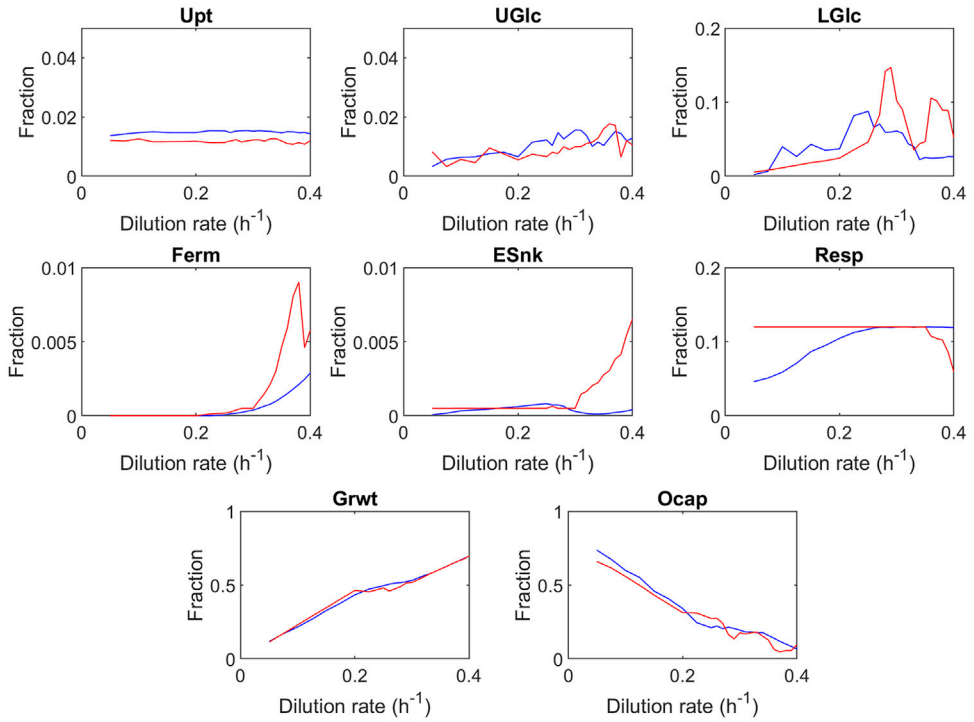


FIGURE 3 | Predicted proteome fractions at steady-state as a function of the dilution rate. Blue represents the best proteome out of 1,000 randomly generated proteomes; red represents the best proteome out of 100 randomly generated proteomes. The values shown are averaged over 40 optimization runs, and the single results are displayed in the **Supplementary Figure S4**. Upt, uptake; UGlc, upper glycolysis; LGlc, lower glycolysis; Ferm, fermentation; Esnk, electron sink/glycerol pathway; Resp, respiration; Grwt, growth; Ocap, overcapacity. The trehalose sector was decreased to zero in all instances of the overcapacity simulation, and therefore not shown in the figure.

shown that the respiratory repression observed by Van Hoek et al. (1998) could be negated upon long-term adaptation (Barford and Hall, 1979; Rieger et al., 1983; Postma et al., 1989), and a stable maximum oxygen uptake rate above a dilution rate of 0.28 h^{-1} was found.

To test the hypothesis of short- vs. long-term evolution, the proteome optimization approach was performed with a reduced number of generated proteomes and compared to the experimental data of Van Hoek et al. (1998) (**Figure 2**, red line). With a high number of generated proteomes for the optimization, the experimental findings of long-term chemostats could be reproduced. From these predictions, we hypothesize that cells not exposed to long-term glucose-limited conditions did not yet reach the "optimal" proteome allocation and respective metabolic phenotype. This set number of 1,000 simulations was chosen because only very limited further optimization of the objective function was observed after this number of simulations. As such, 1,000 simulations were concluded as sufficient to reach the optimum. Work on adapted glucose-grown cultures, at which point glucose repression on respiration disappears, is cultivated for at least 50 generations at the same dilution rate (Barford and Hall, 1979). A work by Van Hoek et al. (1998) describes the Crabtree effect with its typical glucose repression of respiration, by cultivating cultures at the same dilution rate for seven generations. Therefore, a set number of 100 simulations was chosen to reflect this state of limited adaptation of the proteome from batch growth conditions.

Looking into the global trends in the fully evolved proteome allocation at different dilution rates (**Figure 3**, see **Supplementary Material S4** for sensitivity analysis), an increase in the dilution rate can be seen for nearly all sectors

leading to the unused space (in the following called overcapacity sector, last panel). The overcapacity sector accounts for the fraction of the proteome which remains unused within the optimized proteomes. Before discussing specific trends, the high dilution rates will be highlighted. Even close to the maximal growth rate, the model predicts a small overcapacity sector. Nevertheless, please note that the batch and very high dilution rate might still have different optimization criteria; here, in the model, minimal substrate concentration was applied as the objective function. Because of the optimization approach, some robustness is required that was not further tuned as the fraction is rather small (7%) and does not change trends. Additionally, the algorithm samples from an enumerated number of randomly generated proteomes and therefore requires some buffer for robustness.

A major difference between this model and earlier studies (Nilsson and Nielsen, 2016) is that the proteome space limit is not reached at the critical growth rate ($D = 0.28\ h^{-1}$). At the critical dilution rate ($D = 0.28\ h^{-1}$), the overcapacity sector still has a significant fraction (21%). As briefly discussed earlier, Nilsson and Nielsen (2016) postulated that the Crabtree effect could be explained by the catalytic efficiency of the fermentation and respiration pathways expressed as ATP per amount of protein used in the pathway (**Table 2**). To estimate these catalytic efficiencies, Nilsson and Nielsen (2016) used the fluxes and specific enzyme activities for fermentation and respiration, under the assumption that all enzymes operate at half of their maximum specific activity , whereas in this model, the estimation of the catalytic efficiency is based upon the proteome and fluxome dataset, using dynamic saturation of enzymes. The estimation proposed by Nilsson and Nielsen (2016) subsequently produced a proteome composition in which the mass of all respiration proteins is 19 times larger than the protein mass of all glycolysis enzymes , while from proteome measurements it was observed that the mass of all respiration proteins is 0.3 times the size of the mass of all glycolysis proteins (de Godoy et al., 2008; Elsemman et al., 2022). This large difference in proteome allocation between glycolysis and respiration causes the catalytic efficiency of fermentation to be overestimated. The conclusion that the proteome is fully allocated after the critical growth rate is reached leads to the prediction that the "optimal" endpoint of proteome allocation is reached, which cannot explain datasets by Barford and Hall (1979); Rieger et al. (1983). Additional modeling studies by Elsemman et al. (2022) suggest that the decrease in oxygen consumption at higher growth rates observed by Van Hoek et al. (1998) is not caused by a limitation in proteome capacity but rather by a maximum rate of mitochondria biogenesis, in which long-term adaptation could overcome the described glucose repression of respiration.

Prediction of Proteome Allocation Under Dynamic Conditions

The proteome compositions, especially at low dilution rates were characterized by a significant overcapacity sector. The kinetic proteome allocation approach could not yet answer why the cells maintained such an excess proteome. As discussed earlier, the hypothesis for a proteome overcapacity is competitiveness and robustness including dynamic environmental conditions. Overcapacity could enable faster substrate uptake rates and enable a competitive advantage and outcompete slower consuming microbes (Jannasch, 1967). Furthermore, excess capacity could enable a robust, balanced functioning of pathways such as glycolysis (van Heerden et al., 2014) under dynamic substrate conditions.

To test these hypotheses, we studied the predicted proteome allocation under different repetitive substrate-feeding regimes using the proteome-dependent kinetic model, using the minimization of the time-weighted residual substrate concentration as the objective function. With this approach, we were able to select competitive proteomes with fast substrate uptake rates. As a reference dynamic condition, an experimentally explored feeding regime was chosen, i.e., a cycle length of 400 s of which 20 s was used to feed the culture ($D = 2\ h^{-1}$), leading to the average dilution rate of $D = 0.1\ h^{-1}$ over the complete cycle (Suarez-Mendez et al., 2014).

Proteome allocations and respective metabolic phenotypes were then compared to the steady-state at the same (average) growth rate. First, we studied the maximum, minimum, and average enzyme saturation (V/V_{max}) under dynamic conditions compared to the enzyme saturation under chemostat conditions (**Table 3**). Under dynamic conditions, the maximal enzyme saturation is much higher (up to 92% for the respiration reaction) than that under chemostat conditions (77% for respiration). Nevertheless, the average enzyme saturation over the whole cycle is actually lower than that under the reference chemostat state (for respiration, 25% compared to 77% at steady-state). This indicates that the proteome optimization to some extent focuses on the "peak" flux, especially for the large sectors of respiration and growth, indicating high usage of the available flux capacity while on average leaving a large overcapacity over the whole cycle. This enables a rapid consumption of the substrate as soon as it becomes available, which was the optimization criteria.

We were now curious to see how the perturbation strength would influence the proteome allocation. Therefore, the length of the feeding period was varied at the same average dilution rate, resulting in different substrate perturbation intensities. The respective predicted proteome allocations were calculated and compared (**Figure 4**) for the different ratios of feeding time over cycle time (TF/TC). TF/TC values were chosen as log2 increments from the experimentally used TF/TC value of 1/20 (Suarez-Mendez et al., 2014).

The growth sector fraction increased with the perturbation intensity, suggesting that this strategy was the most effective measure to survive the higher substrate concentration variations (from faster feeding) and consequently high flux dynamics. The growth reaction seemed to act as an efficient and fast sink for substrate and ATP. However, in reality, the growth sector does not consist of a single reaction and may not be able to provide a rapid response upon glucose influx. For this reason, two other scenarios were additionally evaluated: 1) the regulation of the trehalose cycle upon repeated substrate pulses and 2) the

TABLE 3 | Enzyme saturation, i.e., v/v_{max} under dynamic feeding conditions compared to steady-state (both at a dilution rate of D = 0.1 h-1). For dynamic conditions, v/v_{max} is calculated at the maximum rate during the cycle and the minimum as well as the average over the cycle. Upt, uptake; UGlc, upper glycolysis; LGlc, lower glycolysis; Ferm, fermentation; Esnk, electron sink/glycerol pathway; Resp, respiration; TrSn, trehalose synthesis; TrDg, trehalose degradation; Grwt, growth.

	Upt	Uglc	LGlc	Ferm	Esnk	Resp	TrSn	TrDg	Grwt
Max V/V_{max} ratio under FF	6%	19%	8%	25%	46%	92%	79%	11%	100%
Min V/V_{max} ratio under FF	<1%	<1%	<1%	<1%	<1%	1%	<1%	4%	<1%
Average V/V_{max} ratio under FF	1%	3%	1%	4%	7%	25%	10%	7%	24%
V/V_{max} ratio under chemostat	<1%	2%	6%	-	13%	77%	-	-	74%

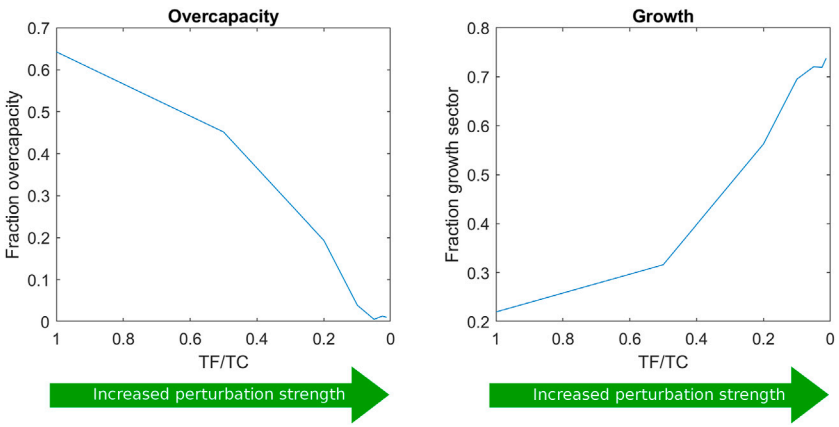


FIGURE 4 | Proteome allocation as a function of the ratio of feeding time over cycle time (TF/TC). Further proteome sector fractions are shown in **Supplementary Material S5**.

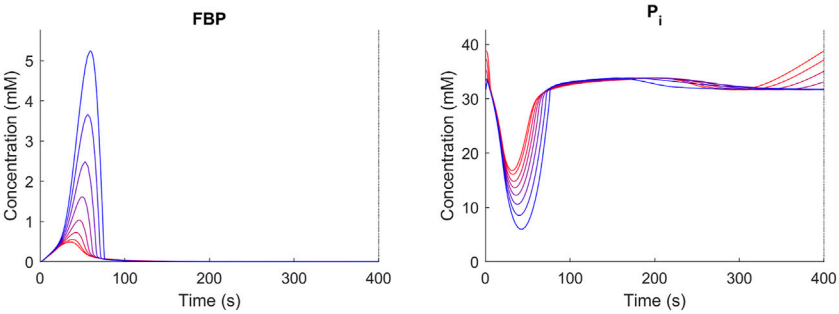


FIGURE 5 | Concentration time course over repetitive cycles (D = 0.1 h^{-1}, TF/TC = 0.05) for different trehalose sector fractions (blue = 0.1 red 1%). Shown are FBP and Pi as representative metabolites. For trehalose sector fractions <0.1%, no stable cycles were obtained.

regulation of the ratio between upper and lower glycolysis (see **Supplementary Material S6**).

Impact of the Proteome Fraction on the Trehalose Cycle

The trehalose cycle has been described to function as a "safety valve" upon large changes in the glycolytic flux (Thevelein and Hohmann, 1995; Blomberg, 2000; van Heerden et al., 2014; Vicente et al., 2018). Under dynamic conditions in yeast, it was found that a significant amount of imported glucose was recycled through the trehalose cycle, especially during periods of high flux changes (Suarez-Mendez et al., 2017). To evaluate the effect of storage metabolism activity under dynamic conditions, the reference condition [D = 0.1 h-1, TF/TC = 0.05, (Suarez-Mendez et al., 2014)], was further analyzed. We varied the trehalose sector size between 0 and 1% (**Figure 5**) and compared the response of metabolism using FBP and Pi as indicators. A balanced metabolic response will lead to repetitive cycles in FBP and Pi. Such repetitive response was observed for proteomes with a trehalose sector larger than 0.1%. Increasing the trehalose sector above 0.1% leads to reduced

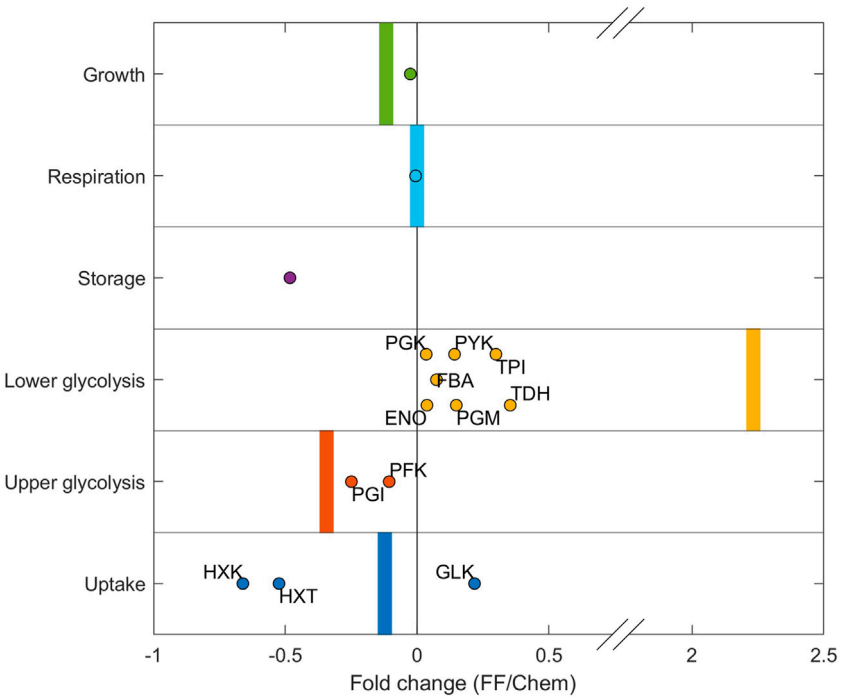


FIGURE 6 | Protein concentration fold change from chemostat to feast/famine cultivation. The experimental fold change individual proteins are displayed as dots. Proteins of trehalose/glycogen storage, ribosomes, and oxidative phosphorylation were used as proxies for the storage, growth, and respiration sectors, respectively. Simulation fold changes for each sector are shown as vertical bars. The simulated storage sector for steady-state conditions was 0 and increased under simulated feast/famine conditions. As such, no fold change could be calculated, and therefore this fold change is not shown.

fluctuations in G6P/FBP and Pi, suggesting a more robust metabolic response. Simulated changes in FBP and Pi are in line with results from previous work by van Heerden et al. (2014).

Comparison of the Model Predictions to Experimental Proteomes

To evaluate the prediction accuracy and trends of the predicted proteomes under dynamic conditions, the simulated proteome adaptation from chemostat to feast/famine conditions was compared with the experimentally measured proteome fold changes between chemostat and feast/famine conditions (Verhagen et al., 2022) (**Figure 6**). Proteins of trehalose/glycogen storage, ribosomes, and oxidative phosphorylation were used as proxies for the storage, growth, and respiration sectors, respectively (proteins categorized in the same way as calibration approach, see Methods).

The model predicted the experimentally observed changes in upper and lower glycolysis (**Figure 6**). The enzyme TDH catalyzes the glyceraldehyde dehydrogenase reaction (TDH), which forms ATP using Pi. However, if upper and lower glycolysis reactions are imbalanced during high fluxes, this reaction becomes a bottleneck, leading to the accumulation of FBP and subsequently to an imbalanced metabolism. Therefore, it was expected that TDH had to be upregulated under substrate-fluctuating conditions to facilitate balanced intermediates, which was reflected in both the model predictions as well as in the experimental dataset. The predicted change in lower glycolysis is larger than that in the experimental data. This is likely caused by the fact that simulated proteomes for chemostat conditions contain no overcapacity in the lower glycolysis sector, whereas experimental proteomes under chemostat conditions appear to contain more overcapacity in this proteome sector. As such, the fold change between measured and simulated values is higher. Furthermore, the model reproduced the average change observed for the uptake sector, although it should be noted that effects of individual iso-enzymes (especially with regard to HXK/GLK, which catalyzes the first step of glycolysis) were not taken into account in the current model.

Significant deviations between experimental and predicted fractions were observed for the storage sector. This was significantly decreased experimentally, while the resource-dependent kinetic model predicted an increase. Experimentally, a decrease of 28%, from 0.25 to 0.2% of the proteome, was observed, while an increase to 0.2% of the proteome was predicted in the model. Possible reasons for this difference in sector size could be: 1) the synthesis of trehalose has additional functions in the cell which are not represented in the model—it is described that trehalose plays an important role in different stress responses, including severe substrate limitation at low dilution rates (see also **Supplementary Figure S3**). 2) The measured and predicted proteomes do not include neither post-translational modifications, which are known to significantly affect the k_{cat}'s of enzymes in the trehalose cycle (Sengupta

et al., 2011), nor changes that could occur during cell-cycle progression.

Furthermore, there could also be a bias from the experimental setup–the differences in the trehalose sector, combined with the observed increase of the lower glycolysis sector, suggest that the experimental chemostat proteome is potentially already primed for dynamic environments, compared to experimental conditions, and as such is more robust than the predicted optimized chemostat proteomes.

CONCLUSION AND OUTLOOK

In this work, we developed a proteome-dependent kinetic modeling framework that predicts the optimal proteome composition for defined extracellular dynamic conditions. The approach could reproduce observed complex metabolic phenomena, such as the Crabtree effect, including long-term adjustments under chemostat conditions.

Analysis of the predicted proteomes showed that under substrate-limiting conditions (i.e., low dilution rates) with close to constant extracellular concentrations, a significant part of the optimized proteome is not required (thus a lot of overcapacity). With increasing substrate availability and/or concentration fluctuations, this overcapacity is shown to decrease. Cells optimized for steady-state conditions were not able to survive these substrate perturbations. This suggests that in reality, when conditions are never as ideal and "optimal" as presented in the model simulations, cells already possess proteome adjustments to create a more robust metabolism, allowing them to cope effectively with external perturbations such as substrate gradients.

Such adjustments to perturbations were found when comparing steady-state and feast/famine condition predictions. The approach generated a stable phenotype and the predicted changes in proteome allocation, i.e., downregulation of uptake and upper glycolysis sectors and upregulation of the lower glycolysis sector were also found experimentally. This complex and strongly kinetics-dependent prediction highlights the relevance of kinetic properties also for the regulation of protein expression. Nevertheless, to achieve this prediction, some constraints, which had to be derived from experimental observations, had to be included: the maximum mitochondrial fraction and the glucose repression on fermentation. These boundaries seemed to be only stretched after very long-term evolution, as observed by Barford and Hall (1979). Following this observation, the model was used to predict the proteome composition and metabolic behavior of cells at different stages of adaptation, able to simulate differences in cultivation history. Thus, the modeling approach was able to cover a large range of conditions and evolution outcomes, which could be specifically relevant for the prediction of production process regimes running over a long time span.

DATA AVAILABILITY STATEMENT

The dataset and the MATLAB scripts for the model presented in this study can be found in online repositories at: https://doi.org/10.4121/19008833, https://doi.org/10.4121/19074791.

AUTHOR CONTRIBUTIONS

KV: conceptualization, model development, writing–original draft, and writing–review and editing, and visualization. SE: conceptualization and model development. BS: conceptualization and model development. SW: conceptualization, writing–review and editing, and supervision.

FUNDING

This work was supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) (project number 737.016.001).

ACKNOWLEDGMENTS

We thank Pascale Daran-Lapujade, David Lao-Martil, Joep Schmitz, and Bas Teusink for their input and useful discussions. In addition, we thank André Vente for his help in the analysis of the proteomics data.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.863470/full#supplementary-material

REFERENCES

Abulesz, E.-M., and Lyberatos, G. (1989). Periodic Operation of a Continuous Culture of Baker's Yeast. *Biotechnol. Bioeng.* 34, 741–749. doi:10.1002/bit.260340603

Barford, J. P., and Hall, R. J. (1979). An Examination of the Crabtree Effect in *Saccharomyces cerevisiae*: the Role of Respiratory Adaptation. *J. Gen. Microbiol.* 114, 267–275. doi:10.1099/00221287-114-2-267

Basan, M., Hui, S., Okano, H., Zhang, Z., Shen, Y., Williamson, J. R., et al. (2015). Overflow Metabolism in *Escherichia coli* Results from Efficient Proteome Allocation. *Nature* 528, 99–104. doi:10.1038/nature15765

Basan, M. (2018). Resource Allocation and Metabolism: the Search for Governing Principles. *Curr. Opin. Microbiol.* 45, 77–83. doi:10.1016/j.mib.2018.02.008

Blomberg, A. (2000). Metabolic Surprises inSaccharomyces Cerevisiaeduring Adaptation to saline Conditions: Questions, Some Answers and a Model. *FEMS Microbiol. Lett.* 182, 1–8. doi:10.1111/j.1574-6968.2000.tb08864.x

Boer, V. M., Crutchfield, C. A., Bradley, P. H., Botstein, D., and Rabinowitz, J. D. (2010). Growth-limiting Intracellular Metabolites in Yeast Growing under Diverse Nutrient Limitations. *MBoC* 21, 198–211. doi:10.1091/mbc.e09-07-0597

Brejning, J., and Jespersen, L. (2002). Protein Expression during Lag Phase and Growth Initiation in *Saccharomyces cerevisiae*. *Int. J. Food Microbiol.* 75, 27–38. doi:10.1016/S0168-1605(01)00726-7

Chu, D., and Barnes, D. J. (2016). The Lag-phase during Diauxic Growth Is a Trade-Off between Fast Adaptation and High Growth Rate. *Sci. Rep.* 6, 25191. doi:10.1038/srep25191

Chubukov, V., Gerosa, L., Kochanowski, K., and Sauer, U. (2014). Coordination of Microbial Metabolism. *Nat. Rev. Microbiol.* 12, 327–340. doi:10.1038/nrmicro3238

Cohen, A. A., Geva-Zatorsky, N., Eden, E., Frenkel-Morgenstern, M., Issaeva, I., Sigal, A., et al. (2008). Dynamic Proteomics of Individual Cancer Cells in Response to a Drug. *Science* 322, 1511–1516. doi:10.1126/science.1160165

de Godoy, L. M. F., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Fröhlich, F., et al. (2008). Comprehensive Mass-Spectrometry-Based Proteome Quantification of Haploid versus Diploid Yeast. *Nature* 455, 1251–1254. doi:10.1038/nature07341

Dekel, E., and Alon, U. (2005). Optimality and Evolutionary Tuning of the Expression Level of a Protein. *Nature* 436, 588–592. doi:10.1038/nature03842

Della-Bianca, B. E., de Hulster, E., Pronk, J. T., van Maris, A. J. A., and Gombert, A. K. (2014). Physiology of the Fuel Ethanol strainSaccharomyces cerevisiaePE-2 at Low pH Indicates a Context-dependent Performance Relevant for Industrial Applications. *FEMS Yeast Res.* 14, 1196–1205. doi:10.1111/1567-1364.12217

Elsemman, I. E., Rodriguez Prado, A., Grigaitis, P., Garcia Albornoz, M., Harman, V., Holman, S. W., et al. (2022). Whole-cell Modeling in Yeast Predicts Compartment-specific Proteome Constraints that Drive Metabolic Strategies. *Nat. Commun.* 13, 801. doi:10.1038/s41467-022-28467-6

Ertugay, N., and Hamamci, H. (1997). Continuous Cultivation of Bakers' Yeast: Change in Cell Composition at Different Dilution Rates and Effect of Heat Stress on Trehalose Level. *Folia Microbiol.* 42, 463–467. doi:10.1007/BF02826554

Gerosa, L., and Sauer, U. (2011). Regulation and Control of Metabolic Fluxes in Microbes. *Curr. Opin. Biotechnol.* 22, 566–575. doi:10.1016/j.copbio.2011.04.016

Goelzer, A., Muntel, J., Chubukov, V., Jules, M., Prestel, E., Nölker, R., et al. (2015). Quantitative Prediction of Genome-wide Resource Allocation in Bacteria. *Metab. Eng.* 32, 232–243. doi:10.1016/j.ymben.2015.10.003

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 Genes. *Science* 274, 546–567. doi:10.1126/science.274.5287.546

Heyland, J., Fu, J., and Blank, L. M. (2009). Correlation between TCA Cycle Flux and Glucose Uptake Rate during Respiro-Fermentative Growth of *Saccharomyces cerevisiae*. *Microbiology* 155, 3827–3837. doi:10.1099/mic.0.030213-0

Jannasch, H. W. (1967). Enrichments of Aquatic Bacteria in Continuous Culture. *Archiv. Mikrobiol.* 59, 165–173. doi:10.1007/BF00406328

Jõers, A., and Tenson, T. (2016). Growth Resumption from Stationary Phase Reveals Memory in *Escherichia coli* Cultures. *Sci. Rep.* 6, 24055. doi:10.1038/srep24055

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a Reference Resource for Gene and Protein Annotation. *Nucleic Acids Res.* 44, D457–D462. doi:10.1093/nar/gkv1070

Karim, M. R., Kawanago, H., and Kadowaki, M. (2014). A Quick Signal of Starvation Induced Autophagy: Transcription versus post-translational Modification of LC3. *Anal. Biochem.* 465, 28–34. doi:10.1016/j.ab.2014.07.007

Liebermeister, W., Noor, E., Flamholz, A., Davidi, D., Bernhardt, J., and Milo, R. (2014). Visual Account of Protein Investment in Cellular Functions. *Proc. Natl. Acad. Sci. U.S.A.* 111, 8488–8493. doi:10.1073/pnas.1314810111

Litsios, A., Ortega, Á. D., Wit, E. C., and Heinemann, M. (2018). Metabolic-flux Dependent Regulation of Microbial Physiology. *Curr. Opin. Microbiol.* 42, 71–78. doi:10.1016/j.mib.2017.10.029

Mahadevan, R., Edwards, J. S., and Doyle, F. J. (2002). Dynamic Flux Balance Analysis of Diauxic Growth in *Escherichia coli*. *Biophysical J.* 83, 1331–1340. doi:10.1016/S0006-3495(02)73903-9

Molenaar, D., van Berlo, R., de Ridder, D., and Teusink, B. (2009). Shifts in Growth Strategies Reflect Tradeoffs in Cellular Economics. *Mol. Syst. Biol.* 5, 323. doi:10.1038/msb.2009.82

Mori, M., Marinari, E., and De Martino, A. (2019). A Yield-Cost Tradeoff Governs *Escherichia coli*'s Decision between Fermentation and Respiration in Carbon-Limited Growth. *Npj Syst. Biol. Appl.* 5, 16. doi:10.1038/s41540-019-0093-4

Nagaraj, N., Alexander Kulak, N., Cox, J., Neuhauser, N., Mayr, K., Hoerning, O., et al. (2012). System-wide Perturbation Analysis with Nearly Complete Coverage of the Yeast Proteome by Single-Shot Ultra HPLC Runs on a Bench Top Orbitrap. *Mol. Cell Proteomics* 11, M111.013722. doi:10.1074/mcp.M111.013722

Nilsson, A., and Nielsen, J. (2016). Metabolic Trade-Offs in Yeast Are Caused by F1F0-ATP Synthase. *Sci. Rep.* 6, 22264. doi:10.1038/srep22264

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What Is Flux Balance Analysis? *Nat. Biotechnol.* 28, 245–248. doi:10.1038/nbt.1614

Pincus, D., Resnekov, O., and Reynolds, K. A. (2017). An Evolution-Based Strategy for Engineering Allosteric Regulation. *Phys. Biol.* 14, 025002. doi:10.1088/1478-3975/aa64a4

Postma, E., Verduyn, C., Scheffers, W. A., and Van Dijken, J. P. (1989). Enzymic Analysis of the crabtree Effect in Glucose-Limited Chemostat Cultures of *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.* 55, 468–477. doi:10.1128/aem.55.2.468-477.1989

Reimers, A.-M., Knoop, H., Bockmayr, A., and Steuer, R. (2017). Cellular Trade-Offs and Optimal Resource Allocation during Cyanobacterial Diurnal Growth. *Proc. Natl. Acad. Sci. U.S.A.* 114, E6457–E6465. doi:10.1073/pnas.1617508114

Rieger, M., KAPpeli, O., and Fiechter, A. (1983). The Role of Limited Respiration in the Incomplete Oxidation of Glucose by Saccharomyces Cerevisiae. *Microbiology* 129, 653–661. doi:10.1099/00221287-129-3-653

Rügen, M., Bockmayr, A., and Steuer, R. (2015). Elucidating Temporal Resource Allocation and Diurnal Dynamics in Phototrophic Metabolism Using Conditional FBA. *Sci. Rep.* 5, 15247. doi:10.1038/srep15247

Schuetz, R., Kuepfer, L., and Sauer, U. (2007). Systematic Evaluation of Objective Functions for Predicting Intracellular Fluxes in *Escherichia coli*. *Mol. Syst. Biol.* 3, 119. doi:10.1038/msb4100162

Sengupta, S., Chaudhuri, P., Lahiri, S., Dutta, T., Banerjee, S., Majhi, R., et al. (2011). Possible Regulation of Trehalose Metabolism by Methylation in *Saccharomyces cerevisiae*. *J. Cel. Physiol.* 226, 158–164. doi:10.1002/jcp.22317

Smallbone, K., Malys, N., Messiha, H. L., Wishart, J. A., and Simeonidis, E. (2011). "Building a Kinetic Model of Trehalose Biosynthesis in *Saccharomyces cerevisiae*," in *Methods in Enzymology*. 1st ed (Elsevier), 355–370. doi:10.1016/B978-0-12-385118-5.00018-9

Suarez-Mendez, C., Sousa, A., Heijnen, J., and Wahl, A. (2014). Fast "Feast/Famine" Cycles for Studying Microbial Physiology under Dynamic Conditions: A Case Study with *Saccharomyces cerevisiae*. *Metabolites* 4, 347–372. doi:10.3390/metabo4020347

Suarez-Mendez, C. A., Hanemaaijer, M., ten Pierick, A., Wolters, J. C., Heijnen, J. J., and Wahl, S. A. (2016). Interaction of Storage Carbohydrates and Other Cyclic Fluxes with central Metabolism: A Quantitative Approach by Non-stationary 13 C Metabolic Flux Analysis. *Metab. Eng. Commun.* 3, 52–63. doi:10.1016/j.meteno.2016.01.001

Suarez-Mendez, C. A., Ras, C., and Wahl, S. A. (2017). Metabolic Adjustment upon Repetitive Substrate Perturbations Using Dynamic 13C-Tracing in Yeast. *Microb. Cel Fact* 16, 161. doi:10.1186/s12934-017-0778-6

Suarez-Mendez, C. A. (2015). *Dynamics of Storage Carbohydrates Metabolism in* Saccharomyces cerevisiae. Delft, Netherlands: Delft University of Technology. doi:10.4233/UUID:2504BD76-9811-4D3C-A66B-3AAE7FBB40B5

Teusink, B., Passarge, J., Reijenga, C. A., Esgalhado, E., Van Der Weijden, C. C., Schepper, M., et al. (2000). Can Yeast Glycolysis Be Understood in Terms of *In Vitro* Kinetics of the Constituent Enzymes? Testing Biochemistry. *Eur. J. Biochem.* 267, 5313–5329. doi:10.1046/j.1432-1327.2000.01527.x

Thevelein, J. M., and Hohmann, S. (1995). Trehalose Synthase: Guard to the Gate of Glycolysis in Yeast? *Trends Biochem. Sci.* 20, 3–10. doi:10.1016/S0968-0004(00)88938-0

van Heerden, J. H., Wortel, M. T., Bruggeman, F. J., Heijnen, J. J., Bollen, Y. J. M., Planqué, R., et al. (2014). Lost in Transition: Start-Up of Glycolysis Yields Subpopulations of Nongrowing Cells. *Science* 343, 1245114. doi:10.1126/science.1245114

Van Hoek, P., Van Dijken, J. P., and Pronk, J. T. (1998). Effect of Specific Growth Rate on Fermentative Capacity of Baker's Yeast. *Appl. Environ. Microbiol.* 64, 4226–4233. doi:10.1128/AEM.64.11.4226-4233.1998

Varma, A., and Palsson, B. O. (1994). Stoichiometric Flux Balance Models Quantitatively Predict Growth and Metabolic By-Product Secretion in Wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.* 60, 3724–3731. doi:10.1128/aem.60.10.3724-3731.1994

Vazquez, A., Beg, Q. K., deMenezes, M. A., Ernst, J., Bar-Joseph, Z., Barabási, A.-L., et al. (2008). Impact of the Solvent Capacity Constraint on *E. coli* Metabolism. *BMC Syst. Biol.* 2, 7. doi:10.1186/1752-0509-2-7

Verhagen, K. J. A., Eerden, S. A., and Wahl, S. A. (2022). Data from: Dataset Proteomics: Analysis of Change in Protein Expression in *Saccharomyces cerevisiae* upon Shift from Glucose Chemostat to Feast/famine Regime. *4TU.ResearchData*. doi:10.4121/19008833

Vicente, R. L., Spina, L., Gómez, J. P. L., Dejean, S., Parrou, J.-L., and François, J. M. (2018). Trehalose-6-phosphate Promotes Fermentation and Glucose Repression in *Saccharomyces cerevisiae*. *Microb. Cel* 5, 444–459. doi:10.15698/mic2018.10.651

Vos, T., Hakkaart, X. D. V., de Hulster, E. A. F., van Maris, A. J. A., Pronk, J. T., and Daran-Lapujade, P. (2016). Maintenance-energy Requirements and Robustness of *Saccharomyces cerevisiae* at Aerobic Near-Zero Specific Growth Rates. *Microb. Cel Fact* 15, 1–20. doi:10.1186/s12934-016-0501-z

Wegner, A., Meiser, J., Weindl, D., and Hiller, K. (2015). How Metabolites Modulate Metabolic Flux. *Curr. Opin. Biotechnol.* 34, 16–22. doi:10.1016/j.copbio.2014.11.008

Wu, L., van Dam, J., Schipper, D., Kresnowati, M. T. A. P., Proell, A. M., Ras, C., et al. (2006). Short-Term Metabolome Dynamics and Carbon, Electron, and ATP Balances in Chemostat-Grown *Saccharomyces cerevisiae* CEN.PK 113-7D Following a Glucose Pulse. *Appl. Environ. Microbiol.* 72, 3566–3577. doi:10.1128/AEM.72.5.3566-3577.2006

Zhao, Q., Stettner, A. I., Reznik, E., Paschalidis, I. C., and Segrè, D. (2016). Mapping the Landscape of Metabolic Goals of a Cell. *Genome Biol.* 17, 109. doi:10.1186/s13059-016-0968-2

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Verhagen, Eerden, Sikkema and Wahl. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 18 May 2022
doi: 10.3389/fmolb.2022.856030

Modeling Calcium Signaling in *S. cerevisiae* Highlights the Role and Regulation of the Calmodulin-Calcineurin Pathway in Response to Hypotonic Shock

Simone Spolaor[1], Mattia Rovetta[1], Marco S. Nobile[2,3,4], Paolo Cazzaniga[5,3,4], Renata Tisi[6] and Daniela Besozzi[1,3,4]**

[1]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy, [2]Department of Environmental Sciences, Informatics and Statistics, Ca' Foscari University of Venice, Venice, Italy, [3]Bicocca Bioinformatics, Biostatistics and Bioimaging Centre—B4, Milan, Italy, [4]SYSBIO/ISBE.IT Centre of Systems Biology, Milan, Italy, [5]Department of Human and Social Sciences, University of Bergamo, Bergamo, Italy, [6]Department of Biotechnology and Biosciences, University of Milano-Bicocca, Milan, Italy

OPEN ACCESS

Edited by:
Alberto Jesus Martin, Universidad Mayor, Chile

Reviewed by:
Laszlo Csernoch, University of Debrecen, Hungary
Shivendra Gajraj Tewari, Biotechnology HPC Software Applications Institute (BHSAI), United States

***Correspondence:**
Renata Tisi
renata.tisi@unimib.it
Daniela Besozzi
daniela.besozzi@unimib.it

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *16 January 2022*
Accepted: *04 April 2022*
Published: *18 May 2022*

Citation:
Spolaor S, Rovetta M, Nobile MS, Cazzaniga P, Tisi R and Besozzi D (2022) Modeling Calcium Signaling in S. cerevisiae Highlights the Role and Regulation of the Calmodulin-Calcineurin Pathway in Response to Hypotonic Shock. Front. Mol. Biosci. 9:856030. doi: 10.3389/fmolb.2022.856030

Calcium homeostasis and signaling processes in *Saccharomyces cerevisiae*, as well as in any eukaryotic organism, depend on various transporters and channels located on both the plasma and intracellular membranes. The activity of these proteins is regulated by a number of feedback mechanisms that act through the calmodulin-calcineurin pathway. When exposed to hypotonic shock (HTS), yeast cells respond with an increased cytosolic calcium transient, which seems to be conditioned by the opening of stretch-activated channels. To better understand the role of each channel and transporter involved in the generation and recovery of the calcium transient—and of their feedback regulations—we defined and analyzed a mathematical model of the calcium signaling response to HTS in yeast cells. The model was validated by comparing the simulation outcomes with calcium concentration variations before and during the HTS response, which were observed experimentally in both wild-type and mutant strains. Our results show that calcium normally enters the cell through the High Affinity Calcium influx System and mechanosensitive channels. The increase of the plasma membrane tension, caused by HTS, boosts the opening probability of mechanosensitive channels. This event causes a sudden calcium pulse that is rapidly dissipated by the activity of the vacuolar transporter Pmc1. According to model simulations, the role of another vacuolar transporter, Vcx1, is instead marginal, unless calcineurin is inhibited or removed. Our results also suggest that the mechanosensitive channels are subject to a calcium-dependent feedback inhibition, possibly involving calmodulin. Noteworthy, the model predictions are in accordance with literature results concerning some aspects of calcium homeostasis and signaling that were not specifically addressed within the model itself, suggesting that it actually depicts all the main cellular components and interactions that constitute the HTS calcium pathway, and thus can correctly reproduce the shaping of the calcium signature by calmodulin- and calcineurin-dependent complex regulations. The model predictions also allowed to provide an interpretation of different regulatory schemes involved in calcium

handling in both wild-type and mutants yeast strains. The model could be easily extended to represent different calcium signals in other eukaryotic cells.

Keywords: calcium signaling, hypotonic shock response, mathematical model, Saccharomyces cerevisiae, calmodulin-calcineurin pathway

1 INTRODUCTION

Calcium ions (Ca^{2+}) have many physiological functions and are ubiquitously used by prokaryotic and eukaryotic unicellular organisms, as well as by multicellular eukaryotes (Plattner and Verkhratsky, 2015b; Plattner and Verkhratsky, 2015a; Giorgi et al., 2018; Bagur and Hajnóczky, 2017). Ca^{2+} represents a universal intracellular messenger that modulates a plethora of processes, such as the control of cell proliferation, programmed cell death, neurotransmission, secretion, vesicular transport, cytoskeleton rearrangement, and transcription (Berridge et al., 2000). Ca^{2+} sequestration in different cellular compartments is the key to maintaining appropriate concentration gradients, which are required to generate the temporal and spatial patterns exploited by cells to encode signals about their own status or the extracellular environment (Berridge et al., 2003; Dupont et al., 2007; Purvis and Lahav, 2013). Specific Ca^{2+} signals are triggered only upon an intra- or extra-cellular stimulus; in the absence of such events, the cell must maintain properly low and non-signaling Ca^{2+} concentrations, in a narrow range of 50–200 nM in eukaryotic cells. This feature is implemented by an extensive and well conserved cellular toolkit comprised of Ca^{2+}-sensing proteins, buffers, channels, pumps, and exchangers (Plattner and Verkhratsky, 2015b). Although Ca^{2+} signals could possibly be generated in all cell compartments, they are mainly studied in the cytosol, and are the result of the release from intracellular stores or the influx from the extracellular environment, or both. The disruption of Ca^{2+} homeostasis system can potentially lead to unwanted (in)activation of signaling cascades, and in turn cause cell defects or even cell death (Carafoli, 2004; Sammels et al., 2010).

To help unraveling the role of channels and transporters in maintaining Ca^{2+} homeostasis, in this work we provide a mathematical model of the hypotonic shock (HTS) response in budding yeast cells. HTS consists in a sudden variation of the osmotic pressure, due to a consistent dilution of the solution concentration to which the cell is exposed. Following a HTS, water flows into the cell, causing an increase in cell volume and turgor pressure that might induce the cell burst. To avoid cytolysis, yeasts have evolved mechanisms to sense and respond to HTS, by rapidly triggering at least three different mechanisms: 1) the cytosolic Ca^{2+} concentration is transiently increased; 2) the osmolyte glycerol is released to the medium in order to relieve osmotic pressure; 3) phospholypase C hydrolizes $PI(4,5)P_2$ generating the major second messengers, diacylglycerol (DAG) and inositol-(3,4,5)-*tris*-phosphate (IP_3). In mammalian cells, DAG is well-known as a protein kinase C activator, while IP_3 is involved in triggering the release of Ca^{2+} from intracellular compartments such as endoplasmic reticulum or Golgi (Pinton et al., 1998; Berridge et al., 2000); it is not clear if this applies to yeast as well, but a role for inositol phosphate in calcium release was previously reported (Belde et al., 1993; Tisi et al., 2002, 2004).

The yeast *S. cerevisiae* has evolved a cell wall that, having less elasticity than the plasma membrane, prevents the cell from excessive expansion (Aguilar-Uscanga and Francois, 2003; Alsteens et al., 2008; Orlean, 2012). The adaptation of the cell wall to the environmental challenges is controlled by the cell wall integrity (CWI) signaling pathway, which shows a complex interrelationship with Ca^{2+} signaling (Hohmann, 2002; Levin, 2011).

The mathematical model presented in this work was defined by integrating well-established experimental evidences with plausible hypotheses on the functioning of HTS response. The main components of the model comprise: 1) the Ca^{2+} membrane transporters, namely, the High Affinity Calcium influx System (HACS) and mechanosensitive channels (MS); 2) the vacuolar transporters Pmc1 and Vcx1; 3) the Ca^{2+}-binding messenger protein calmodulin (CaM), and the Ca^{2+}-CaM-dependant phosphatase calcineurin (CaN). The model is formalised as a system of coupled Ordinary Differential Equations (ODEs), and it can be conceptually divided in two modules: a *biophysical module*, describing the physical properties of yeast cells (e.g., volume, turgor pressure, and membrane tension), and a *biochemical module*, describing the changes in the concentration of protein and molecules involved in Ca^{2+} signaling upon HTS, and including the feedback regulation via the calmodulin-calcineurin pathway.

The model simulations are in accordance with published experimental results, suggesting that the biophysical and the biochemical modules are able to explain the role and interplay among the essential components involved in the HTS response. Our analysis shows that calcium enters the resting cell through the HACS and the MS channels. However, upon HTS, the increase of the plasma membrane tension amplifies the opening probability of MS channels, thus causing a sudden Ca^{2+} pulse. The rapid recovery of the basal Ca^{2+} levels in the cytosol primarily involves Pmc1, differently than in other signalling processes where massive amounts of Ca^{2+} enter the cytoplasm requiring Vcx1 to engage in the recovery (Miseta et al., 1999b). According to our simulations, the involvement of Vcx1 is actually marginal, unless Vcx1 inhibition by calcineurin is removed. Our results also suggest that the MS channels are subject to a calcium-dependent feedback inhibition, possibly involving calmodulin, since it is not relieved by calcineurin removal. This suggests that the very sharp signature of HTS-induced calcium peak would be obtained by the rapid closure of the MS channels triggered by this feedback loop.

The paper is organized as follows: **Section 2** provides a detailed description of the cellular components and processes that were taken into account in the HTS response model; in

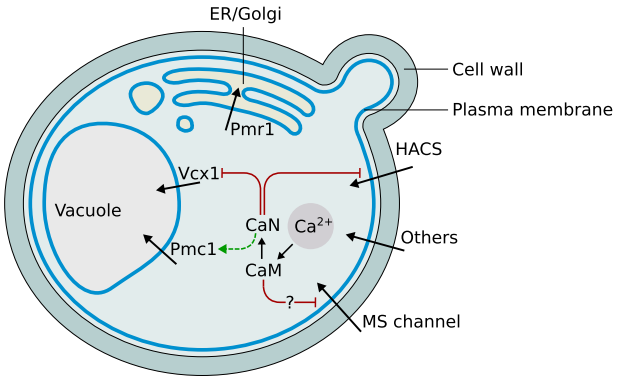


FIGURE 1 | Graphical schematization of the role and localization of the Ca^{2+} signaling components included in the model, and their mutual regulations. Abbreviations: ER = endoplasmic reticulum, HACS = High Affinity Calcium influx System, CaN = calcineurin, CaM = calmodulin, MS = mechanosensitive.

Section 3 we explain how the two modules of the model were formalised and simulated; **Section 4** provides a detailed description of the modules and their parameters; **Section 5** shows the results we obtained by comparing experimental data with our model simulations; finally, we discuss these results in **Section 6** and draw final conclusions in **Section 7**.

2 BIOLOGICAL BACKGROUND

The main cellular components appearing in the mathematical model are represented in **Figure 1**. Their biological function and mutual regulation in controlling Ca^{2+} homeostasis and signaling, especially upon HTS, are described in the following sections.

2.1 Calcium Transport and Homeostasis

2.1.1 Membrane Transporters

The High Affinity Calcium influx System (HACS) is composed of Cch1 and Mid1. Cch1 is a homolog of the pore-forming $\alpha 1$ subunit of mammalian voltage-gated Ca^{2+} channels (Fischer et al., 1997), while Mid1 is a stretch-activated Ca^{2+}-permeable nonselective cation channel whose secondary structure is similar to the non-pore-forming $\alpha 2\delta$ subunit that associates with the mammalian $\alpha 1$ (Iida et al., 1994). Cch1 and Mid1 cooperate in many yeast processes, such as mating pheromone-induced Ca^{2+} uptake (Paidhungat and Garrett, 1997), store-operated Ca^{2+} entry (Locke et al., 2000), ER stress-induced Ca^{2+} uptake (Bonilla et al., 2002), and hyperosmotic stress-induced increase of cytosolic Ca^{2+} (Matsumoto et al., 2002). Cch1 seems to respond to membrane depolarization, as the HACS is dependent on the presence of Kch1 and Kch2, two putative potassium transporters that mediate K^+ influx and, likely, plasma membrane depolarization in pheromone-induced conditions (Stefan C. P. et al., 2013), as well as upon ER stress (Stefan and Cunningham, 2013) or glucose addition (Ma et al., 2021). Some lines of evidence suggest that Mid1 and Cch1 can also function independently from each other (Kanzaki et al., 1999), in different compartments (Yoshimura et al., 2004) or upon different stimuli (Courchesne and Ozturk, 2003).

Besides other yet poorly characterized Ca^{2+} influx systems in the plasma membrane (Eilam and Othman, 1990; Muller et al., 2001; Muller et al., 2003; Groppi et al., 2011), the presence of another plasma membrane Ca^{2+} influx system was implied based on the fact that the hypotonic stress-induced $[Ca^{2+}]_{cyt}$ increase was not inhibited by removing all known transporters (Rigamonti et al., 2015). Its molecular identity is unknown but it is likely to include a Transient Receptor Potential (TRP)-like protein, Flc2, which provides the channel with a $[Ca^{2+}]_{ext}$-dependent inhibition. Flc2 is a member of the fungal spray family, which comprises TRP-like poly-cystic-kidney-disease (PKD)-related calcium channels (Tisi et al., 2016). TRP channels are well conserved and their regulation is polymodal. Almost all TRP channels appear to function as homo- or hetero-tetramers. TRP channels are regulated by a very large spectrum of chemical and physical stimuli such as phosphoinositides, Ca^{2+}, cyclic nucleotides, temperature, voltage, osmotic stress, and membrane shearing. A single TRP channel can exhibit sensitivity to multiple types of stimuli and thus mediates integrated responses (Zheng, 2013).

Differently than in plant and animal cells, Ca^{2+} efflux proteins have not been detected on the plasma membrane of yeast cells. However, Ca^{2+} is presumably excluded from the cytoplasm by yet unknown Ca^{2+} transport mechanisms. Early experiments showed that the presence of potassium or sodium in the medium—and their consequent influx in the cells—induces efflux of Ca^{2+} (Eilam, 1982b). Analogous results were obtained with low external pH, suggesting the presence of a Ca^{2+}/H^+ antiport located on the plasma membrane (Eilam, 1982a; Hong et al., 2013).

2.1.2 Vacuolar Transporters.

Ca^{2+} is massively stored in the vacuole by two active transporters: Pmc1, a high affinity low capacity Ca^{2+}-ATPase (Cunningham K. and Fink G., 1994), and Vcx1, a low affinity high capacity Ca^{2+}/H^+ exchanger. Vcx1 has a major role in shaping the calcium signal, since its high capacity can rapidly attenuate a large burst of cytosolic Ca^{2+} concentration (Miseta et al., 1999b). The free Ca^{2+} concentration in the budding yeast vacuole is estimated to be of about 30 μM, although larger amounts are stored as inorganic phosphates (Dunn et al., 1994). The vacuolar membrane of *S. cerevisiae* also contains Yvc1, a TRP-like calcium channel (Palmer et al., 2001; Denis and Cyert, 2002). Proper Ca^{2+} concentration in the endoplasmic reticulum (ER) organelle is critical to its functions; in *S. cerevisiae* it is maintained at 10 μM (Strayle et al., 1999), well below the concentration found in this compartment in higher eukaryotes, where it is the main internal storage for calcium ions (Stefan C. J. et al., 2013). Cls2/Csg2, an ER-localized protein, was originally proposed to play a role in Ca^{2+} efflux from the ER (Beeler et al., 1994; Tanida et al., 1996). However, Csg2 is likely implicated in the mannosylation of the inositol-phosphoceramide (IPC), corroborating the evidences about the sphingolipid roles in regulating ionic channels (Birchwood et al., 2001; Montefusco et al., 2014). Although initially underestimated, the central role of

the Golgi apparatus in calcium homeostasis and signaling is now well appreciated in eukaryotic cells (Pizzo et al., 2011), also in yeast cells (Miseta et al., 1999a; Wuytack et al., 2003). Pmr1, the budding yeast Golgi P-type Ca^{2+}/Mn^{2+}-ATPase, was the first member of the secretory pathway Ca^{2+}-ATPase (SPCA) subfamily identified (Rudolph et al., 1989; Antebi and Fink, 1992; Sorin et al., 1997); its role is pivotal for the maintenance of proper Ca^{2+} levels in both the Golgi apparatus (Halachmi and Eilam, 1996; Miseta et al., 1999a), and the ER (Durr et al., 1998; Strayle et al., 1999). Early reports suggested that yeast mitochondria have little, if any, role in accumulating Ca^{2+} (Carafoli et al., 1970; Balcavage et al., 1973), and their Ca^{2+} levels appear to be comparable to those of cytosol (Jung et al., 2004; Niedzwiecka et al., 2018).

2.1.3 The Calmodulin-Calcineurin Pathway

Calmodulin (CaM) is a highly conserved and ubiquitous Ca^{2+}-binding protein that modulates the activity of many target enzymes, mainly in response to increasing intracellular Ca^{2+} concentrations (Cyert, 2001), which trigger distinct structural rearrangements, and modes of target activation (Nakashima et al., 2012; Ogura et al., 2012b; Ogura et al., 2012a; Ishida et al., 2002). The number of identified target proteins for mammalian CaM is, according to the Calmodulin Target Database, nearly 300 (Yap et al., 2000), whereas far fewer are known for the yeast CaM (Cyert, 2001). During stress responses, the yeast CaM functions primarily through the activation of a small fraction of its targets: the calmodulin-dependent protein kinases (encoded by *CMK1*, *CMK2*), and calcineurin. Calcineurin is a Ca^{2+}/calmodulin-dependent serine/threonine-specific protein phosphatase, and represents the major Ca^{2+} signaling effector (Cyert and Thorner, 1992; Groppi et al., 2011; Li et al., 2011). Calcineurin regulates Ca^{2+} homeostasis and signaling both at the transcriptional level, for example by regulating the transcription of the Pmc1 encoding gene, and via the transcription factor Crz1 (Yoshimoto et al., 2002; Cyert, 2003), and at the post-translational level, for example by direct dephosphorylation of Vcx1. Some calcineurin targets involved in Ca^{2+} homeostasis are the HACS and the vacuolar Ca^{2+}/H^{+} exchanger Vcx1 (Cunningham and Fink, 1996; Miseta et al., 1999b; Kingsbury and Cunningham, 2000).

2.2 Hypotonic Shock Response

Upon HTS, after the transient osmotic swelling, mammalian cells re-adjust their volume by a mechanism known as regulatory volume decrease (RVD) (Okada et al., 2001). In some cell types, RVD is accomplished by means of stretch-activated Ca^{2+} channels which mediate a rapid increase of the cytosolic Ca^{2+} concentration due to both Ca^{2+} influx and Ca^{2+} release from intracellular stores, which in turn is often the signal that triggers release of osmolytes, reducing osmotic gradients and helping volume regulation (Jakab et al., 2002). In addition, the HTS-induced activation of ion conducting pathways leads to profound changes in the plasma membrane potential, which determines the direction of the ion fluxes depending on the respective equilibrium potentials and the modulation of voltage-gated ionic channels.

Upon HTS, a large fraction of glycerol, an intracellular osmolyte, is released in yeast cells to the extracellular environment within 2–3 min through the activation of Fps1 channels (Tamas et al., 1999), which is inhibited neither by gadolinium, which instead completely blocks Ca^{2+} increase (Batiza et al., 1996), nor by membrane potential alterations (Kayingo et al., 2001). Fps1 was proposed as a mechanosensitive channel directly activated by the induced membrane stretch, whereas the known post-translational modifications probably fine-tune its activity under basal conditions (Ahmadpour et al., 2014).

Phosphoinositides are negatively charged membrane lipids that serve as versatile molecules involved in protein regulation, assembly of actin cytoskeleton, vesicle trafficking and Ca^{2+} signaling. Several phosphoinositides are substrates for phospholipases, thereby generating a number of products that serve as second messenger with biological functions on their own (Strahl and Thorner, 2007). For example, the plasma membrane lipid phosphatidylinositol 4,5-bisphosphate (PtdIns (4,5) P_2) can be depleted by activation of phospholipase C (PLC), producing diacylglycerol (DAG) and the diffusible molecule Ins (1,4,5) P_3 (IP_3), which in mammals is a fundamental signaling molecule. In budding yeast, as in all eukaryotic cells, different phosphoinositide species are generated in a compartment-specific manner and, hence, can be regarded as distinct markers for each organelle (Odorizzi et al., 2000; Strahl and Thorner, 2007; Balla, 2013) and contribute to specific regulation of protein activity, such as ion channels (Hille et al., 2015). Upon HTS, *S. cerevisiae* cells hydrolyze plasma membrane PtdIns (4,5) P_2 with a mechanism dependent on Plc1 but independent on the extracellular Ca^{2+} concentration. This process liberates IP_3, which is then rapidly phosphorylated in IP_6. In addition, another phosphoinositide, PtdIns4P, and is rapidly synthesized and then progressively consumed in the next minutes (Perera et al., 2004). Although a link between these dynamics and Ca^{2+} signaling has not been explored yet, it is tempting to suggest one. In fact, the rapid depletion of the plasma membrane signature lipid PtdIns-(4,5)-P_2 could influence the activity of some channels. Flc2 is the best candidate, since it resides on plasma membrane and contains a putative lipid-binding domain. Channels localized on Golgi or ER could also be influenced by changes in phosphoinositides abundances: PtdIns4P transient increase could regulate Golgi channels, since PtdIns4P is the signature lipid of this organelle. Some Ca^{2+} regulation seems to be at stake in HTS because *PLC1* gene deletion, which abolishes PtdIns-(4,5)-P_2 depletion and IP_6 production, causes a greater increase in calcium influx after HTS compared to a wild type strain (Tisi et al., 2002). It is worth noting, however, that *PLC1* deletion does not affect PtdIns4P dynamics upon HTS (Perera et al., 2004).

The immediate and transient (~2 min) cytosolic Ca^{2+} pulse (Batiza et al., 1996; Rigamonti et al., 2015) triggered by HTS in *S. cerevisiae* cells is generated both by influx from the extracellular medium and efflux from intracellular stores. An early study on yeast cells grown in synthetic medium reported that this increase was mediated by an instantaneous release of Ca^{2+} from intracellular stores and then sustained by influx of

extracellular calcium. In fact, addition of an extracellular Ca^{2+} chelator, BAPTA, affected later stages of the response without affecting the initial, and rapid cytosolic Ca^{2+} rise (Batiza et al., 1996). Moreover, the same study showed that the HTS-induced Ca^{2+} response was dependent on both intensity of the shock and type of growth medium, the latter affecting also the pre-stimulus baseline $[Ca^{2+}]_{cyt}$. The increase of Ca^{2+} was inhibited in a dose-dependent manner by pre-treatment with gadolinium, a blocker of stretch-activated channels, and suggesting that the membrane stretching that occurs following HTS-induced cell swelling is directly sensed by Ca^{2+} channels (Batiza et al., 1996). In yeast cells grown in YPD—a complex, nutrient-rich medium—and challenged with HTS by diluting the medium with distilled water, an estimate of the initial rate of calcium increase at micromolar $[Ca^{2+}]_{ext}$ could be fitted by a Hill function, suggesting that the calcium increase in response to HTS was caused by the activation of a single channel or transporter located on the plasma membrane (Rigamonti et al., 2015).

The HTS response was also measured for mutants lacking proteins known to be involved in calcium signaling and homeostasis. *cch1*Δ mutants, lacking a functional HACS channel on the plasma membrane, responded to HTS with a higher calcium peak at all $[Ca^{2+}]_{ext}$ considered, suggesting a negative regulatory role for this protein during HTS. Mutants in other known influx pathways, on the other hand, and had a response similar to the wild-type. Calcium levels are affected in calcineurin mutants, suggesting that this Ca^{2+}-dependent effector shapes the calcium signal during HTS through dephosphorylation of target transporters and/or by modulating their long-term expression. In addition to display altered resting calcium levels, mutants lacking calcineurin respond to HTS with a dramatically reduced peak. Flc2 was found to be involved in the HTS-induced calcium response, since deletion of *FLC2* increases the initial rate of the Ca^{2+} increase compared with wild-type, suggesting an inhibitory role of this protein on the channel that is activated by the HTS. Based on the experimental evidences described above, a model is proposed that includes only the essential players in the HTS-induced calcium response. In non signaling conditions, Ca^{2+} enters the cell through HACS channel and other unidentified influx pathways. A still unidentified mechanosensitive calcium channel is located on the plasma membrane and activated by the increased membrane tension caused by HTS. In addition, this channel appears to be negatively regulated by Flc2. Since the elevation of $[Ca^{2+}]_{cyt}$ is transient, some intracellular transporters must restore the steady-state levels of cytosolic Ca^{2+}. This signal attenuation is probably performed by the Golgi-localized Pmr1, together with vacuolar-localized Pmc1 and Vcx1.

3 METHODS

3.1 Model Definition and Simulation

The mathematical model of HTS response in *S. cerevisiae* was defined on the basis of available experimental evidences (Rigamonti et al., 2015), and can be conceptually divided in two modules:

1 the *biophysical module* describes the changes in volume and other cell parameters, such as turgor pressure and membrane tension. This module allows for quantitatively following all changes in the physical state of the cell depending on cytosolic and extracellular osmolarities. Such parameters, in turn, regulate the activity of some components of the biochemical module. In particular, stretch-activated channels open following a sudden increase of membrane tension, promoting Ca^{2+} diffusion through them;
2 the *biochemical module* describes all the relevant reactions that take place in the cell—or between the cell and the extracellular environment—during the HTS response. This module comprises two compartments: the cytosol and the extracellular environment.

The model was formalized as a system of coupled ODEs. Mass-action kinetics was used to model the physical interactions between Ca^{2+}, calmodulin, and calcineurin. Most transport reactions were modeled by means of the Michaelis-Menten kinetics, as substantiated by previous studies (Ohsumi and Anraku, 1983; Wei et al., 1999; Takita et al., 2001; Teng et al., 2008).

Stretch-activated channels were modeled differently: they can be viewed as pores, whose opening probability depends on membrane tension. In particular, the opening probability of mechanosensitive channels was shown to follow a Boltzmann distribution (Gustin et al., 1988; Sukharev et al., 1999; Jiang and Sun, 2013). For the sake of simplicity, in this work the Boltzmann equation employs turgor pressure (note that membrane tensions can be calculated from turgor pressure using Laplace's law for a thin-walled sphere (Gustin et al., 1988; Sackin, 1995)). The opening probability (P_{open}) of mechanosensitive channels was thus formally defined as:

$$P_{open} = 1 - \frac{1}{1 + e^{\frac{P - P_{MS}}{g_{MS}}}} \quad (1)$$

where P is turgor pressure, P_{MS} is the value of turgor pressure at which P_{open} is equal to 0.5, and g_{MS} is a slope parameter.

Since ions pass through the channel pore down their electrochemical gradient, we assume that calcium ions flow according to their concentration gradient. The calcium flux j is then given by:

$$j = P_{open} \cdot k \cdot \Delta c \quad (2)$$

where k is a rate parameter and Δc is the Ca^{2+} concentration gradient across the membrane.

The model was simulated with COPASI (version: 4.19) (Hoops et al., 2006), using the LSODA algorithm (Petzold, 1983) with default settings. Simulation outputs consist in time traces of species concentrations over a period of 160 s, in line with the stress response duration of yeast cells. The model is available as an SBML Level 2 Version 5 file (ID MODEL2112030001) in the BioModels repository (Malik-Sheriff et al., 2020): https://www.ebi.ac.uk/biomodels/MODEL2112030001.

3.2 Parameter Estimation

Unknown parameters were estimated with COPASI (Hoops et al., 2006), using the available implementation of the Particle Swarm Optimization (PSO) algorithm (Kennedy and Eberhart, 1995).

PSO is a global optimization algorithm based on the concept of "swarm intelligence", which was shown to be effective in solving optimization problems characterized by multi-modal and noisy fitness landscapes, such as the ones related to the parameter estimation problem of biochemical systems (Nobile et al., 2018; Tangherloni et al., 2019; Besozzi et al., 2020).

To estimate the unknown parameters, simulations were fitted against available experimental time traces of $[Ca_{cyt}]$, measured in different experimental conditions in Rigamonti et al. (2015), including deletion mutants in key proteins of the HTS response and a wide range of extracellular Ca^{2+} concentrations. It is worth mentioning that, in the parameter estimation process, we took into account the well known fact that the intracellular calcium levels are kept in a narrow range despite wide variations in external conditions, thanks to calcium buffering and sensing and feedback mechanisms (Cunningham K. W. and Fink G. R., 1994). Thus, while reaction constants must be the same across different experiments, the concentrations of proteins bound to Ca^{2+} can vary. According to this line of reasoning, these parameters were not forced to be the same for all experiments. Since the model describes a stimulus response, the steady-state pre-stress conditions were also included in the parameter estimation process. Search ranges for all unknown parameters were set to be within biologically plausible numeric intervals.

4 MODEL DEFINITION

4.1 Definition of the Biophysical Module

The biophysical module describes the variation of the physical parameters of the cell, such as the cell volume and turgor pressure. The following mathematical description of volume regulation under osmotic stress is a simplified version of a previously published model, which was carefully parameterized using hyperosmotic shock data (Schaber and Klipp, 2008).

4.1.1 Volume

Assuming that volume changes are only due to water flow and not to solute flow, the total cell volume V (in L) can be defined as the sum of an osmotically active volume (water volume) V_{os} and an osmotically inactive volume (solid volume) V_b:

$$V = V_{os} + V_b, \tag{3}$$

with V_b assumed to be constant. The extracellular volume, V_{ex}, is much greater than the volume occupied by the cell. Therefore, for the sake of simplicity, we assume that $V_{ex} = V(0) \cdot 1{,}000$, where $V(0)$ is the initial cell volume.

The water flow is driven by gradients of water potential and hydrodynamic potential (Griffin, 1981; Kleinhans, 1998), which can be formalized as:

$$\frac{dV}{dt} = \frac{d}{dt}(V_{os} + V_b) = \frac{dV_{os}}{dt} = -L_p A(P + \Delta\Pi_n - \sigma\Delta\Pi_s), \tag{4}$$

where L_p is the hydraulic conductivity (in $dm \cdot MPa^{-1} \cdot s^{-1}$), A is the cell surface area (in dm^2) and P is the intracellular hydrostatic pressure exerted on the cell wall—i.e., the turgor (in MPa)—which equilibrates $\Delta\Pi$ under steady-state conditions ($\frac{dV}{dt} = 0$). $\Delta\Pi$ is the osmotic pressure difference (in MPa) between the outside and the inside of the cell (subscripts n and s denote non-permeable and permeable solutes, respectively), while the dimensionless parameter σ is the reflection coefficient, which depends on the solute permeability (Kedem and Katchalsky, 1958). In *S. cerevisiae*, the main permeable solute is glycerol (Reed et al., 1987), which is released by Fps1 channels (Luyten et al., 1995). When water and solutes are transported by different channels, the reflection coefficient has been shown to be:

$$\sigma = 1 - \frac{k_s \bar{V}}{RTL_p}, \tag{5}$$

where $\bar{V}$ is the partial molar volume (in $m^3 \cdot mol^{-1}$) of the solute, k_s is the membrane solute permeability (in $dm \cdot s^{-1}$), R is the gas constant (in $J \cdot mol^{-1} \cdot K^{-1}$) and T is the temperature (in K) (Kedem and Katchalsky, 1958; Kleinhans, 1998). Since the only solute considered here is glycerol, σ is approximately equal to 1 at room temperature[1], and thus:

$$\frac{dV_{os}}{dt} = -L_p A(P + \Delta\Pi_n - \Delta\Pi_s), \tag{6}$$

with $A \coloneqq A(t) = (36\pi)^{\frac{1}{3}} V(t)^{\frac{2}{3}}$, since the cell has a roughly spherical shape.

The van't Hoff law can be used to express the osmotic pressure in terms of concentration of osmotically active molecules: $\Delta\Pi = c_{PC} RT \Delta c$, where Δc is the concentration and c_{PC} is a conversion factor relating concentrations in M to pressures in MPa. Thus, **Eq. 6** can be written as:

$$\frac{dV_{os}}{dt} = -L_p A(P + c_{PC} RT([Osm_e] - [Osm_i])), \tag{7}$$

where Δc is substituted with $[Osm_e] - [Osm_i]$, where $[Osm_e]$ and $[Osm_i]$ are extracellular and intracellular concentration (in $\mu mol \cdot L^{-1}$) of osmotically active molecules, respectively. If **Eq. 7** is initially at steady-state ($\frac{dV_{os}}{dt} = 0$), the internal osmolarity can be estimated as a function of the initial turgor pressure P_0 and the initial external osmolarity Osm_e (0) as:

$$[Osm_i](0) = [Osm_e](0) + \frac{P_0}{c_{PC} RT}. \tag{8}$$

Both intracellular and extracellular osmolarities are the sum of concentrations of permeable solutes (glycerol) and non-permeable solutes. The extracellular osmolarity is thus:

$$[Osm_e] = c_n^e + [Gly_e] - [Gly_e](0), \tag{9}$$

where c_n^e is the extracellular concentration of non-permeable solutes, which varies in time according to the applied stimulus (explained below), while $[Gly_e]$ is the extracellular glycerol concentration. The initial extracellular glycerol concentration ($[Gly_e](0)$) is assumed to be 1,000 times lower than the initial intracellular glycerol concentration (that is, $[Gly_e](0) = [Gly_i](0)/1{,}000$), providing a gradient for glycerol efflux from the cell. All concentrations are expressed in $\mu mol \cdot L^{-1}$.

[1] $\bar{V}$ for glycerol is $0.071 \cdot 10^{-3}\ m^3/mol$

The intracellular osmolarity, accounting for cell volume variation, is defined as:

$$[Osm_i] = [Gly_i] + \frac{c_n^i V_{os}(0)}{V_{os}}, \tag{10}$$

where $[Gly_i]$ is the intracellular glycerol concentration, $c_n^i = [Osm_i](0) - [Gly_i](0)$ is the concentration of non-permeable solutes, while V_{os} and $V_{os}(0)$ are, respectively, the cytosolic volume and the cytosolic volume at $t = 0$. Concentrations are expressed in μmol·L^{-1}, while volumes are expressed in μm^3.

4.1.2 Turgor Pressure

The water potential gradient maintained by all cells across their membrane is balanced by a hydrostatic pressure called turgor. In walled cells, turgor pressure causes the cell membrane to exert a force on the cell wall, which expands due to its elasticity. The elastic-theory of turgor pressure states that the change in turgor pressure P is proportional to a relative change in cell volume:

$$dP = -\epsilon \frac{dV}{V}, \tag{11}$$

where the proportionality factor ϵ is called volumetric elastic modulus, or Young's modulus (in MPa).

The dependence of turgor pressure on volume can be deduced by integration:

$$\int_{V_0}^{V} dP(V) = -\int_{V_0}^{V} \epsilon \frac{1}{V} \;\Leftrightarrow\; P(V) - P(V_0) = -\epsilon \ln\left(\frac{V}{V_0}\right). \tag{12}$$

By defining V_0 as the volume when turgor becomes zero, the turgor pressure can be expressed as a function of volume:

$$\begin{cases} P(V, V_0, \epsilon) = -\epsilon \ln\left(\frac{V}{V_0}\right) & \text{for } V \geq V_0, \\ 0 & \text{for } V < V_0. \end{cases} \tag{13}$$

It is known that glycerol efflux and synthesis is tightly regulated according to environmental conditions (Talemi et al., 2016). However, given the small time-scales considered here, internal glycerol concentration is assumed to be constant.

4.1.3 Hypotonic Shock

Hypotonic shock is applied to the cell by diluting the medium with distilled water, thus decreasing its osmolarity. This dilution is modeled as follows:

$$\begin{cases} c_n^e = c_0^e & \text{for } t < t_{off}, \\ c_n^e = \left(c_0^e - \frac{c_0^e}{d}\right) e^{\frac{t_{off} - t}{t_m}} + \frac{c_0^e}{d} & \text{otherwise}, \end{cases} \tag{14}$$

where c_0^e is the initial extracellular osmolarity (in μmol·L^{-1}), d is the diluting factor, t is the simulation time instant (in s), t_{off} (in s) is the instant when dilution is applied, and t_m is the mixing time (in s) that regulates the speed of dilution.

4.2 Definition of the Biochemical Module

The concentration of calcium ions in the cytosol changes due to fluxes across different channels and transporters. The cytosol is also provided with calmodulin—a protein involved in the binding and sensing of calcium ions (Cyert, 2001)—that can activate calcineurin, the main calmodulin effector. HTS is applied by diluting the medium in which cells grow with distilled water, which has also the effect of reducing the availability of calcium ions. Since the extracellular volume is way larger than the volume occupied by all cells, the reduction of extracellular calcium ions caused by the cell uptake can be neglected. Therefore, the extracellular Ca^{2+} concentration depends only on dilution factor and mixing time (see **Eq. 14**):

$$\begin{cases} [\mathrm{Ca_{ex}}] = [\mathrm{Ca_{ex}}](0) & \text{for } t < t_{off}, \\ [\mathrm{Ca_{ex}}] = \left([\mathrm{Ca_{ex}}](0) - \frac{[\mathrm{Ca_{ex}}](0)}{d}\right) e^{\frac{t_{off} - t}{t_m}} + \frac{[\mathrm{Ca_{ex}}](0)}{d} & \text{otherwise}, \end{cases} \tag{15}$$

with all concentrations expressed in nM, and t, t_{off} and t_m expressed in seconds.

The rate of change of the Ca^{2+} concentration in the cytosol can be written as the sum of fluxes of the relevant channels and transporters (described below):

$$\frac{d[\mathrm{Ca_{cyt}}]}{dt} = j_{\mathrm{IN}} + j_{\mathrm{Cch1}} + j_{\mathrm{MS}} - j_{\mathrm{Pmr1}} - j_{\mathrm{Vcx1}} - j_{\mathrm{Pmc1}}, \tag{16}$$

where the js are Ca^{2+} fluxes (in nM·s^{-1}) across calcium channels and transporters.

Very often, channels and transporters show kinetics that can be described with the Michaelis-Menten equation (i.e. the transporter saturates at high substrate concentrations) (Christopher, 2002; Weijiu, 2012; Fridlyand et al., 2003). Indeed, this is the case for the intracellular transporters Pmr1, Vcx1, Pmc1 (see references in **Table 2**). Pmr1 indirectly replenishes the endoplasmic reticulum with Ca^{2+}, while Vcx1 and Pmc1 are responsible for its sequestration into the vacuole. Ca^{2+} enters the cell through the plasma membrane-located HACS channel and other transporters whose molecular identities are yet unknown (Batiza et al., 1996; Locke et al., 2000; Tisi et al., 2002; Cui et al., 2009a). Here, the influx associated with this unknown transport is simply called j_{IN}—to recall its function—and is assumed to have a Michaelis-Menten kinetics.

Experimental evidences strongly suggest that the increase of cytosolic Ca^{2+} in cells challenged with HTS is caused by the opening of a MS channel on the plasma membrane (Batiza et al., 1996; Rigamonti et al., 2015). This calcium influx pathway has not been molecularly identified yet, and it is here denoted by j_{MS}. The equations describing the fluxes are:

$$j_{\mathrm{IN}} = \frac{v_{\mathrm{IN}}\,[\mathrm{Ca_{ex}}]}{k_{\mathrm{IN}} + [\mathrm{Ca_{ex}}]}, \tag{17}$$

$$j_{\mathrm{Pmr1}} = \frac{v_{\mathrm{Pmr1}}\left[\mathrm{Ca_{cyt}}\right]}{k_{\mathrm{Pmr1}} + \left[\mathrm{Ca_{cyt}}\right]}, \tag{18}$$

$$j_{\mathrm{Vcx1}}^{0} = \frac{v_{\mathrm{Vcx1}}\left[\mathrm{Ca_{cyt}}\right]}{k_{\mathrm{Vcx1}} + \left[\mathrm{Ca_{cyt}}\right]}, \tag{19}$$

$$j_{\mathrm{Pmc1}} = \frac{v_{\mathrm{Pmc1}} \left[\mathrm{Ca}_{\mathrm{cyt}}\right]}{k_{\mathrm{Pmc1}} + \left[\mathrm{Ca}_{\mathrm{cyt}}\right]}, \tag{20}$$

$$j^0_{\mathrm{MS}} = P_{\mathrm{open}}\, k_{\mathrm{MS}} \left([\mathrm{Ca}_{\mathrm{ex}}] - \left[\mathrm{Ca}_{\mathrm{cyt}}\right] \right), \tag{21}$$

$$j^0_{\mathrm{Cch1}} = k_{\mathrm{Cch1}} \left([\mathrm{Ca}_{\mathrm{ex}}] - \left[\mathrm{Ca}_{\mathrm{cyt}}\right] \right), \tag{22}$$

where P_{open} is the opening probability of the MS channel (see below), k_{Cch1} and k_{MS} are rate parameters (in s^{-1}), *v*s. are rate constants (in $nM{\cdot}s^{-1}$), and all other *k*s are Michaelis constants (in nM).

The opening probability of MS channels follows a Boltzmann distribution (Gustin et al., 1988; Sukharev et al., 1999; Jiang and Sun, 2013). MS channels are gated by membrane tension (Gustin et al., 1988; Sackin, 1995) but here, for the sake of simplicity, turgor pressure is used instead (see **Section 3.1** and **Eq. 1** for a justification). The opening probability of the MS channel is then:

$$P_{\mathrm{open}} = 1 - \frac{1}{1 + e^{\frac{P - P_{\mathrm{MS}}}{g_{\mathrm{MS}}}}}, \tag{23}$$

where P is the cell turgor pressure in MPa (see **Eq. 13**), P_{MS} is the turgor pressure (in MPa) at which P_{open} is equal to 0.5, and g_{MS} is a slope parameter (in MPa).

4.2.1 Feedback Regulation

Inside the cytosol, yeast calmodulin binds three Ca^{2+} ions with high cooperativity (Davis et al., 1986; Nakashima et al., 1999):

$$\mathrm{CaM} + 3\mathrm{Ca}^{2+} \underset{k^-_m}{\overset{k^+_m}{\rightleftharpoons}} \mathrm{CaMCa}. \tag{24}$$

By using the law of mass action, the rate equation for Ca^{2+}-bound calmodulin can be formalized as:

$$\frac{d[\mathrm{CaMb}]}{dt} = k^+_m \left([\mathrm{CaMt}] - [\mathrm{CaMb}] \right) \left[\mathrm{Ca}_{\mathrm{cyt}}\right]^3 - k^-_m [\mathrm{CaMb}], \tag{25}$$

where [CaMb] is the concentration of Ca^{2+}-bound calmodulin (in nM), CaMt is the total calmodulin concentration (in nM), k^+_m and k^-_m are the forward and backward rate constants (in $nM^{-4}{\cdot}s^{-1}$ and s^{-1}, respectively).

Calcineurin, a protein phosphatase, is activated upon binding with the Ca^{2+}-bound calmodulin. By the law of mass action we can state that:

$$\frac{d[\mathrm{CaNb}]}{dt} = k^+_n \left([\mathrm{CaNt}] - [\mathrm{CaNb}] \right) [\mathrm{CaMb}] - k^-_n [\mathrm{CaNb}], \tag{26}$$

where [CaNb] is the concentration of calmodulin-bound calcineurin (in nM), CaNt is the total calcineurin concentration (in nM), k^+_n and k^-_n are the forward and backward rate constants (in $nM^{-2}{\cdot}s^{-1}$ and s^{-1}, respectively).

Following HTS, calcium ions enter the yeast cell through MS channels located on the plasma membrane. After the initial rise in Ca^{2+}, the signal dissipation observed in the successive seconds must be the result of either an increased activity of one of the intracellular transporters or a feedback inhibition on the MS channels. Since no positive regulation is known for any of the relevant transporters, in the present model the latter mechanism is assumed. This assumption is supported by circumstantial evidence suggesting that the yeast MS channel interacts with a homologous of TRP proteins (Rigamonti et al., 2015). TRP proteins form tetrameric ion channels which frequently interact with—and are inhibited by—calmodulin (Rhoads and Friedberg, 1997; Zhu, 2005). It is also known that activated calcineurin post-transcriptionally inhibits Vcx1 and HACS activity (Cunningham and Fink, 1996; Miseta et al., 1999b; Locke et al., 2000). The following equations are thus used to model the feedback inhibition:

$$j_{\mathrm{Cch1}} = \frac{j^0_{\mathrm{Cch1}}}{1 + k_{I\mathrm{Cch1}} [\mathrm{CaNb}]}, \tag{27}$$

$$j_{\mathrm{MS}} = \frac{j^0_{\mathrm{MS}}}{1 + k_{I\mathrm{MS}} [\mathrm{CaMb}]}, \tag{28}$$

$$j_{\mathrm{Vcx1}} = \frac{j^0_{\mathrm{Vcx1}}}{1 + k_{I\mathrm{Vcx1}} [\mathrm{CaNb}]}, \tag{29}$$

where $k_{I\mathrm{Cch1}}$, $k_{I\mathrm{MS}}$ and $k_{I\mathrm{Vcx1}}$ are inhibition constants (in nM^{-1}).

4.3 Estimation of the Unknown Parameters

Unknown parameters of the model were fitted against time traces of cytosolic Ca^{2+} measurements from HTS experiments conducted by Rigamonti et al. (2015) and by R. Tisi, unpublished results (**Figure 2**), as described in Materials and Methods. In these experiments, wild-type *S. cerevisiae* cells were grown in YPD medium, which has an estimated osmolarity of 0.26 Osm/L (Schaber et al., 2010). Specifically, the parameters of the biophysical module (**Table 1**) were set to reflect those experimental conditions.

In addition to the wild-type strain, we simulated three mutant strains: *cnb1*Δ, lacking functional calcineurin; *flc2*Δ, lacking a putative TRP-like channel subunit; and *cch1*Δ, lacking a functional HACS channel. Within the model, a mutant can be simulated by setting to zero the parameter associated with the function carried out by the removed gene. Hence, the *cnb1*Δ strain was obtained by setting to zero the total amount of calcineurin in the cell (CaNt), and the *cch1*Δ by setting to zero the rate parameter k_{Cch1}. However, these parameter settings were not sufficient to reproduce the behavior displayed by the three mutants. Thus, we performed additional parameter estimations exploiting available experimental time traces of the mutant strains (from Rigamonti et al. (2015) and R. Tisi, unpublished results, **Figure 2**).

In particular, *PMC1* expression strongly depends on calcineurin (Cunningham and Fink, 1996), but the model can not automatically adjust *PMC1* expression in the *cnb1*Δ mutant, since it lacks any description of transcriptional processes. To take into account the reduction of *PMC1* expression in this mutant, we performed a separate parameter estimation to infer the value of the v_{Pmc1} parameter (i.e., the $V_{\max}$ parameter of a Michaelis-Menten equation (**Eq. 20**), which is proportional to the enzyme abundance) in the *cnb1*Δ model. By doing so, it was possible to assess the reduction of Pmc1 transporters in this mutant with respect to the wild-type. We argue that the estimated value of v_{Pmc1} for the *cnb1*Δ mutant could be regarded as an estimate of the basal expression level of the Pmc1 protein, when Cnb1 is not stimulating the *PMC1* gene transcription. Thus, we used the same

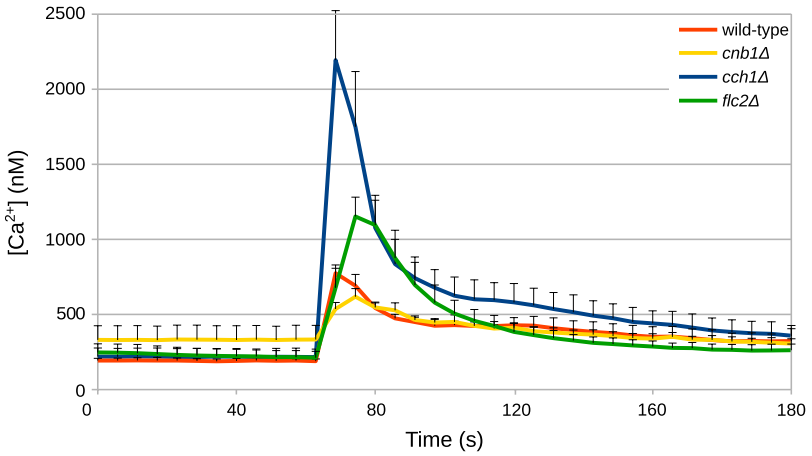


FIGURE 2 | Example of time traces of cytosolic Ca^{2+} concentration in *S. cerevisiae* cells challenged with HTS. HTS was applied by diluting the growth medium with four volumes of distilled water at t =60. The final concentration of Ca^{2+} in the medium was 5.9 μM. Time traces were taken from Rigamonti et al. (2015) and from Tisi R., unpublished results.

TABLE 1 | List of parameters of the biophysical module.

Parameters	Value	Unit	References
V_b	$20.5 \cdot 10^{-15}$	L	Schaber et al. (2010)
V (0)	$50 \cdot 10^{-15}$	L	
V_{ex}	1,000 · V (0)	L	
V_{os}(0)	V (0) – V_b	L	
L_p	$1.3 \cdot 10^{-7}$	dm·MPa^{-1}·s^{-1}	Talemi et al. (2016)
P_0	0.61	MPa	Schaber et al. (2010)
c_{PC}	10^{-9}	–	
R	8.314	J·mol^{-1}·K^{-1}	
T	303.15	K	
c_0^e	260,000	μmol · L^{-1}	Schaber et al. (2010)
D	variable	–	Rigamonti et al. (2015)
t_{off}	30	s	Rigamonti et al. (2015)
t_m	10	s	
[Gly_i]	180,000	μmol · L^{-1}	Schaber et al. (2010)
ϵ	14.3	MPa	

value of v_{Pmc1} to simulate the *cch1*Δ model. In fact, in order to maintain a physiological Ca^{2+} level when the Ca^{2+} income is lower, a novel steady state balance has to be achieved, where Pmc1 abundance has to be adjusted so that Ca^{2+} remains available to be provided to the secretory pathway as well. This is achieved by finely tuning the homeostasis regulatory circuit, through calcineurin transcriptional control on *PMC1* gene, leading to a Pmc1 activity nearby the basal level estimated for the *cnb1*Δ mutant.

Adopting a similar approach to the one described above, we estimated the value of k_{MS} for the *flc2*Δ strain, in order to account for the observed increased activity of the MS channel in this strain (Rigamonti et al., 2015). It is not known whether the increased activity of this channel is due to an increased channel abundance or simply to an increased channel activity. In any case, the rate parameter k_{MS} is—like the $V_{\max}$ of a Michaelis-Menten equation—proportional to the number of channels.

Analysis of the parameter estimation results led to the conclusion that the model could be simplified without affecting the simulations. In particular, removal of (*a*) the feedback inhibition on the HACS channel, mediated by calcineurin (**Eq. 27**), and of (*b*) the influx pathway that was called "IN" (**Eq. 17**), did not cause any significant difference on the simulation outcomes. Therefore, the feedback on HACS and the j_{IN} influx were removed and all simulations and parameter values reported here are related to this simplified model. **Table 2** lists all the parameters used for the wild-type model, while **Table 3** contains only the values that changed depending on the strain.

5 RESULTS

5.1 Model Simulations

The simulations of the full model of HTS response correctly reproduce the dynamics of the Ca^{2+} transients (**Figures 3A–D**), as well as the steady-state levels and peak values of cytosolic Ca^{2+} (**Figures 3E,F**) of both wild-type and all mutant strains.

The results of the simulations for the wild-type strain are shown in **Figure 4**. According to the model, after the HTS the HACS activity decreases due to the sudden shortage of the extracellular Ca^{2+} (**Figure 4A**), while the MS channels on the plasma membrane open rapidly (**Figure 4B**, inset). The cytosolic Ca^{2+}-dependent inhibition on the MS channels provides a way to decrease the activity of this channel, thus helping to restore low intracellular calcium levels after the stimulus. In fact, the dynamics reported in **Figure 4B** shows that only in the presence of feedback inhibition the flux through this channel is still decreasing at time t > 280s. These outcomes also suggest that the main pump responsible for signal attenuation is Pmc1, the Ca^{2+}-ATPase located on the vacuole, while Pmr1 and Vcx1 have a negligible role in this respect (**Figures 4C,D**). In particular, while the predicted rate of

TABLE 2 | List of parameters of the wild-type calcium model.

Parameters	Value	Range	Unit	References
$[Ca_{cyt}](0)$	100		nM	Cunningham and Fink, (1994a)
$[Ca_{ex}](0)$	29,500		nM	Rigamonti et al. (2015)
t_m	9.1	1—10	s	
v_{Vcx1}	2,820,420	$100—5 \cdot 10^6$	$nM \cdot s^{-1}$	
v_{Pmc1}	280,870	$100—5 \cdot 10^5$	$nM \cdot s^{-1}$	
v_{Pmr1}	813	$100—5 \cdot 10^5$	$nM \cdot s^{-1}$	
k_{MS}	132,184	$0.1—10^6$	s^{-1}	
k_{Cch1}	0.37	$0.1—10^6$	s^{-1}	
k_{Pmr1}	70		nM	(Sorin et al., 1997; Wei et al., 1999)
k_{Vcx1}	100,000		nM	Ohsumi and Anraku, (1983)
k_{Pmc1}	4,300		nM	Takita et al. (2001)
P_{MS}	0.76	0.6—0.8	MPa	
g_{MS}	0.039	0.00001—1	MPa	
k_m^+	$1.8 \cdot 10^{-14}$	$10^{-20}—0.01$	$nM^{-4} \cdot s^{-1}$	
k_m^-	$k_m^+ \cdot 9{,}000$		s^{-1}	Starovasnik et al. (1993)
k_n^+	0.1	$10^{-8}—10$	$nM^{-2} \cdot s^{-1}$	
k_n^-	1,000	$10^{-4}—10^4$	s^{-1}	
[CaMt]	2,600		nM	Ho et al. (2018)
[CaNt]	310		nM	Ho et al. (2018)
k_{fMS}	26,395	$100—10^8$	nM^{-1}	
k_{fVcx1}	9,348,540	$100—10^8$	nM^{-1}	

Known parameters were fixed, while parameters that needed to be estimated were allowed to change inside the values shown in the "Range" column. Simulations of the mutant strains were done with the parameters listed here, except for those listed in ***Table 3****.*

TABLE 3 | List of parameters that change depending on the modeled mutant strain. All other parameters were kept as in the wild-type model (**Table 2**).

	v_{Pmc1} ($nM \cdot s^{-1}$)	k_{Cch1} (s^{-1})	k_{MS} (s^{-1})	[CaNt] (nM)
wild-type	280,870	0.37	132,184	310
*flc2*Δ	280,870	0.37	238,519	310
*cch1*Δ	41,079	0	132,184	310
*cnb1*Δ	41,079	0.37	132,184	0

Ca^{2+} sequestration by Pmr1 is predicted to be consistently slow, the Vcx1 activity is kept down by the feedback inhibition mediated by calcineurin (**Figure 4D**).

As described above, the peak response of *cch1*Δ mutants—lacking functional HACS—is considerably higher than in the wild-type strain (**Figure 3F**). According to the model, the peak difference is not ascribable to an increased activity of the MS channels in this strain with respect to the wild-type. In fact, in this mutant, Ca^{2+} fluxes through both MS channels and Pmc1 are considerably lower than both wild-type and *flc2*Δ strains (**Figures 5A,C**). This behaviour is the result of the reduced *PMC1* expression (**Table 3**):

Since Pmc1 is the main transporter responsible for Ca^{2+} sequestration from the cytosol, reducing its activity increases the level of Ca^{2+}, which in turn inhibits the MS channels through the activated calmodulin (**Figure 5B**). This feedback process ensures that the steady-state calcium levels are comparable to those of the wild-type. However, the reduced activity of Pmc1 causes a higher peak when the MS channels open following HTS. Rigamonti *et al.* suggested that the increased Ca^{2+} peak in *flc2*Δ is caused by the removal of an inhibitory effect on the MS channels (Rigamonti et al., 2015). To test this hypothesis, during the parameter estimation process we let the k_{MS} parameter of *flc2*Δ mutant assume different values compared to other strains (**Table 3**). The simulations show that, in line with this, a difference in the MS channels rate compared to the wild-type is sufficient to explain the increased peak observed during HTS (**Figure 5A**).

The *cnb1*Δ mutants show an increased steady-state level of cytosolic Ca^{2+} and a lower peak compared to all other strains (**Figure 3**). As described in **Section 4.3**, for this mutant cells we estimated the reduction of *PMC1* expression. Our model predicts that Pmc1 abundance is about 16 times lower compared to the wild-type, which is counterbalanced by the increased activity of Vcx1 that is no longer inhibited by calcineurin. This compensation mechanism, however, fails to fully replace the decreased activity of Pmc1, and thus the mutant displays higher steady-state calcium levels. The higher calcium levels result in more activated calmodulin (**Figure 5B**), which in turn inhibits the MS channels leading to a lower peak (**Figure 5A**).

Figure 6 shows a prediction of HTS response in not yet tested experimental conditions, since the technical difficulty in achieving a tight control of the speed of dilution could not be overcome. Variability in the speed and shape of the signal was observed in different experiments, particularly when an automatic injector was not applied, thus we decided to investigate the effect of this parameter variation with *in silico* experiments. The t_m parameter of the model defines the mixing time, i.e. the speed of dilution. The simulations of the wild-type strain with different mixing times produces an unexpected pattern (**Figure 6A**). In particular, while the speed of the response turns out to be linearly dependent on the mixing time (**Figure 6B**, red curve), the peak Ca^{2+} value is not. The

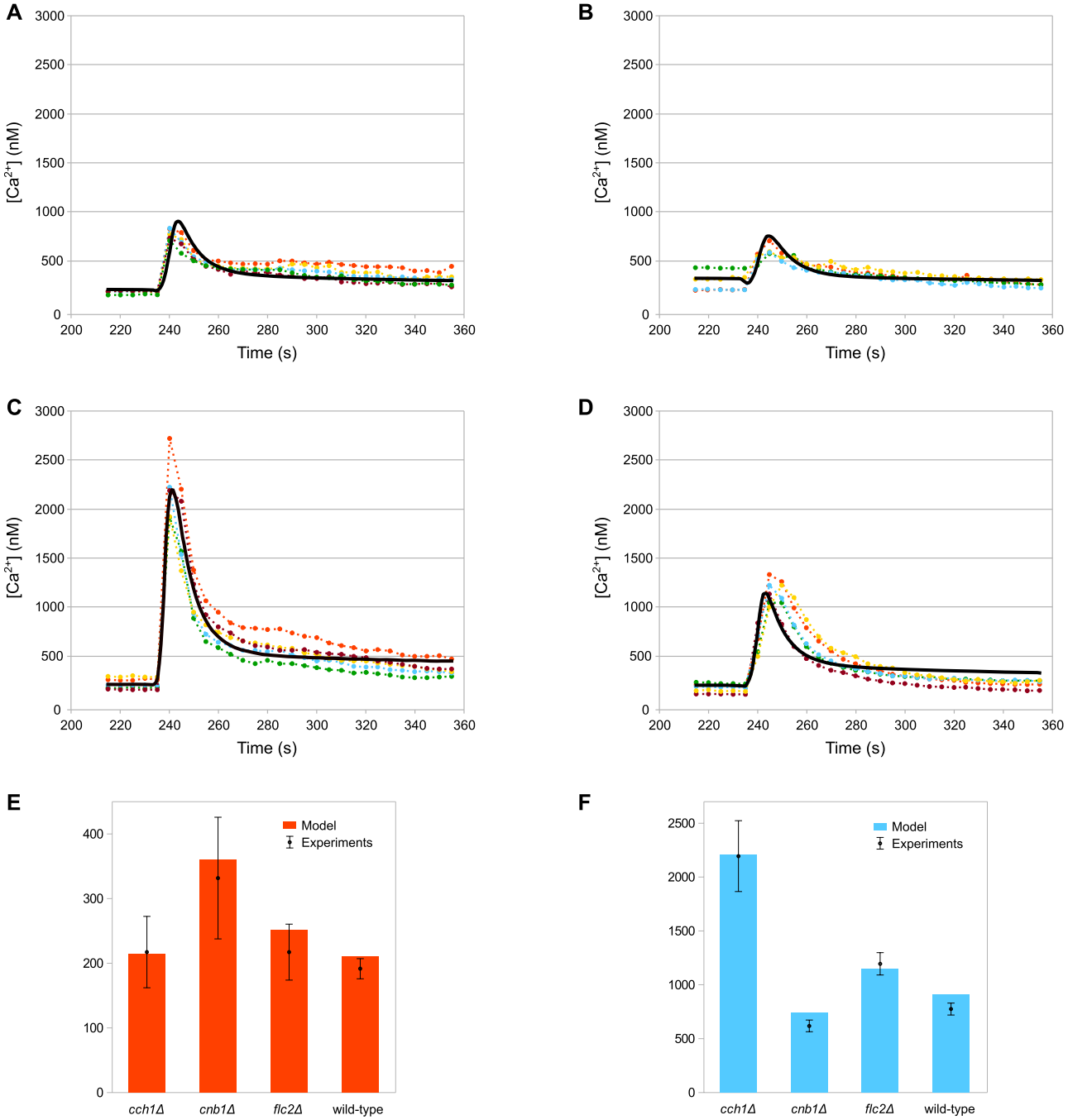


FIGURE 3 | Comparison of simulation outcomes and experimental measurements. In Figures **(A–D)** the dotted lines represent experimental measurements and different colours indicate replicates, while the black lines represent model simulations. **(A)** Wild-type. **(B)** *cnb1*Δ. **(C)** *cch1*Δ. **(D)** *flc2*Δ. **(E)** Baseline cytosolic Ca^{2+} concentration (nM). **(F)** Max peak cytosolic Ca^{2+} concentration (nM). HTS was applied to the cells at 235 s by adding four volumes of distilled water to the medium **(E,F)** Plots constructed using the same data shown in **(A–D)**; bars represent standard deviations.

peak values grow with the mixing time until a maximum at t_m = 15s, before decreasing again (**Figure 6B**, blue curve).

The model also predicts that the inactivation of the MS channels—simulated by setting to zero the parameter k_{MS}—would reduce the steady-state cytosolic Ca^{2+} concentration from 215 to 160 nM.

6 DISCUSSION

S. cerevisiae cells have to adapt to changes in growth conditions that arise both naturally, in the environment where they live, and artificially, during human exploitation. All cells sense extracellular osmolarity and fine-tune their biophysical

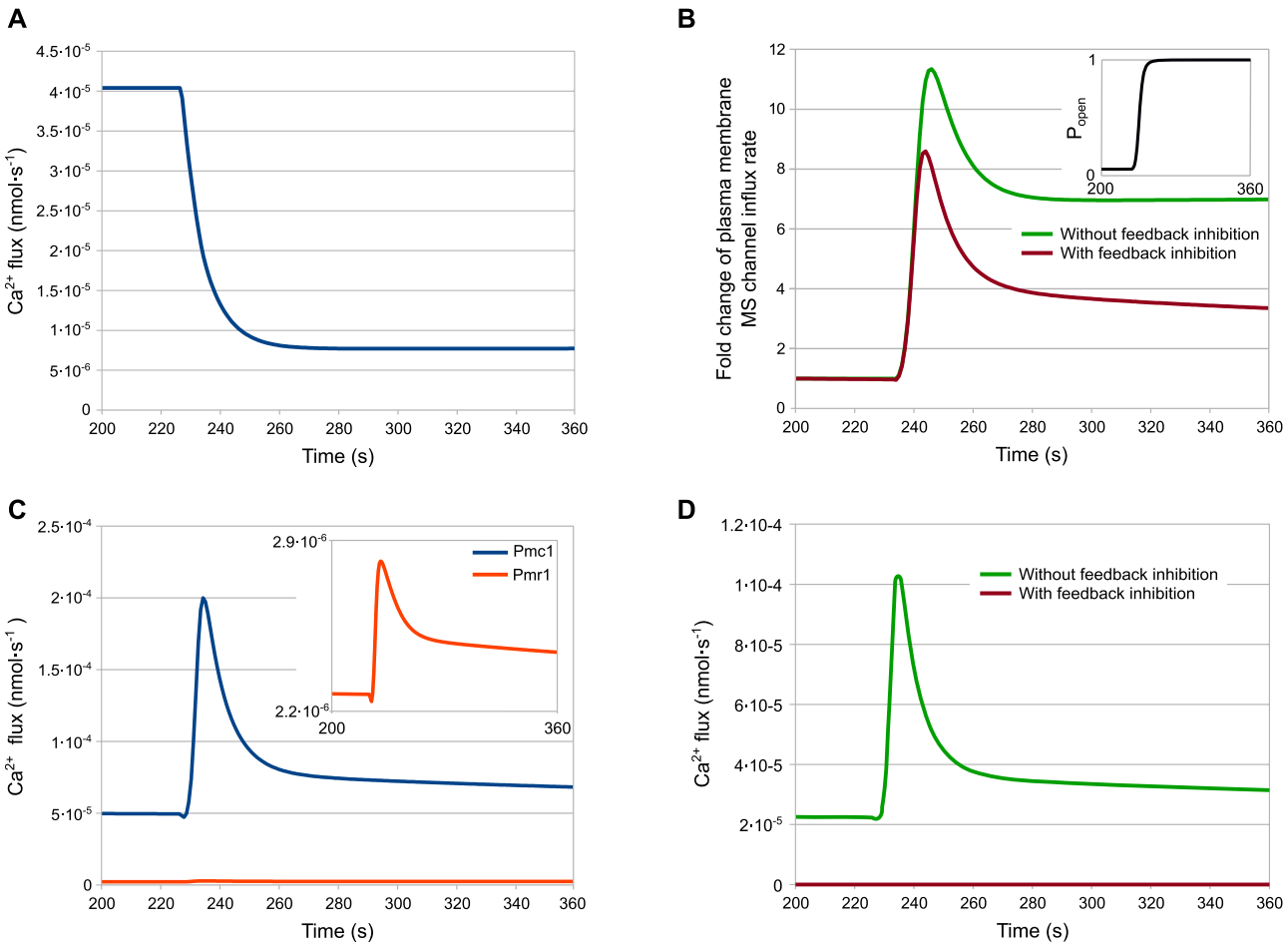


FIGURE 4 | Fluxes involved in the HTS response of the wild-type strain. **(A)** HACS flux. **(B)** Effect of calmodulin inhibition on MS channels. **(C)** Pmc1 and Pmr1 fluxes. **(D)** Effect of calcineurin inhibition on Vcx1 **(A–D)** Ca^{2+} fluxes through channels and transporters as predicted by the model (parameters are given in **Table 2**) **(B,D)** Fluxes with or without inhibition are depicted by plotting j_{MS} and j^{0}_{MS} for MS channel, and j_{Vcx1} and j^{0}_{Vcx1} for the Vcx1 transporter. Inset in **(B)** shows how the opening probability of MS channels changes with time (note that it is not equal to zero before the stimulus). Fluxes in **(B)** are normalised to their steady-state values to better appreciate the long-term effect of channel inhibition.

parameters to ensure survival (Pedersen et al., 2011). Generation of a Ca^{2+} transient in response to HTS seems to be an ubiquitous phenomenon shared by yeast, plant and mammalian cells, and it is caused by the opening of MS channels (Sachs and Morris, 1998; Cox et al., 2013). In the yeast *S. cerevisiae*, many proteins involved in Ca^{2+} handling have been identified (Cui et al., 2009b; Tisi et al., 2016). However, only recently the specific roles of each channel or transporter in building, attenuating and shaping the observed Ca^{2+} dynamics are being investigated. The model presented here can help in bridging the gap between genetic analyses and phenotypic observations regarding Ca^{2+} signaling in yeast cells.

Simulations of the biophysical module alone showed that the volume of yeast cells increases only by a few percent relative to the pre-stress volume. This is in contrast to non-walled cells, whose volume can increase up to 50% (Weskamp et al., 2000). This difference is not surprising, as yeast cells must face harsher and unpredictable osmotic conditions than single cells inside a multicellular organism, and have therefore evolved a rigid cell wall.

In *S. cerevisiae* cells the vacuole is the main organelle for Ca^{2+} storage (Dunn et al., 1994), while in mammalian cells this role is played by the ER (Montero et al., 1995; Strayle et al., 1999). In fact, our simulations shows that most of the Ca^{2+} is transported into the vacuole. In all modeled strains, with the exception of *cnb1*Δ, Pmc1 appears to be the main vacuolar transporter. As expected, removal of calcineurin have opposite effects on vacuolar transporters: it reduces the expression of *PMC1*, but it increases the activity of Vcx1. However, our simulations suggest that the loss of Pmc1 activity is not entirely counterbalanced by Vcx1, and thus the cytosolic calcium level is higher than in the wild-type. These results are in accordance with calcium accumulation measurements showing that *pmc1*Δ mutants accumulate only 20% calcium compared to wild-type (Cunningham and Fink, 1996). In addition, the predictions made by our model—namely, that Vcx1 has a small role in calcium

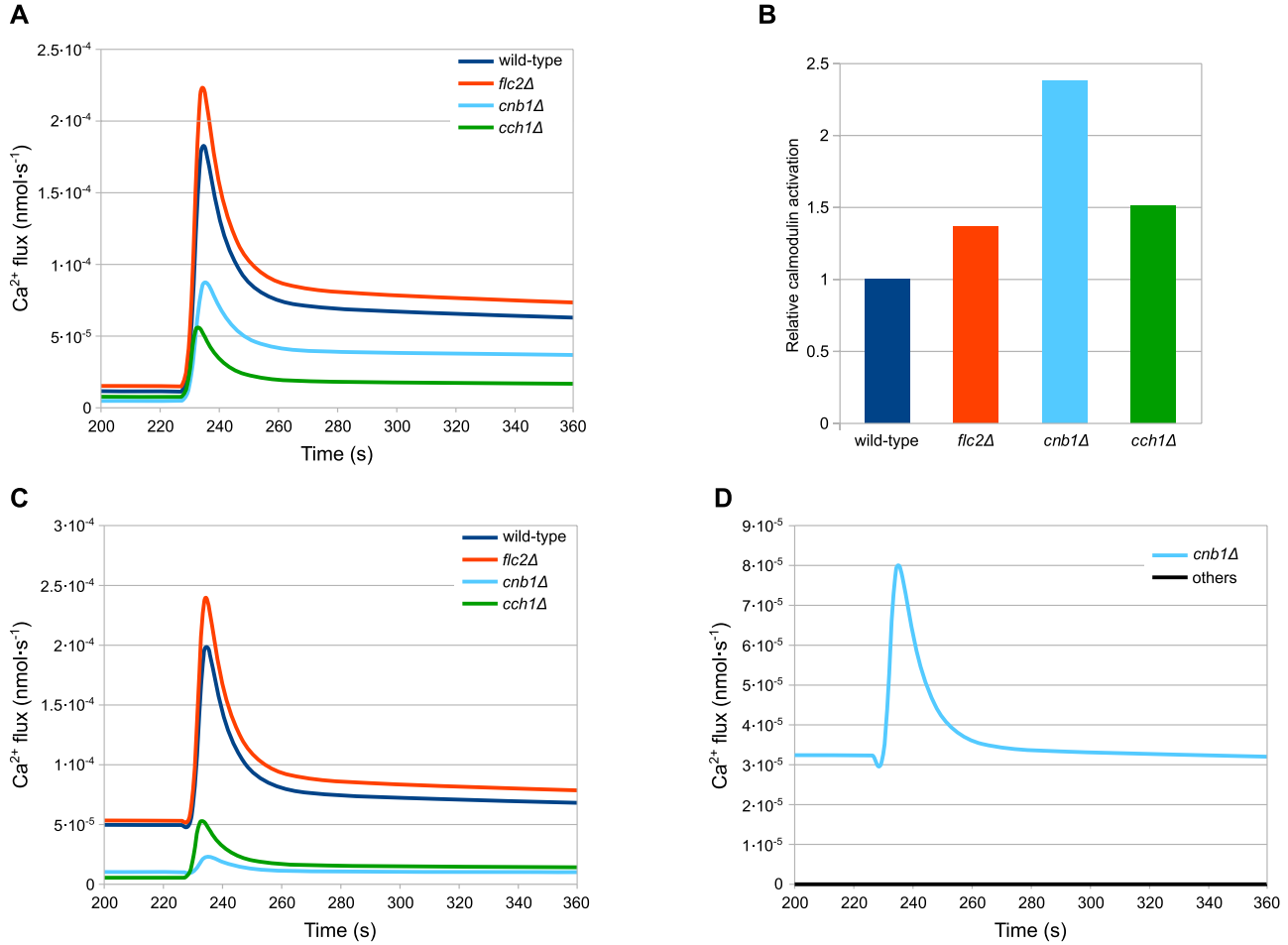


FIGURE 5 | Simulations showing the main differences among the considered strains. **(A)** MS influx. **(B)** Calmodulin activation. **(C)** Pmc1 flux. **(D)** Vcx1 flux **(A,C,D)** Fluxes through the most relevant channels. **(B)** Calmodulin activation, shown relative to the wild-type level, before HTS is applied. Time traces of calcineurin are similar and are not shown for the sake of figure readability. **(D)** Vcx1 fluxes of all mutants but *cnb1*Δ are overlapping and therefore are coloured black.

accumulation unless calcineurin is removed—are supported by the fact that *vcx1*Δ mutants accumulate the same amount of calcium as the wild-type, and that Vcx1 has a larger role in Ca^{2+} sequestration in *cnb1*Δ mutants (Cunningham and Fink, 1996). Pmr1 is important for the maintenance of proper calcium levels in the ER, where the steady-state free concentration is only 10 μM (Strayle et al., 1999). Our simulations suggest also that Pmr1 has a marginal role during the HTS response, as well as in maintaining low steady-state calcium level into the cytosol. However, no definitive comparison can be drawn between model predictions and experimental results because *pmr1*Δ mutants display pleiotropic defects that are probably direct consequences of ER calcium depletion (Rudolph et al., 1989; Antebi and Fink, 1992).

Yeast cells are equipped with a variety of calcium influx pathways, each with seemingly different functions (Eilam and Othman, 1990; Iida et al., 1994; Fischer et al., 1997; Muller et al., 2003). Our model initially included three calcium entry pathways: HACS (the high affinity calcium system composed of at least Cch1 and Mid1), MS channels, and another calcium influx of unknown molecular identity. The presence of MS channels on the yeast plasma membrane was demonstrated by patch-clamp experiments (Gustin et al., 1988), but their molecular identities are still unknown. Parameter estimation revealed that the additional influx (here called "IN") was unnecessary in this model, and that HACS, and the MS channels alone are sufficient to explain the observed data. The removal of "IN", which was the only active transporter on the plasma membrane introduced in the model, implies that in the simulations Ca^{2+} enters the cell only by passive transport, driven by the gradient in concentration and by the membrane potential. In fact, experiments demonstrated that in energy-depleted cells, where maintenance of membrane potential is defective, there was no Ca^{2+} influx (Eilam and Chernichovsky, 1987).

According to our simulations, even during steady-state conditions, and Ca^{2+} enters the cell through unstimulated MS channels. Indeed, most channels are known to be "leaky", i.e., they stochastically open even in non signaling conditions. Since the flux through a single channel can be several orders of magnitude higher than that of a single transporter (Milo et al.,

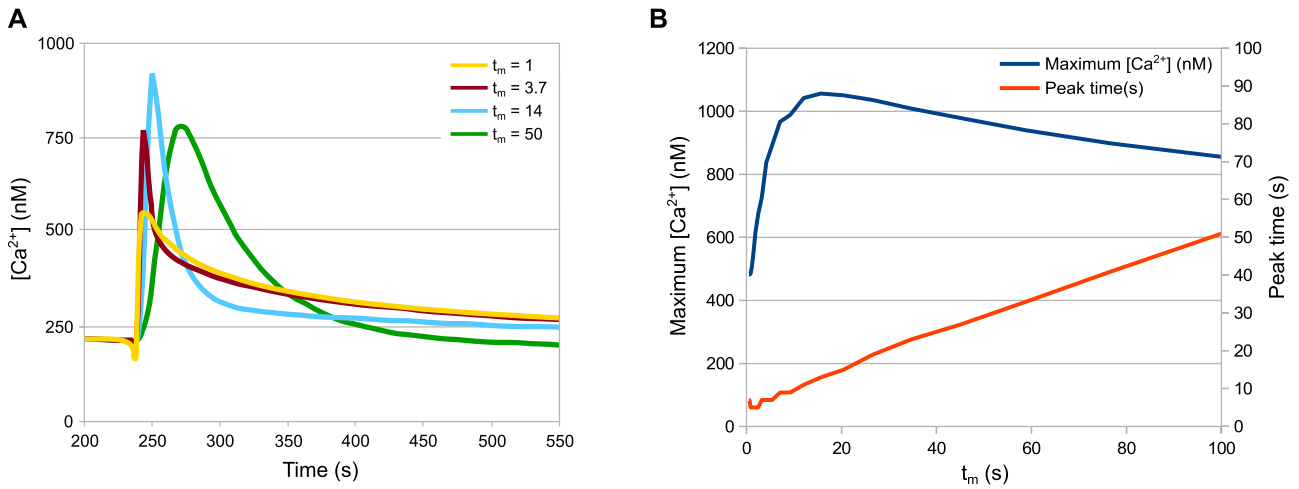


FIGURE 6 | Behaviour of the system depending on the speed of the HTS. **(A)** Ca^{2+} peaks in response to different dilution speeds. **(B)** Maximum Ca^{2+} values and time of maximum Ca^{2+} values depending on the dilution speed. The t_m parameter is inversely proportional to the speed of dilution. Increasing the speed of dilution increases the speed of response but not necessarily the peak height.

2010), even brief channel openings can theoretically supplement the cytosol with a significant amount of Ca^{2+}. It has been shown that treatment with gadolinium, a blocker of stretch-activated channels, eliminates the Ca^{2+} rise in response to HTS without significantly affecting the steady-state Ca^{2+} concentration (Batiza et al., 1996). Accordingly, removing MS channels from our model decreases steady-state calcium levels from 215 to 160 nM, a value which is still well within the physiological range (Cunningham K. W. and Fink G. R., 1994). During model definition, a feedback inhibition was introduced on MS channels, mediated by calmodulin. This is the only assumption made in the model that is not yet supported by strong experimental evidence. We suggest that some kind of calcium-dependent feedback must exist, which either increases the activity of intracellular transporters or decreases the activity of MS channels. Many ion channels are inhibited by calmodulin (Saimi and Kung, 2002) and performing a parameter estimation with the model lacking this feedback inhibition produced poor results, in particular for the *cnb1*Δ mutant.

HACS channel has been shown to be activated following a number of external stimuli, such as alkaline stress and mating pheromone (Iida et al., 1994; Viladevall et al., 2004), but it is also involved in calcium uptake during normal growth, as evidenced by long-term calcium accumulation studies (Muller et al., 2001). In addition, HACS seems to physically interact with, and be inhibited by, calcineurin (Muller et al., 2001; Bonilla and Cunningham, 2003). After the parameter estimation process, it emerged that in our model the equation describing this negative regulation was unnecessary to reproduce the experimental data. This result suggests that, in the growth conditions considered here, feedback inhibition by calcineurin does not change significantly, since calcineurin activity is low in the chosen cultural conditions (Groppi et al., 2011).

Flc2 belongs to a TRP-like fungal family of putative yeast calcium transporters, together with Yor365c, Flc1, and Flc3 (Tisi et al., 2016). This raises the possibility that also the MS channels, with which Flc2 seems to interact, might belong to the aforementioned family. Our simulations support the hypothesis that Flc2 exerts an inhibitory role on the MS channels located on the plasma membrane (Rigamonti et al., 2015). In fact, the observed increase of calcium response and calcineurin hyperactivation in the *flc2*Δ mutant (Rigamonti et al., 2015) could be reproduced with the model just by increasing the activity of the MS channels. Flc2 may directly or indirectly influence the MS channel by affecting its stability, function or localization.

Most of the model predictions are in accordance with experimental results that were not used for the parameter estimation process, suggesting that the model provides an accurate description of the HTS response in yeast cells as well as the most relevant transcriptional and post-transcriptional regulations involved in calcium handling. In particular, the model suggests the following regulatory scheme (**Figure 7**). Normally, Ca^{2+} enters the cell via HACS and MS channel leak. Pmc1 is the main intracellular transporter that keeps Ca^{2+} cytosolic level within the physiological range by pumping it into the vacuole. In the wild-type, Vcx1 is almost completely inhibited by calcineurin (**Figure 7A**). Mutants defective in HACS channel are still able to maintain physiological levels of cytosolic Ca^{2+} because calcineurin-dependent expression of Pmc1 is decreased. For the same reason, when this strain is challenged with HTS, the sudden Ca^{2+} influx is more slowly attenuated by Pmc1 (**Figure 7C**). In *flc2*Δ mutants, the flux through the MS channels is increased by removal of the inhibitory effect of Flc2 (**Figure 7B**). Lastly, mutants lacking calcineurin express less Pmc1, but the inhibitory effect on Vcx1 is relieved. Vcx1 only partially compensates for the reduced Pmc1 activity and the resulting higher Ca^{2+} levels activate calmodulin, that in turn inhibits MS channels (**Figure 7D**).

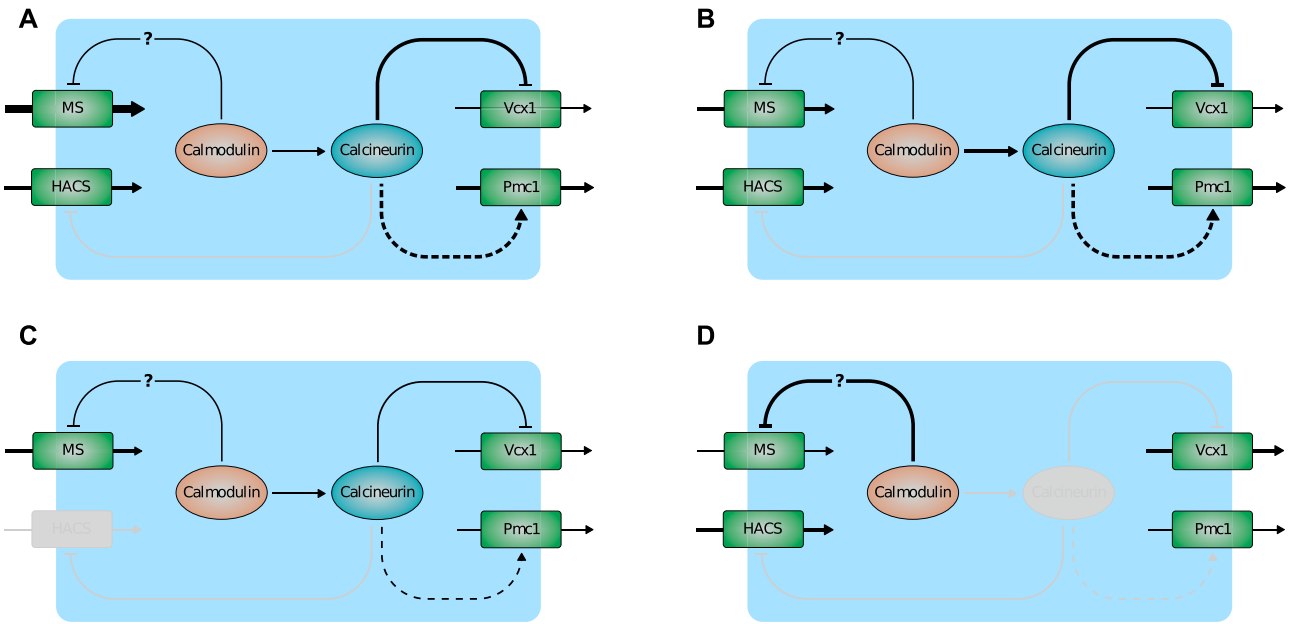


FIGURE 7 | Scheme depicting the main predictions of the model. **(A)** Wild-type. **(B)** *flc2*Δ. **(C)** *cch1*Δ. **(D)** *cnb1*Δ. Activation is indicated by arrows, inactivation by T-bar arrows. Plain and dashed lines indicate post-transcriptional and transcriptional regulations. Line thickness represents the strenght of fluxes/(in)activations. Inactive components of the pathway are greyed-out.

The function of the Ca^{2+} transient is still unclear. Calcineurin regulates gene expression by promoting Crz1 movement into the nucleus (Stathopoulos and Cyert, 1997; Cai et al., 2008). During HTS, Crz1 stays in the nucleus for about 5 mins (Oh et al., 2012), but its activity is only slightly increased (Rigamonti et al., 2015). In mammalian cells, calcium is often required for regulatory volume decrease (Jakab et al., 2002), while in yeast cells it could be implicated in the regulation of the cell wall integrity pathway (Levin, 2011; Rigamonti et al., 2015). In addition to the transient increase of Ca^{2+}, HTS rapidly stimulates the Plc1-dependent hydrolysis of phosphatidylinositol 4,5-bisphosphate (Perera et al., 2004), an event that in many cell types elicits complex calcium responses (Hille et al., 2015). However, yeast *plc1*Δ mutants are still able to generate the Ca^{2+} signal during HTS (Tisi et al., 2002) and, conversely, Plc1 activity is independent from calcium availability in the medium (Perera et al., 2004), suggesting that in yeast these events might be independent.

7 CONCLUSION

As extensively described, the simulations of our model are in agreement with the experimental data, suggesting that the cellular components included in the model and their interactions are sufficient to explain the biological data in our hands. The model was defined on the basis of experimental evidences from the literature, and the parameter estimation process was performed against data from both the wild-type and mutant cells, in order to maximise the prediction's reliability. The model presented in this work could be extended to include the response to several other stimuli known to elicit a calcium signal in yeast (Iida et al., 1990; Courchesne and Ozturk, 2003; Groppi et al., 2011). Such extensions of the model can be achieved either by modifying the current system of ODEs, or by adopting hybrid modeling approaches (Spolaor et al., 2019b; Nobile et al., 2020), in order to include other functional modules in the model that can better describe the different cellular processes involved in Ca^{2+} signalling (including, e.g., gene regulation and expression). Extended versions of the model could be useful for understanding the role of known channels and transporters, suggesting novel putative regulatory mechanisms and gaining new insights on Ca^{2+} signaling that can be further investigated experimentally.

Being Ca^{2+} an essential signaling ion, proteins involved in its homeostasis are increasingly studied as potential targets of antifungal drugs (Kwun et al., 2021; Li et al., 2021; Wang et al., 2021). A model of Ca^{2+} signaling could be used to predict the outcomes of the inhibition of a particular protein on the calcium physiology of fungal cells, thus helping to identify the best drug targets. Moreover, such models could be used to study *in silico* the relationships between pathogenic fungi and human cells (Spolaor et al., 2019a), accelerating the discovery of new antifungal treatments. Since Ca^{2+} signaling is well conserved among fungi (Tisi et al., 2016), a model of *S. cerevisiae* would require only slight modifications to describe what happens in other, pathogenic fungi.

Finally, as a future development, we plan to perform a large-scale sensitivity analysis of the model, in order to determine the most relevant parameters governing the response to hypotonic shock. Global sensitivity analysis requires a massive amount of

independent simulations, which can lead to a huge computational effort (Nobile and Mauri, 2017). In order to make this analysis feasible, we will re-implement the model using the Python library ginSODA (Nobile et al., 2019), which provides the possibility of offloading a massive number of simulations of an ODE-based model to the GPU, thus strongly reducing the overall running time.

DATA AVAILABILITY STATEMENT

The model presented in this study can be found in the BioModels repository, with accession number ID MODEL2112030001, at the following link https://www.ebi.ac.uk/biomodels/MODEL2112030001.

AUTHOR CONTRIBUTIONS

RT and DB conceived and designed the study. RT produced the experimental data. MR defined the model, performed the simulations and analysis. SS oversaw data analysis and curation, and model sharing. SS, MR, and RT wrote the first draft of the manuscript. MN, PC, and DB critically reviewed and edited the manuscript. RT and DB acquired funding. All authors reviewed and approved the manuscript.

FUNDING

This work was partially supported by the SYSBIO/ISBE. IT Research Centre of Systems Biology.

REFERENCES

Aguilar-Uscanga, B., and Francois, J. M. (2003). A Study of the Yeast Cell wall Composition and Structure in Response to Growth Conditions and Mode of Cultivation. *Lett. Appl. Microbiol.* 37, 268–274. doi:10.1046/j.1472-765x.2003.01394.x

Ahmadpour, D., Geijer, C., Tamás, M. J., Lindkvist-Petersson, K., and Hohmann, S. (2014). Yeast Reveals Unexpected Roles and Regulatory Features of Aquaporins and Aquaglyceroporins. *Biochim. Biophys. Acta (BBA) - Gen. Subjects* 1840, 1482–1491. doi:10.1016/j.bbagen.2013.09.027

Alsteens, D., Dupres, V., Mc Evoy, K., Wildling, L., Gruber, H. J., and Dufrêne, Y. F. (2008). Structure, Cell wall Elasticity and Polysaccharide Properties of Living Yeast Cells, as Probed by AFM. *Nanotechnology* 19, 384005. doi:10.1088/0957-4484/19/38/384005

Antebi, A., and Fink, G. R. (1992). The Yeast $Ca^{(2+)}$-ATPase Homologue, *PMR1*, Is Required for normal Golgi Function and Localizes in a Novel Golgi-like Distribution. *MBoC* 3, 633–654. doi:10.1091/mbc.3.6.633

Bagur, R., and Hajnóczky, G. (2017). Intracellular Ca^{2+} Sensing: Its Role in Calcium Homeostasis and Signaling. *Mol. Cel* 66, 780–788. doi:10.1016/j.molcel.2017.05.028

Balcavage, W. X., Lloyd, J. L., Mattoon, J. R., Ohnishi, T., and Scarpa, A. (1973). Cation Movements and Respiratory Response in Yeast Mitochondria Treated with High Ca^{2+} Concentrations. *Biochim. Biophys. Acta (BBA) - Bioenerg.* 305, 41–51. doi:10.1016/0005-2728(73)90229-6

Balla, T. (2013). Phosphoinositides: Tiny Lipids with Giant Impact on Cell Regulation. *Physiol. Rev.* 93, 1019–1137. doi:10.1152/physrev.00028.2012

Batiza, A. F., Schulz, T., Masson, P. H., and Masson, P. H. (1996). Yeast Respond to Hypotonic Shock with a Calcium Pulse. *J. Biol. Chem.* 271, 23357–23362. doi:10.1074/jbc.271.38.23357

Beeler, T., Gable, K., Zhao, C., and Dunn, T. (1994). A Novel Protein, CSG2p, Is Required for Ca^{2+} Regulation in *Saccharomyces cerevisiae*. *J. Biol. Chem.* 269, 7279–7284. doi:10.1016/s0021-9258(17)37280-0

Belde, P. J. M., Vossen, J. H., Borst-Pauwels, G. W. F. H., and Theuvenet, A. P. R. (1993). Inositol 1,4,5-trisphosphate Releases Ca^{2+} from Vacuolar Membrane Vesicles of *Saccharomyces cerevisiae*. *FEBS Lett.* 323, 113–118. doi:10.1016/0014-5793(93)81460-h

Berridge, M. J., Bootman, M. D., and Roderick, H. L. (2003). Calcium Signalling: Dynamics, Homeostasis and Remodelling. *Nat. Rev. Mol. Cel Biol.* 4, 517–529. doi:10.1038/nrm1155

Berridge, M. J., Lipp, P., and Bootman, M. D. (2000). The Versatility and Universality of Calcium Signalling. *Nat. Rev. Mol. Cel Biol.* 1, 11–21. doi:10.1038/35036035

Besozzi, D., Manzoni, L., Nobile, M. S., Spolaor, S., Castelli, M., Vanneschi, L., et al. (2020). Computational Intelligence for Life Sciences. *Fundamenta Informaticae* 171, 57–80.

Birchwood, C. J., Saba, J. D., Dickson, R. C., and Cunningham, K. W. (2001). Calcium Influx and Signaling in Yeast Stimulated by Intracellular Sphingosine 1-phosphate Accumulation. *J. Biol. Chem.* 276, 11712–11718. doi:10.1074/jbc.m010221200

Bonilla, M., and Cunningham, K. W. (2003). Mitogen-activated Protein Kinase Stimulation of Ca^{2+} Signaling Is Required for Survival of Endoplasmic Reticulum Stress in Yeast. *MBoC* 14, 4296–4305. doi:10.1091/mbc.e03-02-0113

Bonilla, M., Nastase, K. K., and Cunningham, K. W. (2002). Essential Role of Calcineurin in Response to Endoplasmic Reticulum Stress. *EMBO J.* 21, 2343–2353. doi:10.1093/emboj/21.10.2343

Cai, L., Dalal, C. K., and Elowitz, M. B. (2008). Frequency-modulated Nuclear Localization Bursts Coordinate Gene Regulation. *Nature* 455, 485–490. doi:10.1038/nature07292

Carafoli, E., Balcavage, W. X., Lehninger, A. L., and Mattoon, J. R. (1970). Ca^{2+} Metabolism in Yeast Cells and Mitochondria. *Biochim. Biophys. Acta (BBA) - Bioenerg.* 205, 18–26. doi:10.1016/0005-2728(70)90057-5

Carafoli, E. (2004). Calcium-mediated Cellular Signals: a story of Failures. *Trends Biochem. Sci.* 29, 371–379. doi:10.1016/j.tibs.2004.05.006

Christopher, P. F. (2002). *Computational Cell Biology*. In *Interdisciplinary Applied Mathematics*; V. *20*. Springer-Verlag: New York.

Courchesne, W. E., and Ozturk, S. (2003). Amiodarone Induces a caffeine-inhibited, MID1-dependent Rise in Free Cytoplasmic Calcium in *Saccharomyces cerevisiae*. *Mol. Microbiol.* 47, 223–234. doi:10.1046/j.1365-2958.2003.03291.x

Cox, C. D., Nomura, T., Ziegler, C. S., Campbell, A. K., Wann, K. T., and Martinac, B. (2013). Selectivity Mechanism of the Mechanosensitive Channel MscS Revealed by Probing Channel Subconducting States. *Nat. Commun.* 4, 2137. doi:10.1038/ncomms3137

Cui, J., Kaandorp, J. A., Ositelu, O. O., Beaudry, V., Knight, A., Nanfack, Y. F., et al. (2009a). Simulating Calcium Influx and Free Calcium Concentrations in Yeast. *Cell Calcium* 45, 123–132. doi:10.1016/j.ceca.2008.07.005

Cui, J., Kaandorp, J. A., Sloot, P. M. A., Lloyd, C. M., and Filatov, M. V. (2009b). Calcium Homeostasis and Signaling in Yeast Cells and Cardiac Myocytes. *FEMS Yeast Res.* 9, 1137–1147. doi:10.1111/j.1567-1364.2009.00552.x

Cunningham, K. W., and Fink, G. R. (1994b). Calcineurin-dependent Growth Control in *Saccharomyces cerevisiae* Mutants Lacking PMC1, a Homolog of Plasma Membrane Ca^{2+} ATPases. *J. Cel Biol.* 124, 351–363. doi:10.1083/jcb.124.3.351

Cunningham, K. W., and Fink, G. R. (1994a). Ca^{2+} Transport in *Saccharomyces cerevisiae*. *J. Exp. Biol.* 196, 157–166. doi:10.1242/jeb.196.1.157

Cunningham, K. W., and Fink, G. R. (1996). Calcineurin Inhibits VCX1-dependent H+/Ca^{2+} Exchange and Induces Ca^{2+} ATPases in *Saccharomyces cerevisiae*. *Mol. Cel. Biol.* 16, 2226–2237. doi:10.1128/mcb.16.5.2226

Cyert, M. S. (2003). Calcineurin Signaling in *Saccharomyces cerevisiae*: How Yeast Go Crazy in Response to Stress. *Biochem. Biophysical Res. Commun.* 311, 1143–1150. doi:10.1016/s0006-291x(03)01552-3

Cyert, M. S. (2001). Genetic Analysis of Calmodulin and its Targets in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.* 35, 647–672. doi:10.1146/annurev.genet.35.102401.091302

Cyert, M. S., and Thorner, J. (1992). Regulatory Subunit (*CNB1* Gene Product) of Yeast Ca^{2+}/calmodulin-dependent Phosphoprotein Phosphatases Is Required for Adaptation to Pheromone. *Mol. Cel. Biol.* 12, 3460–3469. doi:10.1128/mcb.12.8.3460

Davis, T. N., Urdea, M. S., Masiarz, F. R., and Thorner, J. (1986). Isolation of the Yeast Calmodulin Gene: Calmodulin Is an Essential Protein. *Cell* 47, 423–431. doi:10.1016/0092-8674(86)90599-4

Denis, V., and Cyert, M. S. (2002). Internal Ca^{2+} Release in Yeast Is Triggered by Hypertonic Shock and Mediated by a TRP Channel Homologue. *J. Cel Biol.* 156, 29–34. doi:10.1083/jcb.200111004

Dunn, T., Gable, K., and Beeler, T. (1994). Regulation of Cellular Ca^{2+} by Yeast Vacuoles. *J. Biol. Chem.* 269, 7273–7278. doi:10.1016/s0021-9258(17)37279-4

Dupont, G., Combettes, L., and Leybaert, L. (2007). Calcium Dynamics: Spatio-Temporal Organization from the Subcellular to the Organ Level. *Int. Rev. Cytol.* 261, 193–245. doi:10.1016/s0074-7696(07)61005-5

Dürr, G., Strayle, J., Plemper, R., Elbs, S., Klee, S. K., Catty, P., et al. (1998). Themedial-Golgi Ion Pump Pmr1 Supplies the Yeast Secretory Pathway with Ca^{2+} and Mn^{2+} Required for Glycosylation, Sorting, and Endoplasmic Reticulum-Associated Protein Degradation. *MBoC* 9, 1149–1162. doi:10.1091/mbc.9.5.1149

Eilam, Y. (1982a). Studies on Calcium Efflux in the Yeast *Saccharomyces cerevisiae*. *Microbios* 35, 99–110.

Eilam, Y., and Chernichovsky, D. (1987). Uptake of Ca^{2+} Driven by the Membrane Potential in Energy-Depleted Yeast Cells. *Microbiology* 133, 1641–1649. doi:10.1099/00221287-133-6-1641

Eilam, Y., and Othman, M. (1990). Activation of Ca^{2+} influx by Metabolic Substrates in *Saccharomyces cerevisiae*: Role of Membrane Potential and Cellular ATP Levels. *J. Gen. Microbiol.* 136, 861–866. doi:10.1099/00221287-136-5-861

Eilam, Y. (1982b). The Effect of Monovalent Cations on Calcium Efflux in Yeasts. *Biochim. Biophys. Acta (BBA) - Biomembranes* 687, 8–16. doi:10.1016/0005-2736(82)90164-x

Fischer, M., Schnell, N., Chattaway, J., Davies, P., Dixon, G., and Sanders, D. (1997). The *Saccharomyces cerevisiae CCH1* Gene Is Involved in Calcium Influx and Mating. *FEBS Lett.* 419, 259–262. doi:10.1016/s0014-5793(97)01466-x

Fridlyand, L. E., Tamarina, N., and Philipson, L. H. (2003). Modeling of Ca^{2+} Flux in Pancreatic β-cells: Role of the Plasma Membrane and Intracellular Stores. *Am. J. Physiology-Endocrinology Metab.* 285, E138–E154. doi:10.1152/ajpendo.00194.2002

Giorgi, C., Marchi, S., and Pinton, P. (2018). The Machineries, Regulation and Cellular Functions of Mitochondrial Calcium. *Nat. Rev. Mol. Cel Biol* 19, 713–730. doi:10.1038/s41580-018-0052-8

Griffin, D. M. (1981). "Water and Microbial Stress," in *Advances in Microbial Ecology* (Springer), 91–136. doi:10.1007/978-1-4615-8306-6_3

Groppi, S., Belotti, F., Brandão, R. L., Martegani, E., and Tisi, R. (2011). Glucose-induced Calcium Influx in Budding Yeast Involves a Novel Calcium Transport System and Can Activate Calcineurin. *Cell Calcium* 49, 376–386. doi:10.1016/j.ceca.2011.03.006

Gustin, M. C., Zhou, X.-L., Martinac, B., and Kung, C. (1988). A Mechanosensitive Ion Channel in the Yeast Plasma Membrane. *Science* 242, 762–765. doi:10.1126/science.2460920

Halachmi, D., and Eilam, Y. (1996). Elevated Cytosolic Free Ca^{2+} concentrations and Massive Ca^{2+} accumulation within Vacuoles, in Yeast Mutant lacking PMR1, a Homolog of Ca^{2+}-ATPase. *FEBS Lett.* 392, 194–200. doi:10.1016/0014-5793(96)00799-5

Hille, B., Dickson, E. J., Kruse, M., Vivas, O., and Suh, B.-C. (2015). Phosphoinositides Regulate Ion Channels. *Biochim. Biophys. Acta (BBA) - Mol. Cel Biol. Lipids* 1851, 844–856. doi:10.1016/j.bbalip.2014.09.010

Ho, B., Baryshnikova, A., and Brown, G. W. (2018). Unification of Protein Abundance Datasets Yields a Quantitative *Saccharomyces cerevisiae* Proteome. *Cel Syst.* 6, 192–205. doi:10.1016/j.cels.2017.12.004

Hohmann, S. (2002). Osmotic Stress Signaling and Osmoadaptation in Yeasts. *Microbiol. Mol. Biol. Rev.* 66, 300–372. doi:10.1128/mmbr.66.2.300-372.2002

Hong, S., Cong, X., Jing, H., Xia, Z., Huang, X., Hu, X., et al. (2013). Characterization of Ca^{2+}/H^{+} Exchange in the Plasma Membrane of *Saccharomyces cerevisiae*. *Arch. Biochem. Biophys.* 537, 125–132. doi:10.1016/j.abb.2013.07.005

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., et al. (2006). COPASI–a COmplex PAthway SImulator. *Bioinformatics* 22, 3067–3074. doi:10.1093/bioinformatics/btl485

Iida, H., Nakamura, H., Ono, T., Okumura, M. S., and Anraku, Y. (1994). *MID1*, a Novel *Saccharomyces cerevisiae* Gene Encoding a Plasma Membrane Protein, Is Required for Ca^{2+} Influx and Mating. *Mol. Cel. Biol.* 14, 8259–8271. doi:10.1128/mcb.14.12.8259

Iida, H., Yagawa, Y., and Anraku, Y. (1990). Essential Role for Induced Ca^{2+} Influx Followed by $[Ca^{2+}]_i$ Rise in Maintaining Viability of Yeast Cells Late in the Mating Pheromone Response Pathway. A Study of $[Ca^{2+}]_i$ in Single *Saccharomyces cerevisiae* Cells with Imaging of Fura-2. *J. Biol. Chem.* 265, 13391–13399. doi:10.1016/s0021-9258(19)38311-5

Ishida, H., Nakashima, K.-i., Kumaki, Y., Nakata, M., Hikichi, K., and Yazawa, M. (2002). The Solution Structure of Apocalmodulin from *Saccharomyces cerevisiae* Implies a Mechanism for its Unique Ca^{2+} Binding Property. *Biochemistry* 41, 15536–15542. doi:10.1021/bi020330r

Jakab, M., Fuerst, J., Gschwentner, M., Bottà, G., Garavaglia, M.-L., Bazzini, C., et al. (2002). Mechanisms Sensing and Modulating Signals Arising from Cell Swelling. *Cell. Physiol. Biochem.* 12, 235–258. doi:10.1159/000067895

Jiang, H., and Sun, S. X. (2013). Cellular Pressure and Volume Regulation and Implications for Cell Mechanics. *Biophysical J.* 105, 609–619. doi:10.1016/j.bpj.2013.06.021

Jung, D. W., Bradshaw, P. C., Litsky, M., and Pfeiffer, D. R. (2004). Ca^{2+} Transport in Mitochondria from Yeast Expressing Recombinant Aequorin. *Anal. Biochem.* 324, 258–268. doi:10.1016/j.ab.2003.10.029

Kanzaki, M., Nagasawa, M., Kojima, I., Sato, C., Naruse, K., Sokabe, M., et al. (1999). Molecular Identification of a Eukaryotic, Stretch-Activated Nonselective Cation Channel. *Science* 285, 882–886. doi:10.1126/science.285.5429.882

Kayingo, G., Kilian, S., and Prior, B. (2001). Conservation and Release of Osmolytes by Yeasts during Hypo-Osmotic Stress. *Arch. Microbiol.* 177, 29–35. doi:10.1007/s00203-001-0358-2

Kedem, O., and Katchalsky, A. (1958). Thermodynamic Analysis of the Permeability of Biological Membranes to Non-electrolytes. *Biochim. Biophys. Acta* 27, 229–246. doi:10.1016/0006-3002(58)90330-5

Kennedy, J., and Eberhart, R. (1995). "Particle Swarm Optimization," In Proceedings of IEEE International Conference on Neural Networks. IEEE, 1942–1948.

Kingsbury, T. J., and Cunningham, K. W. (2000). A Conserved Family of Calcineurin Regulators. *Genes Dev.* 14, 1595–1604. doi:10.1101/gad.14.13.1595

Kleinhans, F. W. (1998). Membrane Permeability Modeling: Kedem-Katchalsky vs a Two-Parameter Formalism. *Cryobiology* 37, 271–289. doi:10.1006/cryo.1998.2135

Kwun, M. S., Lee, H. J., and Lee, D. G. (2021). *β*-amyrin-induced Apoptosis in *Candida albicans* Triggered by Calcium. *Fungal Biol.*. doi:10.1016/j.funbio.2021.03.006

Levin, D. E. (2011). Regulation of Cell wall Biogenesis in *Saccharomyces cerevisiae*: the Cell wall Integrity Signaling Pathway. *Genetics* 189, 1145–1175. doi:10.1534/genetics.111.128264

Li, H., Rao, A., and Hogan, P. G. (2011). Interaction of Calcineurin with Substrates and Targeting Proteins. *Trends Cel Biol.* 21, 91–103. doi:10.1016/j.tcb.2010.09.011

Li, W., Shrivastava, M., Lu, H., and Jiang, Y. (2021). *Calcium-calcineurin Signaling Pathway in* Candida albicans*: A Potential Drug Target*, 249. Microbiological research, 126786. doi:10.1016/j.micres.2021.126786

Locke, E. G., Bonilla, M., Liang, L., Takita, Y., and Cunningham, K. W. (2000). A Homolog of Voltage-Gated Ca^{2+} Channels Stimulated by Depletion of Secretory Ca^{2+} in Yeast. *Mol. Cel. Biol.* 20, 6686–6694. doi:10.1128/mcb.20.18.6686-6694.2000

Luyten, K., Albertyn, J., Skibbe, W. F., Prior, B. A., Ramos, J., Thevelein, J. M., et al. (1995). Fps1, a Yeast Member of the MIP Family of Channel Proteins, Is a Facilitator for Glycerol Uptake and Efflux and Is Inactive under Osmotic Stress. *EMBO J.* 14, 1360–1371. doi:10.1002/j.1460-2075.1995.tb07122.x

Ma, T.-Y., Deprez, M.-A., Callewaert, G., and Winderickx, J. (2021). Coordinated Glucose-Induced Ca^{2+} and Ph Responses in Yeast *Saccharomyces cerevisiae*. *Cell Calcium* 100, 102479. doi:10.1016/j.ceca.2021.102479

Malik-Sheriff, R. S., Glont, M., Nguyen, T. V. N., Tiwari, K., Roberts, M. G., Xavier, A., et al. (2020). BioModels-15 Years of Sharing Computational Models in Life Science. *Nucleic Acids Res.* 48, D407–D415. doi:10.1093/nar/gkz1055

Matsumoto, T. K., Ellsmore, A. J., Cessna, S. G., Low, P. S., Pardo, J. M., Bressan, R. A., et al. (2002). An Osmotically Induced Cytosolic Ca^{2+} Transient Activates Calcineurin Signaling to Mediate Ion Homeostasis and Salt Tolerance of *Saccharomyces cerevisiae*. *J. Biol. Chem.* 277, 33075–33080. doi:10.1074/jbc.m205037200

Milo, R., Jorgensen, P., Moran, U., Weber, G., and Springer, M. (2010). BioNumbers-the Database of Key Numbers in Molecular and Cell Biology. *Nucleic Acids Res.* 38, D750–D753. doi:10.1093/nar/gkp889

Miseta, A., Fu, L., Kellermayer, R., Buckley, J., and Bedwell, D. M. (1999a). The Golgi Apparatus Plays a Significant Role in the Maintenance of Ca^{2+} Homeostasis in the *vps33*Δ Vacuolar Biogenesis Mutant of *Saccharomyces cerevisiae*. *J. Biol. Chem.* 274, 5939–5947. doi:10.1074/jbc.274.9.5939

Miseta, A., Kellermayer, R., Aiello, D. P., Fu, L., and Bedwell, D. M. (1999b). The Vacuolar Ca^{2+}/H^{+} exchanger Vcx1p/Hum1p Tightly Controls Cytosolic Ca^{2+} levels in *S. cerevisiae*. *FEBS Lett.* 451, 132–136. doi:10.1016/s0014-5793(99)00519-0

Montefusco, D. J., Matmati, N., and Hannun, Y. A. (2014). The Yeast Sphingolipid Signaling Landscape. *Chem. Phys. Lipids* 177, 26–40. doi:10.1016/j.chemphyslip.2013.10.006

Montero, M., Brini, M., Marsault, R., Alvarez, J., Sitia, R., Pozzan, T., et al. (1995). Monitoring Dynamic Changes in Free Ca^{2+} Concentration in the Endoplasmic Reticulum of Intact Cells. *EMBO J.* 14, 5467–5475. doi:10.1002/j.1460-2075.1995.tb00233.x

Muller, E. M., Locke, E. G., and Cunningham, K. W. (2001). Differential Regulation of Two Ca^{2+} Influx Systems by Pheromone Signaling in *Saccharomyces cerevisiae*. *Genetics* 159, 1527–1538. doi:10.1093/genetics/159.4.1527

Muller, E. M., Mackin, N. A., Erdman, S. E., and Cunningham, K. W. (2003). Fig1p Facilitates Ca^{2+} Influx and Cell Fusion during Mating of *Saccharomyces cerevisiae*. *J. Biol. Chem.* 278, 38461–38469. doi:10.1074/jbc.m304089200

Nakashima, K.-i., Ishida, H., Ohki, S.-y., Hikichi, K., and Yazawa, M. (1999). Calcium Binding Induces Interaction between the N- and C-Terminal Domains of Yeast Calmodulin and Modulates its Overall Conformation. *Biochemistry* 38, 98–104. doi:10.1021/bi982067t

Nakashima, K., Ishida, H., Nakatomi, A., and Yazawa, M. (2012). Specific Conformation and Ca^{2+}-Binding Mode of Yeast Calmodulin: Insight into Evolutionary Development. *J. Biochem.* 152, 27–35. doi:10.1093/jb/mvs048

Niedzwiecka, K., Tisi, R., Penna, S., Lichocka, M., Plochocka, D., and Kucharczyk, R. (2018). Two Mutations in Mitochondrial *ATP6* Gene of ATP Synthase, Related to Human Cancer, Affect ROS, Calcium Homeostasis and Mitochondrial Permeability Transition in Yeast. *Biochim. Biophys. Acta (BBA) - Mol. Cel Res.* 1865, 117–131. doi:10.1016/j.bbamcr.2017.10.003

Nobile, M. S., and Mauri, G. (2017). "Accelerated Analysis of Biological Parameters Space Using GPUs," in International Conference on Parallel Computing Technologies (Springer), 70–81. doi:10.1007/978-3-319-62932-2_6

Nobile, M. S., Tangherloni, A., Rundo, L., Spolaor, S., Besozzi, D., Mauri, G., et al. (2018). "Computational Intelligence for Parameter Estimation of Biochemical Systems," in 2018 IEEE Congress on Evolutionary Computation (CEC) (IEEE)–18. doi:10.1109/cec.2018.8477873

Nobile, M. S., Votta, G., Palorini, R., Spolaor, S., De Vitto, H., Cazzaniga, P., et al. (2020). Fuzzy Modeling and Global Optimization to Predict Novel Therapeutic Targets in Cancer Cells. *Bioinformatics* 36, 2181–2188. doi:10.1093/bioinformatics/btz868

Odorizzi, G., Babst, M., and Emr, S. D. (2000). Phosphoinositide Signaling and the Regulation of Membrane Trafficking in Yeast. *Trends Biochem. Sci.* 25, 229–235. doi:10.1016/s0968-0004(00)01543-7

Ogura, K., Kumeta, H., Takahasi, K., Kobashigawa, Y., Yoshida, R., Itoh, H., et al. (2012a). Solution Structures of Yeast *Saccharomyces cerevisiae* Calmodulin in Calcium- and Target Peptide-Bound States Reveal Similarities and Differences to Vertebrate Calmodulin. *Genes Cells* 17, 159–172. doi:10.1111/j.1365-2443.2012.01580.x

Ogura, K., Okamura, H., Katahira, M., Katoh, E., and Inagaki, F. (2012b). Conformational Dynamics of Yeast Calmodulin in the Ca^{2+}-Bound State Probed Using NMR Relaxation Dispersion. *FEBS Lett.* 586, 2548–2554. doi:10.1016/j.febslet.2012.06.031

Oh, S., Lee, S. S., Ryu, H. R., Park, J. W., Peter, M., and Jeon, N. L. (2012). "Unraveling Mechano-Stress Responsive Signaling Networks in Budding Yeast via Microfluidic Devices," in 16th International Conference on Miniaturized Systems for Chemistry and Life Sciences (MicroTAS 2012), 1 449–451.

Ohsumi, Y., and Anraku, Y. (1983). Calcium Transport Driven by a Proton Motive Force in Vacuolar Membrane Vesicles of *Saccharomyces cerevisiae*. *J. Biol. Chem.* 258, 5614–5617. doi:10.1016/s0021-9258(20)81935-8

Okada, Y., Maeno, E., Shimizu, T., Dezaki, K., Wang, J., and Morishima, S. (2001). Receptor-mediated Control of Regulatory Volume Decrease (RVD) and Apoptotic Volume Decrease (AVD). *J. Physiol.* 532, 3–16. doi:10.1111/j.1469-7793.2001.0003g.x

Orlean, P. (2012). Architecture and Biosynthesis of the *Saccharomyces cerevisiae* Cell wall. *Genetics* 192, 775–818. doi:10.1534/genetics.112.144485

Paidhungat, M., and Garrett, S. (1997). A Homolog of Mammalian, Voltage-Gated Calcium Channels Mediates Yeast Pheromone-Stimulated Ca^{2+} Uptake and Exacerbates the *cdc1*(Ts) Growth Defect. *Mol. Cel. Biol.* 17, 6339–6347. doi:10.1128/mcb.17.11.6339

Palmer, C. P., Zhou, X.-L., Lin, J., Loukin, S. H., Kung, C., and Saimi, Y. (2001). A TRP Homolog in *Saccharomyces cerevisiae* Forms an Intracellular Ca^{2+} permeable Channel in the Yeast Vacuolar Membrane. *Proc. Natl. Acad. Sci. U.S.A.* 98, 7801–7805. doi:10.1073/pnas.141036198

Pedersen, S. F., Kapus, A., and Hoffmann, E. K. (2011). Osmosensory Mechanisms in Cellular and Systemic Volume Regulation. *Jasn* 22, 1587–1597. doi:10.1681/asn.2010121284

Perera, N. M., Michell, R. H., and Dove, S. K. (2004). Hypo-osmotic Stress Activates Plc1p-dependent Phosphatidylinositol 4,5-bisphosphate Hydrolysis and Inositol Hexakisphosphate Accumulation in Yeast. *J. Biol. Chem.* 279, 5216–5226. doi:10.1074/jbc.m305068200

Petzold, L. (1983). Automatic Selection of Methods for Solving Stiff and Nonstiff Systems of Ordinary Differential Equations. *SIAM J. Sci. Stat. Comput.* 4, 136–148. doi:10.1137/0904010

Pinton, P., Pozzan, T., and Rizzuto, R. (1998). The Golgi Apparatus Is an Inositol 1,4,5-Trisphosphate-Sensitive Ca^{2+} Store, with Functional Properties Distinct from Those of the Endoplasmic Reticulum. *EMBO J.* 17, 5298–5308. doi:10.1093/emboj/17.18.5298

Pizzo, P., Lissandron, V., Capitanio, P., and Pozzan, T. (2011). Ca^{2+} Signalling in the Golgi Apparatus. *Cell Calcium* 50, 184–192. doi:10.1016/j.ceca.2011.01.006

Plattner, H., and Verkhratsky, A. (2015a). Evolution of Calcium Signalling. *Cell Calcium* 57, 121–122. doi:10.1016/j.ceca.2015.02.007

Plattner, H., and Verkhratsky, A. (2015b). The Ancient Roots of Calcium Signalling Evolutionary Tree. *Cell Calcium* 57, 123–132. doi:10.1016/j.ceca.2014.12.004

Purvis, J. E., and Lahav, G. (2013). Encoding and Decoding Cellular Information through Signaling Dynamics. *Cell* 152, 945–956. doi:10.1016/j.cell.2013.02.005

Reed, R. H., Chudek, J. A., Foster, R., and Gadd, G. M. (1987). Osmotic Significance of Glycerol Accumulation in Exponentially Growing Yeasts. *Appl. Environ. Microbiol.* 53, 2119–2123. doi:10.1128/aem.53.9.2119-2123.1987

Rhoads, A. R., and Friedberg, F. (1997). Sequence Motifs for Calmodulin Recognition. *FASEB j.* 11, 331–340. doi:10.1096/fasebj.11.5.9141499

Rigamonti, M., Groppi, S., Belotti, F., Ambrosini, R., Filippi, G., Martegani, E., et al. (2015). Hypotonic Stress-Induced Calcium Signaling in *Saccharomyces cerevisiae* Involves TRP-like Transporters on the Endoplasmic Reticulum Membrane. *Cell Calcium* 57, 57–68. doi:10.1016/j.ceca.2014.12.003

Rudolph, H. K., Antebi, A., Fink, G. R., Buckley, C. M., Dorman, T. E., LeVitre, J., et al. (1989). The Yeast Secretory Pathway Is Perturbed by Mutations in PMR1, a Member of a Ca^{2+} ATPase Family. *Cell* 58, 133–145. doi:10.1016/0092-8674(89)90410-8

Sachs, F., and Morris, C. E. (1998). "Mechanosensitive Ion Channels in Nonspecialized Cells," in *Reviews of Physiology Biochemistry and Pharmacology* (Springer), 132, 1–77.

Sackin, H. (1995). Stretch-activated Ion Channels. *Kidney Int.* 48, 1134–1147. doi:10.1038/ki.1995.397

Saimi, Y., and Kung, C. (2002). Calmodulin as an Ion Channel Subunit. *Annu. Rev. Physiol.* 64, 289–311. doi:10.1146/annurev.physiol.64.100301.111649

Sammels, E., Parys, J. B., Missiaen, L., De Smedt, H., and Bultynck, G. (2010). Intracellular Ca^{2+} Storage in Health and Disease: a Dynamic Equilibrium. *Cell Calcium* 47, 297–314. doi:10.1016/j.ceca.2010.02.001

Schaber, J., Adrover, M. À., Eriksson, E., Pelet, S., Petelenz-Kurdziel, E., Klein, D., et al. (2010). Biophysical Properties of *Saccharomyces cerevisiae* and Their

Relationship with HOG Pathway Activation. *Eur. Biophys. J.* 39, 1547–1556. doi:10.1007/s00249-010-0612-0

Schaber, J., and Klipp, E. (2008). Short-term Volume and Turgor Regulation in Yeast. *Essays Biochem.* 45, 147–160. doi:10.1042/bse0450147

Nobile, M. S., Cazzaniga, P., Besozzi, D., and Mauri, G. (2019). ginSODA: Massive Parallel Integration of Stiff ODE Systems on GPUs. *J. Supercomput* 75, 7844–7856. doi:10.1007/s11227-018-2549-5

Sorin, A., Rosas, G., and Rao, R. (1997). PMR1, a Ca^{2+}-ATPase in Yeast Golgi, Has Properties Distinct from Sarco/endoplasmic Reticulum and Plasma Membrane Calcium Pumps. *J. Biol. Chem.* 272, 9895–9901. doi:10.1074/jbc.272.15.9895

Spolaor, S., Gribaudo, M., Iacono, M., Kadavy, T., Komínková Oplatková, Z., Mauri, G., et al. (2019a). "Towards Human Cell Simulation," in *High-Performance Modelling and Simulation for Big Data Applications* (Cham: Springer), 221–249. doi:10.1007/978-3-030-16272-6_8

Spolaor, S., Nobile, M. S., Mauri, G., Cazzaniga, P., and Besozzi, D. (2019b). Coupling Mechanistic Approaches and Fuzzy Logic to Model and Simulate Complex Systems. *IEEE Trans. Fuzzy Syst.* 28, 1748–1759.

Starovasnik, M. A., Davis, T. N., and Klevit, R. E. (1993). Similarities and Differences between Yeast and Vertebrate Calmodulin: an Examination of the Calcium-Binding and Structural Properties of Calmodulin from the Yeast *Saccharomyces cerevisiae*. *Biochemistry* 32, 3261–3270. doi:10.1021/bi00064a008

Stathopoulos, A. M., and Cyert, M. S. (1997). Calcineurin Acts through the *CRZ1/TCN1*-Encoded Transcription Factor to Regulate Gene Expression in Yeast. *Genes Dev.* 11, 3432–3444. doi:10.1101/gad.11.24.3432

Stefan, C. J., Manford, A. G., and Emr, S. D. (2013a). ER-PM Connections: Sites of Information Transfer and Inter-organelle Communication. *Curr. Opin. Cel Biol.* 25, 434–442. doi:10.1016/j.ceb.2013.02.020

Stefan, C. P., and Cunningham, K. W. (2013). Kch1 Family Proteins Mediate Essential Responses to Endoplasmic Reticulum Stresses in the Yeasts *Saccharomyces cerevisiae* and Candida Albicans. *J. Biol. Chem.* 288, 34861–34870. doi:10.1074/jbc.m113.508705

Stefan, C. P., Zhang, N., Sokabe, T., Rivetta, A., Slayman, C. L., Montell, C., et al. (2013b). Activation of an Essential Calcium Signaling Pathway in *Saccharomyces cerevisiae* by Kch1 and Kch2, Putative Low-Affinity Potassium Transporters. *Eukaryot. Cel* 12, 204–214. doi:10.1128/ec.00299-12

Strahl, T., and Thorner, J. (2007). Synthesis and Function of Membrane Phosphoinositides in Budding Yeast, *Saccharomyces cerevisiae*. *Biochim. Biophys. Acta (BBA) - Mol. Cel Biol. Lipids* 1771, 353–404. doi:10.1016/j.bbalip.2007.01.015

Strayle, J., Pozzan, T., and Rudolph, H. K. (1999). Steady-state Free Ca^{2+} in the Yeast Endoplasmic Reticulum Reaches Only 10 microM and Is Mainly Controlled by the Secretory Pathway Pump Pmr1. *EMBO J.* 18, 4733–4743. doi:10.1093/emboj/18.17.4733

Sukharev, S. I., Sigurdson, W. J., Kung, C., Sachs, F., and Sukharev, S. I. (1999). Energetic and Spatial Parameters for Gating of the Bacterial Large Conductance Mechanosensitive Channel, MscL. *J. Gen. Physiol.* 113, 525–540. doi:10.1085/jgp.113.4.525

Takita, Y., Engstrom, L., Ungermann, C., and Cunningham, K. W. (2001). Inhibition of the Ca^{2+}-ATPase Pmc1p by the V-SNARE Protein Nyv1p. *J. Biol. Chem.* 276, 6200–6206. doi:10.1074/jbc.m009191200

Talemi, S. R., Tiger, C.-F., Andersson, M., Babazadeh, R., Welkenhuysen, N., Klipp, E., et al. (2016). Systems Level Analysis of the Yeast Osmo-Stat. *Sci. Rep.* 6, 30950. doi:10.1038/srep30950

Tamas, M. J., Luyten, K., Sutherland, F. C. W., Hernandez, A., Albertyn, J., Valadi, H., et al. (1999). Fps1p Controls the Accumulation and Release of the Compatible Solute Glycerol in Yeast Osmoregulation. *Mol. Microbiol.* 31, 1087–1104. doi:10.1046/j.1365-2958.1999.01248.x

Tangherloni, A., Spolaor, S., Cazzaniga, P., Besozzi, D., Rundo, L., Mauri, G., et al. (2019). Biochemical Parameter Estimation vs. Benchmark Functions: A Comparative Study of Optimization Performance and Representation Design. *Appl. Soft Comput.* 81, 105494. doi:10.1016/j.asoc.2019.105494

Tanida, I., Takita, Y., Hasegawa, A., Ohya, Y., and Anraku, Y. (1996). Yeast Cls2p/Csg2p Localized on the Endoplasmic Reticulum Membrane Regulates a Non-exchangeable Intracellular Ca^{2+} Pool Cooperatively with Calcineurin. *FEBS Lett.* 379, 38–42. doi:10.1016/0014-5793(95)01478-0

Teng, J., Goto, R., Iida, K., Kojima, I., and Iida, H. (2008). Ion-channel Blocker Sensitivity of Voltage-Gated Calcium-Channel Homologue Cch1 in *Saccharomyces cerevisiae*. *Microbiology (Reading, Engl.)* 154, 3775–3781. doi:10.1099/mic.0.2008/021089-0

Tisi, R., Baldassa, S., Belotti, F., and Martegani, E. (2002). Phospholipase C Is Required for Glucose-Induced Calcium Influx in Budding Yeast. *FEBS Lett.* 520, 133–138. doi:10.1016/s0014-5793(02)02806-5

Tisi, R., Belotti, F., Wera, S., Winderickx, J., Thevelein, J. M., and Martegani, E. (2004). Evidence for Inositol Triphosphate as a Second Messenger for Glucose-Induced Calcium Signalling in Budding Yeast. *Curr. Genet.* 45, 83–89. doi:10.1007/s00294-003-0465-5

Tisi, R., Rigamonti, M., Groppi, S., and Belotti, F. (2016). Calcium Homeostasis and Signaling in Fungi and Their Relevance For Pathogenicity of Yeasts and Filamentous Fungi. *AIMS Mol. Sci.* 3, 505–549. doi:10.3934/molsci.2016.4.505

Viladevall, L., Serrano, R., Ruiz, A., Domenech, G., Giraldo, J., Barceló, A., et al. (2004). Characterization of the Calcium-Mediated Response to Alkaline Stress in *Saccharomyces cerevisiae*. *J. Biol. Chem.* 279, 43614–43624. doi:10.1074/jbc.m403606200

Wang, L., Zhang, Y., Yang, D., Zhang, S., Han, D., and Luo, Y. (2021). Aureoverticillactam, a Potent Antifungal Macrocyclic Lactam from Streptomyces Aureoverticillatus Nh6, Generates Calcium Dyshomeostasis Induced Cell Apoptosis via the Phospholipase C Pathway in fusarium Oxysporum F. Sp. Cubense Race 4. *Phytopathology*.

Wei, Y., Marchi, V., Wang, R., and Rao, R. (1999). An N-Terminal EF Hand-like Motif Modulates Ion Transport by Pmr1, the Yeast Golgi Ca^{2+}/Mn^{2+}-ATPase. *Biochemistry* 38, 14534–14541. doi:10.1021/bi9911233

Weijiu, L. (2012). *Introduction to Modeling Biological Cellular Control Systems*, 6. Springer Science & Business Media.

Weskamp, M., Seidl, W., and Grissmer, S. (2000). Characterization of the Increase in $[Ca^{2+}]$ I during Hypotonic Shock and the Involvement of Ca^{2+}-Activated K^+ Channels in the Regulatory Volume Decrease in Human Osteoblast-like Cells. *J. Membr. Biol.* 178, 11–20. doi:10.1007/s002320010010

Wuytack, F., Raeymaekers, L., and Missiaen, L. (2003). PMR1/SPCA Ca^{2+} Pumps and the Role of the Golgi Apparatus as a Ca^{2+} Store. *Pflugers Arch. - Eur. J. Physiol.* 446, 148–153. doi:10.1007/s00424-003-1011-5

Yap, K. L., Kim, J., Truong, K., Sherman, M., Yuan, T., and Ikura, M. (2000). Calmodulin Target Database. *J. Struct. Funct. Genomics* 1, 8–14. doi:10.1023/a:1011320027914

Yoshimoto, H., Saltsman, K., Gasch, A. P., Li, H. X., Ogawa, N., Botstein, D., et al. (2002). Genome-wide Analysis of Gene Expression Regulated by the calcineurin/Crz1p Signaling Pathway in *Saccharomyces cerevisiae*. *J. Biol. Chem.* 277, 31079–31088. doi:10.1074/jbc.m202718200

Yoshimura, H., Tada, T., and Iida, H. (2004). Subcellular Localization and Oligomeric Structure of the Yeast Putative Stretch-Activated Ca^{2+} Channel Component Mid1. *Exp. Cel Res.* 293, 185–195. doi:10.1016/j.yexcr.2003.09.020

Zheng, J. (2013). Molecular Mechanism of TRP Channels. *Compr. Physiol.* 3, 221–242. doi:10.1002/cphy.c120001

Zhu, M. X. (2005). Multiple Roles of Calmodulin and Other Ca^{2+}-Binding Proteins in the Functional Regulation of TRP Channels. *Pflugers Arch. - Eur. J. Physiol.* 451, 105–115. doi:10.1007/s00424-005-1427-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Spolaor, Rovetta, Nobile, Cazzaniga, Tisi and Besozzi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 31 May 2022
doi: 10.3389/fmolb.2022.856212

An Epistatic Network Describes *oppA* and *glgB* as Relevant Genes for *Mycobacterium tuberculosis*

Ali-Berenice Posada-Reyes[1,2]*, *Yalbi I. Balderas-Martínez*[3], *Santiago Ávila-Ríos*[3], *Pablo Vinuesa*[4] and *Salvador Fonseca-Coronado*[2]*

[1]Posgrado en Ciencias Biológicas, UNAM, Mexico, Mexico, [2]Facultad de Estudios Superiores Cuautitlán, UNAM, Estado de Mexico, Mexico, [3]Instituto Nacional de Enfermedades Respiratorias "Ismael Cosio Villegas", Ciudad de Mexico, Mexico, [4]Centro de Ciencias Genómicas, UNAM, Cuernavaca, Mexico

Mycobacterium tuberculosis is an acid-fast bacterium that causes tuberculosis worldwide. The role of epistatic interactions among different loci of the *M. tuberculosis* genome under selective pressure may be crucial for understanding the disease and the molecular basis of antibiotic resistance acquisition. Here, we analyzed polymorphic loci interactions by applying a model-free method for epistasis detection, SpydrPick, on a pan–genome-wide alignment created from a set of 254 complete reference genomes. By means of the analysis of an epistatic network created with the detected epistatic interactions, we found that *glgB* (α-1,4-glucan branching enzyme) and *oppA* (oligopeptide-binding protein) are putative targets of co-selection in *M. tuberculosis* as they were associated in the network with *M. tuberculosis* genes related to virulence, pathogenesis, transport system modulators of the immune response, and antibiotic resistance. In addition, our work unveiled potential pharmacological applications for genotypic antibiotic resistance inherent to the mutations of *glgB* and *oppA* as they epistatically interact with *fprA* and *embC*, two genes recently included as antibiotic-resistant genes in the catalog of the World Health Organization. Our findings showed that this approach allows the identification of relevant epistatic interactions that may lead to a better understanding of *M. tuberculosis* by deciphering the complex interactions of molecules involved in its metabolism, virulence, and pathogenesis and that may be applied to different bacterial populations.

Keywords: *glgB*, *oppA*, epistatic network, co-selection, *Mycobacterium tuberculosis*, tuberculosis

OPEN ACCESS

Edited by:
Alberto Jesus Martin, Universidad Mayor, Chile

Reviewed by:
Ramandeep Singh, Translational Health Science and Technology Institute (THSTI), India
Verlaine Joy Timms, The University of Newcastle, Australia

***Correspondence:**
Ali-Berenice Posada-Reyes
bereniceposadareyes@yahoo.com.mx
Salvador Fonseca-Coronado
fonsecacoronado@yahoo.com

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *16 January 2022*
Accepted: *11 March 2022*
Published: *31 May 2022*

Citation:
Posada-Reyes A-B, Balderas-Martínez YI, Ávila-Ríos S, Vinuesa P and Fonseca-Coronado S (2022) An Epistatic Network Describes oppA and glgB as Relevant Genes for Mycobacterium tuberculosis. Front. Mol. Biosci. 9:856212. doi: 10.3389/fmolb.2022.856212

1 INTRODUCTION

In humans, tuberculosis (TB) is a chronic and highly contagious disease that causes more than 10 million human infections and 1.8 million deaths worldwide every year. The constant arrival of drug-resistant strains complicates its control and eradication (Gupta et al., 2018). This disease is mainly caused by members of the *Mycobacterium tuberculosis* complex (MTBC) (Coscolla and Gagneux, 2014) *via* aerosolized bacteria released by patients with TB (Lerner et al., 2015).

Mycobacterium tuberculosis (Mtb) lineages L1–L4 and L7 form a large group of human-adapted strains responsible for the vast majority of global human TB cases, whereas *Mycobacterium africanum* lineages (L5 and L6), which are restricted to humans from West Africa, are phylogenetically linked with the eighth lineage, which comprises various animal-adapted strains (Gonzalo-Asensio et al., 2014).

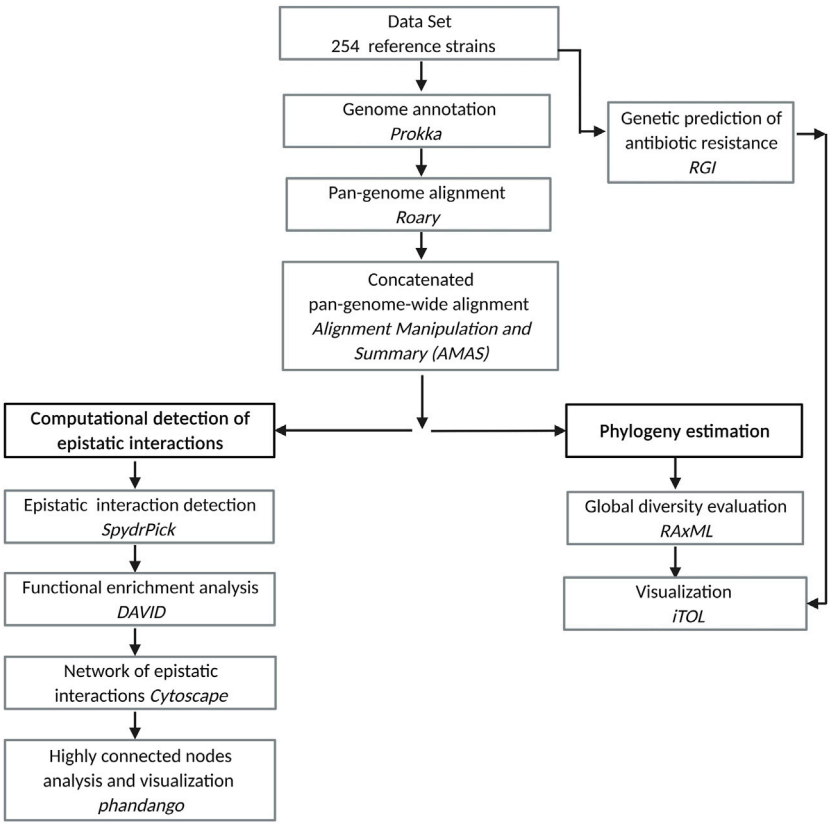


FIGURE 1 | Pipeline for the study of epistatic interactions in Mtb.

The first complete genome sequence of Mtb was described in 1998 (Cole et al., 1998). Since then, whole-genome sequencing (WGS) has been applied to a wide range of clinical scenarios, with the potential to revolutionize TB diagnosis, outbreak investigation, development of drugs and vaccines, and to assist in understanding the evolution and pathogenicity of MTBC (Satta et al., 2018). The increase in genomic data in this new era of big data can be considered a great opportunity to continue with the epidemiological surveillance of Mtb associated with the evaluation of genotypic antibiotic resistance. Moreover, it may allow us to unveil new genes with characteristics that lead us to a better understanding of TB.

Recent advances in the scale and diversity of population genomic data for Mtb provide the potential for revealing whole-genome genetic patterns. Statistical methods combined with recent advances in computational structural biology have identified the polymorphic loci (positions inside a genome) under the strongest co-evolutionary pressures or epistatic interactions (Skwark et al., 2017). Such epistatic interactions describe a functional relationship between genes or polymorphic loci (Sackton and Hartl, 2016). Studies of interactions between mutations in Mtb that result in resistance to diverse drugs have suggested that epistasis may be related to multidrug resistance (Trauner et al., 2014; Kavvas et al., 2018). However, the role of epistatic interactions among many regions of the genome under selection in Mtb remains unknown, and further study will contribute to improving our knowledge of TB.

In this study, we analyzed polymorphic loci interactions for epistatic detection in a set of 254 complete reference genomes from Mtb by the use of the model-free method, SpydrPick (Pensar et al., 2019). SpydrPick is based on calculating the mutual information between two polymorphic loci. This well-annotated reference collection integrates genome annotation, gene characterization, and a sequence variation report with a high certainty of genomic location. First, a pan-genome was created using Roary (Page et al., 2015). Then, using AMAS, a pan–genome-wide alignment was obtained by concatenating individual gene alignments. This pan–genome-wide alignment was the input for SpydrPick.

The application of the method to this data set allowed us to reconstruct an epistatic network. The analysis of this network revealed two putative targets of co-selection (*glgB* and *oppA*) associated with Mtb genes related to virulence, pathogenesis, transport system modulators of the immune response, and antibiotic resistance. This work may have relevant applications in the characterization of new genes involved in the worldwide problem of Mtb drug resistance (WHO, 2021).

2 MATERIALS AND METHODS

An overview of our approach is depicted in **Figure 1**. The steps are described in the following subsections.

2.1 Data Set

We gathered 254 reference strains of Mtb from the NCBI Refseq database that was available as of 4 November 2020. The list of strains is provided as **Supplementary Data S1**.

2.2 Creating Pan–Genome-Wide Alignment

Following the strategy of Pensar et al. (2019), we created a pan–genome-wide alignment of the 254 strains. First, we employed Prokka (Seemann, 2014) to annotate genes and features of interest in the set of strains. This genome annotation (GFF3 format) was the input to create a pan-genome of the strains with Roary (Page et al., 2015).

This tool extracts the gene sequences from the input and then identifies clusters to obtain gene alignments. Roary considers two categories of genes: core and accessory. A gene is considered "core" if it is in at least a certain percentage of strains (isolates) defined by the user. In our study, we followed the approach of Pensar et al. (2019), who set this percentage on 95% strains. The output of Roary is a set of files with individual gene alignments, with one file per gene. These files are concatenated in a matrix using the Alignment Manipulation and Summary (AMAS) tool (Borowiec, 2016). This matrix is formed by gene 1 joined on the right with gene 2 and so on with the rest of the genes [see the example "A: concatenation" from **Figure 1** in Borowiec (2016)]. Thus, the columns of the output matrix are the genes, and the number of rows is the number of strains used to generate the pan–genome-wide alignment (254 in this case).

2.3 Global Diversity Evaluation

The pan–genome-wide alignment was evaluated for global diversity by estimating a phylogeny using RAxML Next Generation (Kozlov et al., 2019). A standard nonparametric bootstrap of 1,000 replicates was performed. Phylogenies were visualized using iTOL v. 6.4.1 (Letunic and Bork, 2021).

2.4 Genetic Prediction of Antibiotic Resistance

We predicted a resistome for the 254 strains using the Resistance Gene Identifier (RGI) tool v. 5.1.1 (Alcock et al., 2020). RGI uses the Comprehensive Antibiotic Resistance Database (CARD) as reference data. Using the output of RGI, we annotated strains for two genotypic characterizations of antibiotic resistance: multidrug-resistant (MDR) strains for those strains with genes resistant to isoniazid and rifampicin and extensively drug-resistant (XDR) strains if they have genes resistant to isoniazid, rifampicin, fluoroquinolone, and at least one of the following three antibiotics: kanamycin, amikacin, or capreomycin. These annotations were incorporated for visualization into the phylogeny displayed by iTOL.

2.5 Computational Detection of Epistatic Interactions

2.5.1 Epistatic Interaction Detection

We utilized SpydrPick (Pensar et al., 2019) to detect the epistatic interactions in the pan–genome-wide alignment. SpydrPick is a model-free method whose computational efficiency enables analysis at the scale of pan-genomes of bacteria. This method facilitates the detection of targets of co-selection related to virulence and antibiotic resistance. The potential of this method is the detection of epistatic interactions in the absence of phenotypic data.

The approach of SpydrPick is based on calculating the mutual information (MI) between two polymorphic loci. MI is an information-theoretic measure of the amount of information that one random variable, X, contains about another random variable, Y. MI is also defined as the reduction in uncertainty in X after observing Y; in other words, MI manifests the reduction in uncertainty of X due to the knowledge of Y (Cover and Thomas, 2006). MI gives a measure of association or correlation between X and Y (Chanda et al., 2020); if the two variables, X and Y, are independent, then the MI is zero. MI is formally defined as follows:

$$MI(X,Y) = \sum_{x\in(X)} \sum_{y\in(Y)} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right), \qquad (1)$$

where $p(x, y)$ is the joint probability and $p(x)$ and $p(y)$ are the marginal probabilities of X and Y. MI has been successfully used for detecting co-selection in bacterial population genomics at a genome-wide scale. Another relevant feature introduced by SpydrPick's approach is the correction for the population structure. This is applied by a sequence reweighting strategy based on how different are the sequences in the pan–genome-wide alignment (Pensar et al., 2019).

SpydrPick detects direct and indirect interactions between loci. A direct interaction occurs between two positions ($P_1 \rightarrow P_2$), whereas an indirect interaction occurs when the two positions (P_1 and P_2) are also linked through a third position ($P_1 \rightarrow P_3 \rightarrow P_2$). In the case of indirect interactions ($P_1 \rightarrow P_2$), SpydrPick removes the interaction if the MI is not larger than the other two interactions ($P_1 \rightarrow P_3$ and $P_3 \rightarrow P_2$).

In addition, SpydrPick performs an analysis to detect outlier interactions. A first criterion to filter outliers is that the distance (bp) between the positions of polymorphic loci must be greater than a linkage disequilibrium (LD) parameter. In this case, a strong LD refers to a close genetic distance between two nucleotide positions. Due to a strong LD hiding a prospective signal of shared co-evolutionary selection pressure, SpydrPick filters out pairs of positions with strong LD to select outlier interactions. According to the SpydrPick's documentation (https://github.com/santeripuranen/SpydrPick), for bacterial genomes, the typical values of the LD are in the 500–20,000 bp range, and the default approach to filtering out strong LD pairs is using a simple distance-based cut-off (20,000 in our case). The second criterion is that the MI must be greater than a threshold obtained from Tukey's outlier test $Q_3 + 1.5 \times (Q_3 - Q_1)$ (Tukey, 1977).

The output of SpydrPick is a table of epistatic interactions that includes the pair of positions of two interacting polymorphic loci in the pan–genome-wide alignment, the genome distance between the two positions, the type of interaction (direct/indirect), and the MI score. When SpydrPick detects outliers,

they are reported in another table, including three additional fields: the MI score without gaps, the gap effect, and if the outlier is considered an extreme outlier ($MI > Q_3 + 3 \times (Q_3—Q_1)$). From the input alignment, SpydPick categorizes any character different from A, C, G, and T as a gap. Gaps are considered in the default MI calculation, so X and Y have an outcome space of five categories. As the gaps may not be informative, SpydrPick calculates for each pair of positions in the outliers another MI score considering only those strains without gaps in either of the two positions. This MI score is named mutual information without gaps (MI_wo_gaps). Using the MI score without gaps, the gap effect is calculated as $(1—MI_wo_gaps/MI) \times 100$ to quantify the positive or negative effect on the MI by discarding strains with gaps in the two positions.

Comparing MI scores without gaps in a meaningful way is difficult due to the fact that the set of strains without gaps in the two positions varies between pairs of positions (Pensar et al., 2019). However, a high value of the gap effect for a given pair of positions may indicate a gap-driven interaction, and a manual analysis of the pair should be required. Thus, following the analysis performed by Pensar et al. (2019), we used the default MI, leaving the analysis of the MI_wo_gaps for a future in-depth study.

The loci of epistatic interaction were annotated with gene id and gene name. Gene names were obtained from the partitions generated by AMAS using an R script. Afterward, using another R script (https://github.com/biotb/epitb-net) and the R Biomartr library (Drost and Paszkowski, 2017), we retrieved the ENTREZ gene id by searching the gene name in the GFF file of the Mtb H37Rv reference genome (GCF_000195955.2).

If there is no gene name detected by Roary during the pan-genome creation, then Roary gives a unique generic name formed by the prefix group and a consecutive number. These generic names also appear in partitions of AMAS; however, no ENTREZ id could be associated with these generic names as these names did not exist in the reference genome GFF file.

On the other hand, Prokka was indicated by a numeric suffix different annotation for the same gene, such as carB_1 and carB_2 (carbamoyl-phosphate synthase large chain). These names were also not found in the reference genome GFF file. In these cases, we eliminated the numeric suffix to find the gene name in the reference genome file. For example, we were able to find the gene id 886,253 for *carB*.

2.5.2 Functional Enrichment Analysis

We used the database for annotation, visualization, and integrated discovery (DAVID) v6.8 (Huang et al., 2009) to obtain a functional annotation of Gene Ontology (GO) terms and KEGG pathways of the genes participating in the epistatic interactions. Specifically, we used the DAVID Web Service Python Script (Jiao et al., 2012) to generate a chart report.

2.5.3 Network of Epistatic Interactions

The set of epistatic interactions can be seen as a model of complex epistatic relations that may be analyzed and displayed as a network. Here, we used Cytoscape (Kohl et al., 2011) to study our set of epistatic interactions. This tool has been utilized for studying diverse types of genetic networks. Cytoscape includes an Analyze Network Tool that calculates several network parameters, such as node degree and betweenness centrality. Another useful tool of Cytoscape is the set of layout algorithms based on the yFile Layout Algorithm App. These algorithms visually organized a network by aligning and rotating groups of nodes.

2.5.4 Highly Connected Nodes Analysis and Visualization

We focused on the most highly connected genes (the highest degree) for analyzing our epistatic network. Functional characterization of these genes was performed by literature curation and showing enriched GO terms for genes interacting with them. In addition, we used the R package SeqinR (Gouy et al., 1985) to upload the pan–genome-wide alignment and extract the allele distribution at loci involved in their epistatic interactions. We used the interactive web tool Phandango, which is used to visualize phylogenetic trees and associated genomic information (Hadfield et al., 2017), to show the estimated phylogeny and allele distribution of loci.

3 RESULTS AND DISCUSSION

3.1 Pan–Genome-Wide Alignment

A total of 6,205 individual genes were aligned by Roary, including 3,659 core genes. After concatenating all individual genes with AMAS, a pan–genome-wide alignment of 6,751,593 bp was obtained.

3.2 Estimated Phylogeny and Antibiotic Resistance Prediction

Based on Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), which are theoretical information criteria to penalize complex models, we selected the estimated phylogeny using a GTR model with four free rates (GTR-R4-FO). The comparison of models is provided in **Supplementary Table S1**. Convergence using the extended majority rule (MRE) criterion (Pattengale et al., 2010) with a 3% cutoff for the bootstrapping was reached after 400 trees.

The prediction of antibiotic resistance by RGI reported that 100% of the 254 strains were MDR (**Figure 2**) and, within this, 15% were XDR. This result indicates that bacterial strains, perhaps currently circulating, present a high level of resistance to first-line treatments, hindering the successful response to treatment and facilitating the dissemination of strains with drug resistance mutations. Thus, detecting epistatic interactions to elucidate polymorphic loci under the strongest co-evolutionary pressure is of utmost importance for molecular surveillance with bioinformatic tools that help us characterize them promptly. Currently, it is reported that 3.4% of the new TB patients and 20% of the patients with a history of previous treatment for TB were diagnosed with MDR TB worldwide (WHO, 2021).

The phylogenetic tree (**Figure 2**) shows the nucleotide diversity of Mtb (254 strain collection). In this study, there is

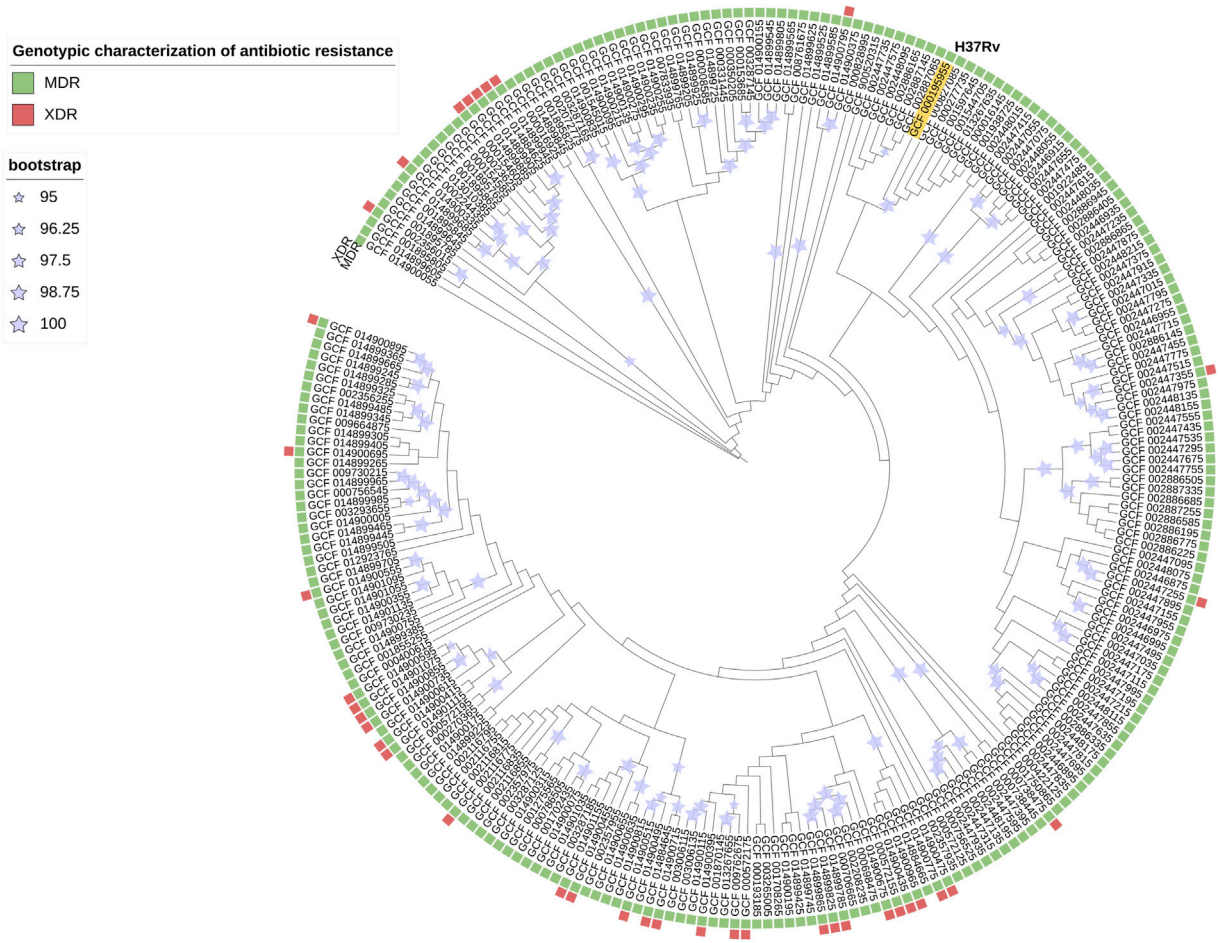


FIGURE 2 | Phylogenetic tree pan-genome Mtb is an iTOL circular visualization with the branch length and the bootstrap values displayed. The tree is based on the Maximum Likelihood topology of 254 strains representative of Mtb diversity and shows that length is proportional to nucleotide topology. Bootstrap values for clades corresponding to the main Mtb clades are shown. The colors correspond to the different genotypic characterizations of antibiotic resistance (MDR = Multidrug Resistant; XDR = Extensively Drug-Resistant).

heterogeneity of submitters, 100% of the strains present genes linked to antibiotic resistance and with respect to the H37Rv strain (reference), and most of the strains present greater genetic diversity.

3.3 Detected Epistatic Interactions

SpydrPick detected 10,573 outlier epistatic interactions (5,484 directed and 5,089 indirect). These interactions describe polymorphic loci under the strongest co-evolutionary pressure. A table with the complete list of outliers is provided as **Supplementary Data S2**. This table includes the fields described in **subsection 2.5.1**, that is, the pair of positions of the two interacting polymorphic loci in the pan–genome-wide alignment, genome distance between the two positions, type of interaction (direct/indirect), MI score, MI score without gaps, gap effect, and whether the outlier interaction is considered an extreme outlier.

After gene annotation, we generated a new table of epistatic interactions that excluded the generic gene names given by Roary. The new table of outliers included 1,940 epistatic interactions among 107 unique genes. From this set of genes, we only found 70 in the reference genome GFF file, and they were associated with their ENTREZ id. Filtering only those interactions, including these 70 genes, we obtained a final table with 890 outlier interactions. The remaining interactions that were not considered in our study will be included in a future analysis.

The final table of outlier interactions includes the two positions of the two interacting polymorphic loci in the pan–genome-wide alignment, gene ENTREZ id and gene name for each position, distance between the two positions (bp), type of interaction (direct = 1, indirect = 0), MI score of the interaction, and if the interaction outlier is an extreme outlier (yes = 1, no = 0). This final table of outliers (**Supplementary Data S3**) was used for enrichment analysis, network reconstruction, and analyses.

SpydrPick was able to find long-distance interactions surpassing the two million bp (**Table 1**). This fact confirms that our study has a whole-genome scale. However, because

TABLE 1 | Statistics of the distance between positions of the two interacting polymorphic loci and statistics of the MI scores, both for the final outlier interactions.

Statistic of the final outlier interactions	Values
Minimum distance (bp)	20,870
Maximum distance (bp)	2,328,291
Median distance (bp)	721,980
Mean distance (bp)	846,454
Minimum MI score	0.4130
Maximum MI score	0.5020
Median MI score	0.4202
Mean MI score	0.4509

we identified epistatic interactions from a pan–genome-wide alignment constructed by concatenating individual gene alignments, the positions are not straightforward whole-genome loci. The minimum distance (bp) between the positions of the two interacting polymorphic loci nearly surpassed the LD criterion of 20,000 bp. The mean of the distance between polymorphic loci in the outlier interactions was 846,454 bp; considering that it is greater than the median (721,980 bp), there may be a slight skewness to distances lower than the mean. On the other hand, the range of MI scores was short, from 0.4130 to 0.5020 (**Table 1**). The MI scores might show a skewness to low values as the mean (0.4509) was higher than the median (0.4202).

All loci in the 890 interactions were found in the described single-nucleotide polymorphisms (SNPs) when we used the pan–genome-wide alignment with the tool SNP-sites v. 2.5.1, which can rapidly identify SNPs from a multi-FASTA alignment (Page et al., 2016). This additional step was developed to identify polymorphisms involved in the detected epistatic interactions.

3.4 Epistatic Network Analysis

3.4.1 *glgB* and *oppA* as Putative Targets of Co-selection

The network of epistatic interactions was analyzed to figure out those genes with a high node degree (the number of edges), that is, a high level of connectivity of the gene with other genes. The most highly connected genes were *glgB* (ENTREZ:886,893, degree = 56), a α-1,4-glucan branching enzyme (GlgB), and *oppA* (ENTREZ:886,985, degree = 37), an oligopeptide-binding protein (OppA) (**Figure 3**).

These two genes also have the highest value of betweenness centrality (*glgB* = 0.493, *oppA* = 0.219). Betweenness centrality is higher for those nodes that join subnets (communities) than those located inside the subnets. Here, we observed three subnets connected by these two genes. An interesting pattern is that each subnet has a different distribution of values of MI (see gradient color of edges in **Figure 3**). The subnet at the bottom has higher MI values (median MI = 0.502) than the other two, the top subnet has a median MI of 0.463, and the subnet at the middle has lower values (median MI = 0.420). A further study is required to elucidate the cause of this pattern. In addition, future analysis will be required to identify the patterns associated with the isolated subnets (*mmpL1-mmpS4* and *lipR-ponA1*).

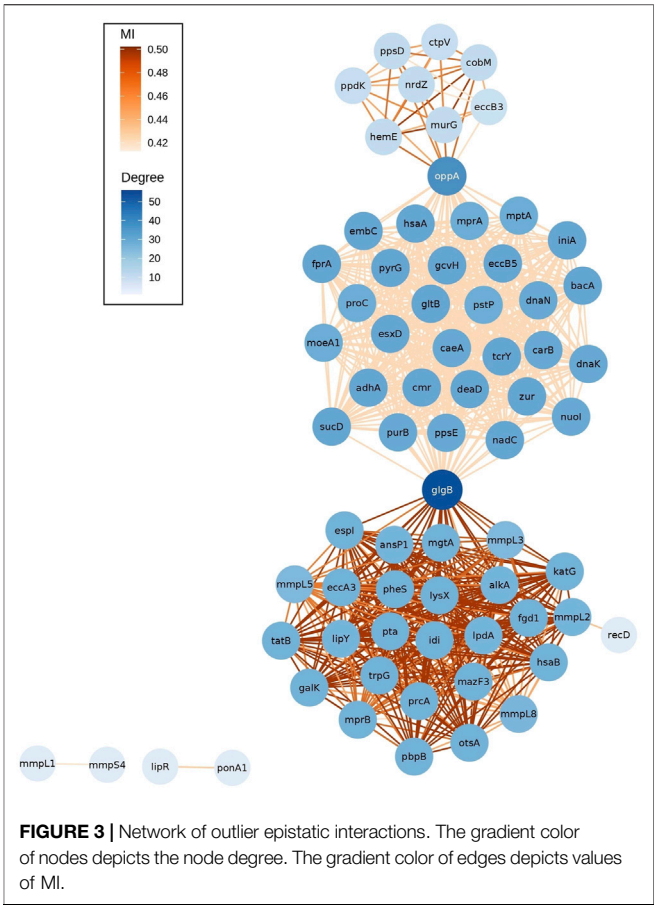


FIGURE 3 | Network of outlier epistatic interactions. The gradient color of nodes depicts the node degree. The gradient color of edges depicts values of MI.

Thus, we consider these two genes as relevant putative targets of co-selection because they may be associated with several genes related to potential pharmacological applications. The GlgB enzyme (encoded by Rv1326c) is the key enzyme involved in the biosynthesis of α-glucan, which plays a significant role in the virulence and pathogenesis of Mtb. Recently, enzymes that participate in the biosynthesis of trehalose have gained major attention as drug targets, especially in Mtb (Dkhar et al., 2015), as capsular polysaccharides of bacteria have been found to modulate the host immune response. The importance of the metabolism of GlgB has been described (De Smet et al., 2000), but the epistatic interactions with other genes remain unknown.

On the other hand, the gene *oppA* (oligopeptide-binding protein) works as a substrate-binding protein for the oligopeptide transport system (Opp), which is responsible for peptide importation. The Opp system is an ATP-binding cassette transporter. This helps in peptide absorption, giving pathogens the essential nutrients as a source of carbon, nitrogen, and amino acids. The Opp system affects many cellular processes, including internalization of quorum-sensing peptides, biofilm production, cell surface modification, and antibiotic resistance (Hopfe et al., 2011). The relevance of the characterization of the peptide transporter system has been described by Dasgupta et al. (2010). Previous studies uncovered the novel observation that this peptide transporter modulates the innate immune response of macrophages infected (Cassio Barreto de Oliveira and Balan,

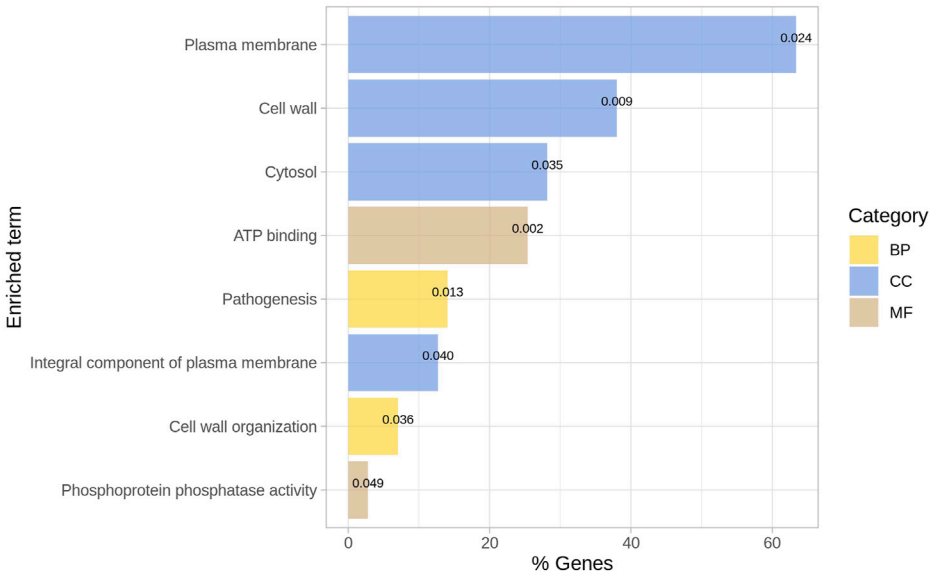


FIGURE 4 | Enriched GO terms for outliers. The *p*-value is indicated for each term. BP = Biological process; CC = Cellular Component; MF = Molecular function.

2020) with Mtb, but the epistatic interactions of *oppA* with other loci remain unknown.

A bacterium is able to adapt its response to host conditions, such as intracellular residence in phagocytic cells, oxidative stress, hypoxia, and carbon and nitrogen source. For this reason, evaluating interactions by bioinformatics experiments is necessary for the identification of new epistatic interactions in genes that have been previously reported in databases, such as the catalog of the WHO, or for the understanding of the epistatic interactions in Mtb before the development of new therapies.

3.4.2 Enriched GO Terms for the Epistatic Network

The list of ENTREZ ids of the genes of the network was used to perform a functional enrichment analysis with DAVID. From the DAVID chart report, we only considered those terms as relevant with p-value < 0.05 (see **Supplementary Data S4** for details of the functional enrichment analysis). Biological processes of pathogenesis (GO:0009405) and cell wall organization (GO: 0071555) were enriched in a subset of genes (**Figure 4**). The cell wall (GO:0005618), plasma membrane (GO:0005886), cytosol (GO:0005829), and integral components of the plasma membrane (GO:0005887) were the more abundant cellular components; in this case, 63% of the genes are in the plasma membrane. Regarding molecular functions, we obtained enrichment for ATP binding (GO:0005524) and phosphoprotein phosphatase activity (GO:0004721) for some genes.

Recently, the biomarkers of Mtb that regulate immune response have been identified to potentially develop drugs for TB. It has been previously described that the functionality of cellular components was associated with infection and verified the regulation of these cellular components as relevant regulators of the immune response in the host (Li et al., 2020). Thus, describing the genes involved in cellular components is crucial for understanding the interactions of bacteria with host molecules that regulate immune response.

In recent studies, the relevance of the structure and biogenesis-related genes of Mtb encoding glycoconjugates has been confirmed, with particular emphasis on the molecules across the different layers of the cell envelope (Angala et al., 2014). In addition, it has been previously stressed that ATP production is crucial for antibiotic resistance in bacteria (Black et al., 2014).

We show the enriched GO terms of genes interacting with *glgB* (**Figure 5B**) and *oppA* (**Figure 5F**) using circular layouts. In addition, **Supplementary Data S5** also contains in table format the genes interacting with *glgB*, their product, and enriched GO terms; the same information is provided for *oppA* in **Supplementary Data S6**. The layouts were generated using the start and end positions of genes reported in partitions generated by AMAS, so the arrangement and size of genes in the layout and positions of interactions reflect the pan–genome-wide alignment. We have highlighted the interactions of *glgB* and *oppA* in red to distinguish them from the interaction of other genes (shown in gray).

Both *glgB* and *oppA* epistatically interact with genes enriched with the biological processes of pathogenesis and cell wall organization (**Figures 5D,H**). One of these genes is *embC*, which codifies for an arabinosyltransferase involved in the biosynthesis of a major component of the mycobacterial cell wall lipoarabinomannan (LAM). The characteristic manosse-capped LAM of Mtb acts as a pathogen-associated molecular pattern (PAMP), modulating the activation of phagocytic cells to control the strength of the host inflammatory immune response, while representing one of the main components in the cell wall organization. In addition, it has been described that *embC* is expressed as part of a polycistronic mRNA controlled by a promoting region differentially expressed depending on the stationary or hypoxia-induced persistence phase of the bacilli,

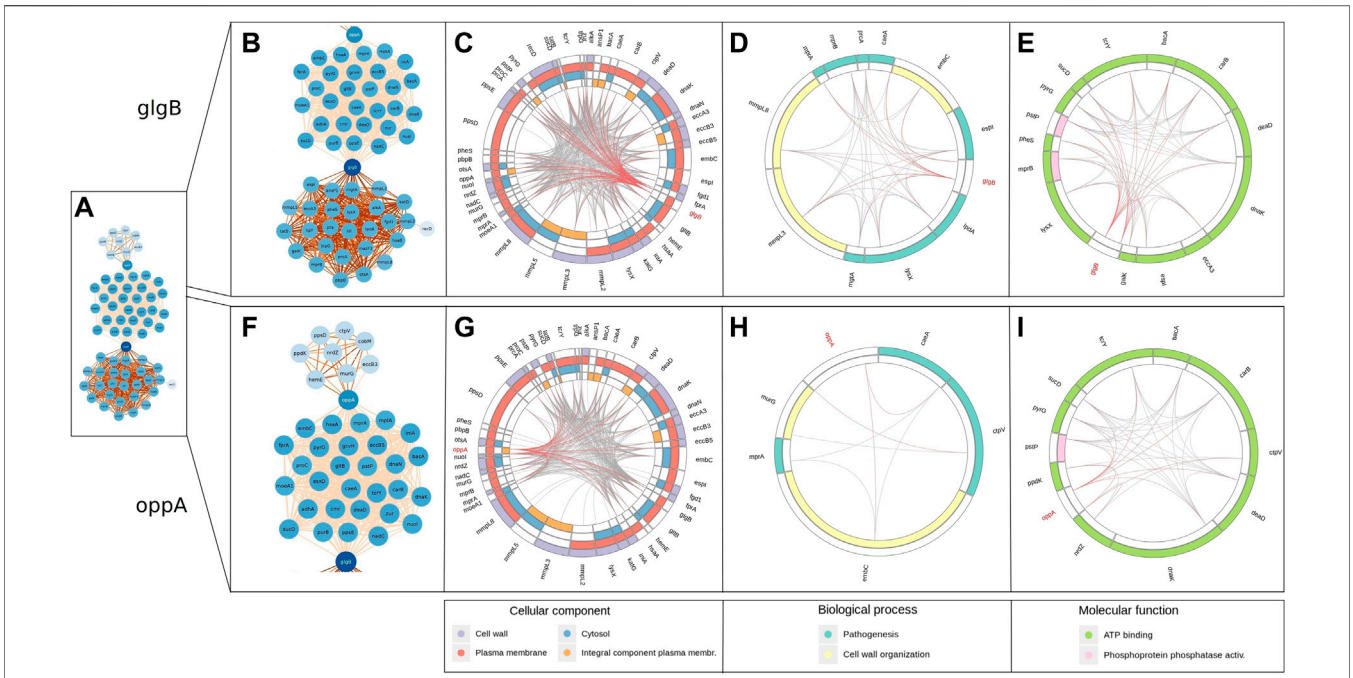


FIGURE 5 | Enriched GO terms of genes that interact with *glgB* and *oppA*. The arrangement and size of genes and positions of interactions reflect the pan–genome-wide alignment. Red lines indicate interactions of *glgB* and *oppA*, whereas gray indicates the interaction of other genes. **(A)** Epistatic network. **(B)** *glgB* interactions (subnet). **(C)** Enriched cellular components of genes interacting with *glgB*. **(D)** Enriched biological processes of genes interacting with *glgB*. **(E)** Enriched molecular functions of genes interacting with *glgB*. **(F)** *oppA* interactions (subnet). **(G)** Enriched cellular components of genes interacting with *oppA*. **(H)** Enriched biological processes of genes interacting with *oppA*. **(I)** Enriched molecular functions of genes interacting with *oppA*.

highlighting the important role of this protein in the biological functions of Mtb and the complex interaction involved in cell wall regulation. Whether or not a direct interaction of *embC* with *glgB* and/or *oppA* exists remains an exciting question to be addressed (Goude et al., 2008).

About molecular functions, *oppA* and *glgB* interact with several genes enriched with ATP binding and with three genes enriched with phosphoprotein phosphatase activity (**Figures 5E, I**). From these genes, *bacA* is another gene found to be interacting at the highest scores with both *glgB* and *oppA* in the network, and *bacA* encodes for a protein of the type IV family of ABC transporter–type exporters; despite the structure, their function as an importer of multi-solute hydrophilic compounds, such as vitamin B12, bleomycin, and aminoglycosides, has been demonstrated due to a large occluded water-filled cavity that spans across the whole lipid membrane. In addition, it has also been demonstrated that this transporter is implicated in the maintenance of chronic infection in murine models by mediating the transport of a molecule that can directly or indirectly modulate the proinflammatory host response. Despite having different structures, BacA and OppA shared their ability to transport a wide range of substrates; in particular, the shared capacity of import peptides related to the innate immune response suggests a complex regulation and interaction of these transporters, guaranteeing the need to carry out studies at the level of gene regulation and function in the near future (Domenech et al., 2009; Cassio Barreto de Oliveira and Balan, 2020; Rempel et al., 2020).

The enriched GO terms that may be related to those associated with ATP synthase in mycobacteria are of particular interest because they contribute to efficient ATP production, and this enzyme has been validated as a target for potential pharmacological applications. In addition, mycobacterial ATP synthase and its characteristics may provide information on adaptations of bacterial energy metabolism. Mtb can survive in human macrophages for an extended time. For Mtb and other pathogenic mycobacteria strains, the blocking of ATP hydrolysis is relevant as it may represent an adaptation to its internal and external human phagosomes, where ATP, once produced, must not be used (Lu et al., 2014). Thus, the importance of epistatic interactions associated with ATP production in pathogenic bacteria may face exceptional challenges as a variety of pathogens need to deal with low energy conditions, such as low oxygen tensions or nutrient limitation inside the host.

Regarding gene interactions enriched with phosphoprotein phosphatase activity, both *oppA* and *glgB* established epistatic interaction with the *pstP* gene, which encodes the Serine/Threonine Protein Phosphatase PstP of Mtb. Signal sensing and transduction *via* phosphorylation and dephosphorylation of specific target proteins are essential for the survival of both eukaryotic and prokaryotic organisms. In the case of Mtb, 11 serine/threonine protein kinases have been described, but only the serine/threonine phosphatase, PstP, has been identified, highlighting the central role of this protein in the control of vital processes as a negative regulator of kinase activity and global serine and threonine phosphorylation (Iswahyudi et al., 2019).

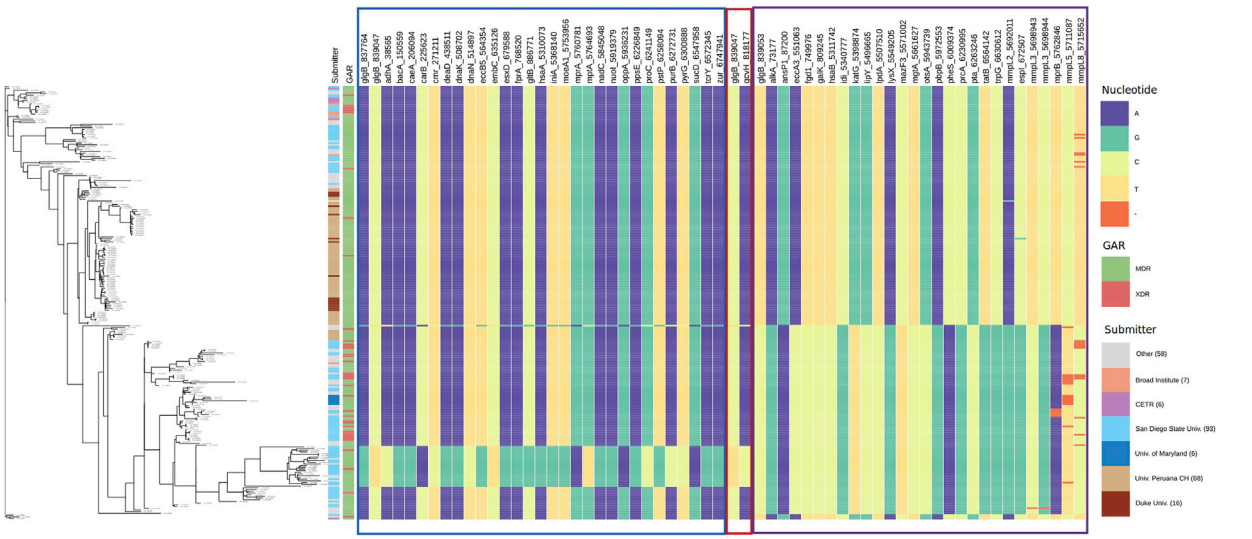


FIGURE 6 | Allele distribution at the loci of genes interacting with *glgB*. Labels of the column indicate gene name and locus. Estimated phylogeny is included on the left. Interacting loci are organized in sections with borders. The border color corresponds to the color of the sections in **Supplementary Data S7**. GAR = Genotypic Antibiotic Resistance.

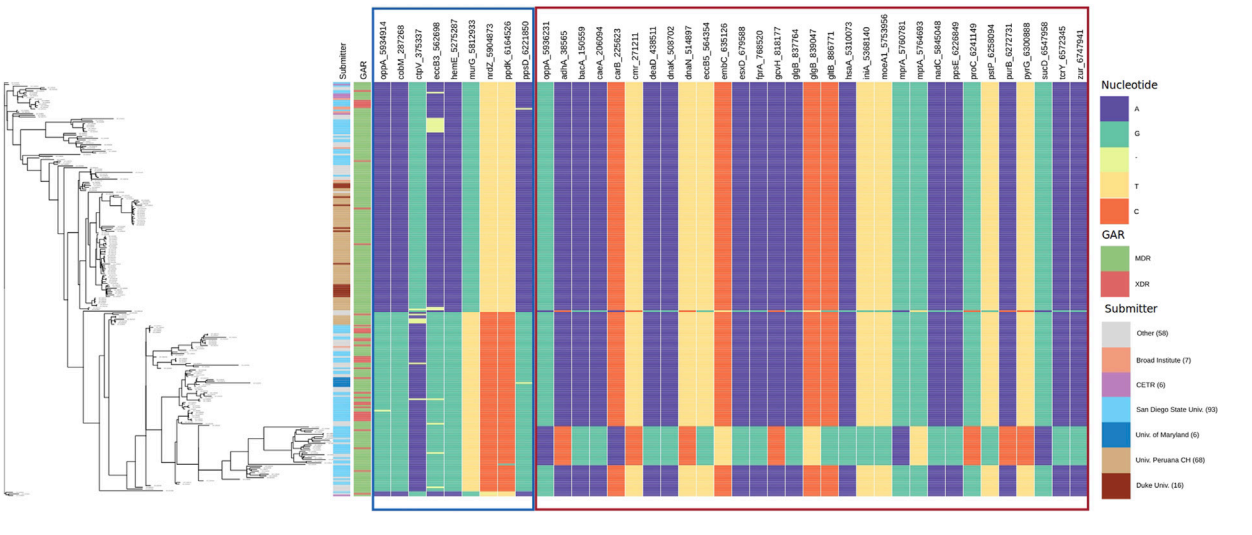


FIGURE 7 | Allele distribution at the loci of genes interacting with *oppA*. Labels of the column indicate gene name and locus. Estimated phylogeny is included on the left. Interacting loci are organized in sections with borders. The border color corresponds to the color of the sections in **Supplementary Data S8**. GAR = Genotypic Antibiotic Resistance.

Recently, other functions of PstP as a regulator of cell wall synthesis and cell division by dephosphorylation of key substrates implicated in both pathways have been described (Sharma et al., 2016). PstP is co-transcribed in an operon with genes involved in peptidoglycan synthesis, with protein kinases PknA and PknB that regulate cell growth and cell division and with *fhaA* and *fhaB*, which encode phosphothreonine recognition proteins that also regulate cell growth and cell division. The involvement of *pstP* with elements necessary for cell wall biosynthesis and their strict dependence on Mn^{2+} for function suggests that the interaction encountered by the computational approach could serve as a starting point for initiating investigations into the molecular interactions that regulate these common processes.

PstP is present as a transmembrane phosphatase and contains a 240–amino acid intracellular catalytic domain, tethered *via* a single transmembrane helix to the 196-amino acid-long extracellular domain (Boitel et al., 2003); it remains to be elucidated if during infection and activation of the innate immune responses (e.g., respiratory burst activation), the degraded bacteria retain the phosphatase activity in the

TABLE 2 | Antibiotic-resistant genes reported by the WHO catalogue are found in the epistatic network.

Gene	Antibiotic-resistant gene (drug)
glgB and *oppA*	*fprA* (AMI)
glgB and *oppA*	*fprA* (CAP)
glgB and *oppA*	*embC* (EMB)
glgB	*katG* (INH)
glgB	*fgd1* (DLM)
glgB	*mmpL5* (BDQ)
glgB	*mmpL5* (CFZ)

[a]*AMI = amikacin; BDQ = bedaquiline; CAP = capreomycin; CFZ = clofazimine; DLM = delamanid; EMB = ethambutol; INH = isoniazid WHO (2021).*

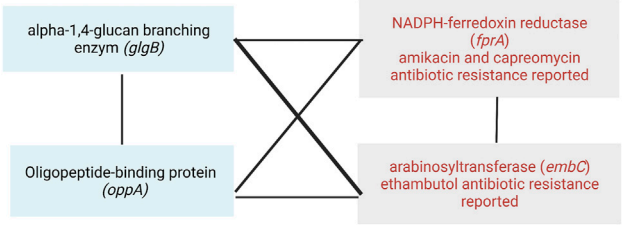


FIGURE 8 | Epistatic interactions between the antibiotic-resistant genes *fprA* and *embC*, and the putative targets of co-selection genes *glgB* and *oppA*.

membrane fragments, and these could contribute to the dephosphorylation of the signaling pathways of the innate system, contributing to the evasion of the immune response.

Four cellular components were enriched for genes interacting with *glgB* (**Figure 5C**) and *oppA* (**Figure 5G**). The majority of genes encode proteins in the plasma membrane. For example, four genes of the group of mycobacterial membrane protein large (MmpL), *mmpL2*, *mmpL3*, *mmpL5,* and *mmpL8* have epistatic interactions with *glgB* and *oppA*. MmpL proteins export cell envelop components (such as virulence-associated lipids and siderophores) to the periplasmic space, contributing at a high level to the persistence of Mtb in the host (Melly and Purdy, 2019). A further study will be required to investigate the fine regulation between the import and export systems of genes identified under epistatic interaction by our approach in order to establish their relevance and biological implications.

3.4.3 Allele Distribution at Loci of Genes Interacting With *glgB* and *oppA*

To observe the patterns of alleles of the interacting polymorphic loci, we show the allele distribution at interacting loci with the loci of *glgB* and *oppA* using Phandango (Hadfield et al., 2017). SpydrPick detected that three loci of *glgB* (837,764, 839,047, and 839,053) interact with 57 polymorphic loci of 56 genes. For *oppA*, two polymorphic loci (5,934,914 and 5,936,231) were found interacting with 38 loci of 37 genes. Tables with loci and genes are available in the **Supplementary Datas S7, S8**.

Interacting loci and gene names are displayed as labels of columns in **Figure 6** for *glgB* and in **Figure 7** for *oppA*. Interacting loci are organized in sections with borders. Each section includes the interactions for each interacting locus. The border color for each section corresponds to the color of the sections in **Supplementary Data S7, S8**. In **Figure 6**, the first section starts with the loci 837,764 and 839,047 of *glgB* (glgB_837764 and glgB_839047) followed by the 29 polymorphic loci that interact with them, that is, these two loci epistatically interact with each one of the 29 loci.

By observing the allele distribution of pairs of loci, we confirm that SpydrPick is able to detect, using the MI score, predictable patterns of alleles at the two loci. See, for example, the second section in **Figure 6**, which only depicts the allele distribution of the locus 839,047 of *glgB* (glgB_839047) and the interacting locus 818,177 of *gcvH* (gcvH_818177). It can be noticed that when there is a C in the locus glgB_839047, there is an A in the locus gcvH_818,177, and when there is a T in the locus glgB_839047, there is a C in the locus gcvH_818,177. This predictability is quantitatively depicted by the value of *MI* = 0.4201. The third section in **Figure 6** exposes the allele distribution of interactions between the locus 839,053 of *glgB* (glgB_839053) and 27 loci of 26 genes (two loci of the gene *mmpL3* interact with the locus glgB_839053). In this section, we observe gaps (-) in the allele distribution of the interacting locus 5,762,846 of *mprB* (two component histidine-protein kinase/phosphatase MprB, *MI* = 0.4743), locus 5,711,087 of *mmpL5* (transmembrane transport protein MmpL5 *MI* = 0.4670) and locus 5,715,652 of *mmpL8* (integral membrane transport protein MmpL8 *MI* = 0.4435).

In **Figure 7**, we present the allele distribution of loci interacting with loci 5,934,914 and 5,936,231 of the gene *oppA*. The first section contains the interactions with the locus oppA_5,934,914. We notice the presence of gaps in the positions ctpV_375337 (*MI* = 0.4516) and eccB3_562,698 (*MI* = 0.417). **Figure 7** also presents well-defined patterns of allele distribution between interacting loci.

The first column in both figures indicates the submitter institution (we included institutions with less than six submitted strains in the category other), and the second column points to genotypic antibiotic resistance (GAR) predicted with RGI. Estimated phylogeny is included on the left to show the diversity of the strain collection that we analyzed. For example, a clade at the bottom of the tree stands out due to its change of nucleotide in relation to the rest of the strains.

The application of this methodology also allowed the identification within the network of various loci in genes associated with resistance. Recently, the WHO published the first catalog of resistance-associated genetic variants for predicting relevant resistance phenotypes based on more than 38,000 WGS phenotyped isolates (WHO, 2021). This has allowed the identification of multiple positions associated with resistance and their classification into five groups. We use this recent classification to find antibiotic-resistant genes in the epistatic network (**Table 2**).

In our network, an interaction was found between *glgB* and *katG*; mutations conferring monoresistance to isoniazid (INH) are common due to INH having been in clinical use since the 1950s. Nevertheless, INH resistance testing is only recently included in some specialized cartridges (e.g., Xpert MTB/XDR) and is not routinely available in such a way that if INH resistance

is not detected, patients are treated as pan-suceptible, which represents a high risk of treatment failure and a greater propensity to acquire further resistance (Sulis and Pai, 2020).

From the antibiotic-resistance genes that we found in the WHO catalog, we observed that *fprA* (resistant to amikacin and capreomycin) and *embC* (resistant to ethambutol) interact with both genes *glgB* and *oppA* (**Table 2**). Moreover, they interact between them, forming a clique of four genes (**Figure 8**). A clique depicts a network where all nodes are fully connected to each other, creating a strong interaction mechanism. This kind of epistatic interactions motivated us to visualize future studies to test new experimental hypotheses to elucidate their biological and pharmacological explanations, and the MI score seems to be a very successful approach to drive so.

4 CONCLUSION

Here, we have presented the reconstruction and analysis of an epistatic network for Mtb from a pan–genome-wide alignment by using the model-free method SpydrPick. Our approach allowed us identifying new epistatic interactions with implications in virulence, pathogenesis, transport system modulators of the immune response, and genotypic antibiotic resistance. By the analysis of the epistatic network, we identified *glgB* and *oppA* as putative targets of co-selection. These two genes epistatically interact with *fprA* and *embC*, two antibiotic-resistant genes reported in the catalog of the WHO, as resistant to ethambutol (*embC*) and amikacin and capreomycin (*fprA*). Our results highlight the importance of implementing computational approaches to elucidate new genes associated to putative epistatic interactions in Mtb.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**; further inquiries can be directed to the corresponding authors. R scripts are available at https://github.com/biotb/epitb-net.

AUTHOR CONTRIBUTIONS

ABPR and SFC conceptualized, designed, and conducted the investigation. ABPR analyzed the data and generated the epistatic network. YBM, SAR, PV, and SFC supervised the research. All authors wrote the manuscript. ABPR and SFC take responsibility for the manuscript. All authors read and approved the final content.

FUNDING

This research was partially supported by the Programa Interno de Apoyo para Proyectos de Investigación, FESC, UNAM (PIAPI 2018).

ACKNOWLEDGMENTS

ABPR is a doctoral student from the *Programa de Doctorado en Ciencias Biológicas, Universidad Nacional Autónoma de México* (UNAM) and received a fellowship from CONACyT (631936). PV gratefully acknowledges support by CONACyT A1-S-11242 and PAPIIT-UNAM IN209321. ABPR gratefully acknowledge the *Programa de Ingeniería Genómica*, CCG-UNAM, for the support during her stay (February–December 2019). We acknowledge the support of Miguel Ángel Pérez León of the *Laboratorio de cómputo de alto rendimiento, Coordinación del departamento de Matemáticas* and of the Lansbiodyt (*Laboratorio Nacional de Soluciones Biomiméticas para Diagnóstico y Terapia*), *Facultad de Ciencias, UNAM*. **Figures 1**, **8** were created with BioRender.com.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.856212/full#supplementary-material

REFERENCES

Alcock, B. P., Raphenya, A. R., Lau, T. T. Y., Tsang, K. K., Bouchard, M., Edalatmand, A., et al. (2020). Card 2020: Antibiotic Resistome Surveillance with the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* 48, D517–D525. doi:10.1093/nar/gkz935

Angala, S. K., Belardinelli, J. M., Huc-Claustre, E., Wheat, W. H., and Jackson, M. (2014). The Cell Envelope Glycoconjugates of Mycobacterium Tuberculosis. *Crit. Rev. Biochem. Mol. Biol.* 49, 361–399. doi:10.3109/10409238.2014.925420

Black, P. A., Warren, R. M., Louw, G. E., van Helden, P. D., Victor, T. C., and Kana, B. D. (2014). Energy Metabolism and Drug Efflux in Mycobacterium Tuberculosis. *Antimicrob. Agents Chemother.* 58, 2491–2503. doi:10.1128/aac.02293-13

Boitel, B., Ortiz-Lombardía, M., Durán, R., Pompeo, F., Cole, S. T., Cerveñansky, C., et al. (2003). Pknb Kinase Activity Is Regulated by Phosphorylation in Two Thr Residues and Dephosphorylation by Pstp, the Cognate Phospho-Ser/thr Phosphatase, in mycobacterium Tuberculosis. *Mol. Microbiol.* 49, 1493–1508. doi:10.1046/j.1365-2958.2003.03657.x

Borowiec, M. L. (2016). Amas: a Fast Tool for Alignment Manipulation and Computing of Summary Statistics. *PeerJ* 4, e1660. doi:10.7717/peerj.1660

Cassio Barreto de Oliveira, M., and Balan, A. (2020). The Atp-Binding Cassette (Abc) Transport Systems in Mycobacterium Tuberculosis: Structure, Function, and Possible Targets for Therapeutics. *Biology* 9, 443. doi:10.3390/biology9120443

Chanda, P., Costa, E., Hu, J., Sukumar, S., Van Hemert, J., and Walia, R. (2020). Information Theory in Computational Biology: Where We Stand Today. *Entropy* 22, 627. doi:10.3390/e22060627

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., et al. (1998). Deciphering the Biology of mycobacterium Tuberculosis from the Complete Genome Sequence. *Nature* 393, 537–544. doi:10.1038/31159

Coscolla, M., and Gagneux, S. (2014). Consequences of Genomic Diversity in mycobacterium Tuberculosis. *Semin. Immunol.* 26, 431–444. doi:10.1016/j.smim.2014.09.012

Cover, T. M., and Thomas, J. (2006). *Elements of Information Theory*. New Jersey: Wiley-Interscience.

Dasgupta, A., Sureka, K., Mitra, D., Saha, B., Sanyal, S., Das, A. K., et al. (2010). An Oligopeptide Transporter of Mycobacterium Tuberculosis Regulates Cytokine

Release and Apoptosis of Infected Macrophages. *PloS one* 5, e12225. doi:10.1371/journal.pone.0012225

De Smet, K. A. L., Weston, A., Brown, I. N., Young, D. B., and Robertson, B. D. (2000). Three Pathways for Trehalose Biosynthesis in Mycobacteria. *Microbiology* 146, 199–208. doi:10.1099/00221287-146-1-199

Dkhar, H. K., Gopalsamy, A., Loharch, S., Kaur, A., Bhutani, I., Saminathan, K., et al. (2015). Discovery of *Mycobacterium tuberculosis* α-1,4-Glucan Branching Enzyme (GlgB) Inhibitors by Structure- and Ligand-Based Virtual Screening. *J. Biol. Chem.* 290, 76–89. doi:10.1074/jbc.m114.589200

Domenech, P., Kobayashi, H., LeVier, K., Walker, G. C., and Barry, C. E. (2009). BacA, an ABC Transporter Involved in Maintenance of Chronic Murine Infections with *Mycobacterium tuberculosis*. *J. Bacteriol.* 191, 477–485. doi:10.1128/JB.01132-08

Drost, H. G., and Paszkowski, J. (2017). Biomartr: Genomic Data Retrieval with R. *Bioinformatics* 33, 1216–1217. doi:10.1093/bioinformatics/btw821

Gonzalo-Asensio, J., Malaga, W., Pawlik, A., Astarie-Dequeker, C., Passemar, C., Moreau, F., et al. (2014). Evolutionary History of Tuberculosis Shaped by Conserved Mutations in the Phopr Virulence Regulator. *Proc. Natl. Acad. Sci. U.S.A.* 111, 11491–11496. doi:10.1073/pnas.1406693111

Goude, R., Amin, A. G., Chatterjee, D., and Parish, T. (2008). The Critical Role of Embc in mycobacterium Tuberculosis. *J. Bacteriol.* 190, 4335–4341. doi:10.1128/jb.01825-07

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C., and Paola, G. d. (1985). Acnuc – A Portable Retrieval System For Nucleic Acid Sequence Databases: Logical And Physical Designs And Usage. *Bioinformatics* 1, 167–172. doi:10.1093/bioinformatics/1.3.167

Gupta, N., Kumar, R., and Agrawal, B. (2018). New Players in Immunity to Tuberculosis: the Host Microbiome, Lung Epithelium, and Innate Immune Cells. *Front. Immunol.* 9, 709. doi:10.3389/fimmu.2018.00709

Hadfield, J., Croucher, N. J., Goater, R. J., Abudahab, K., Aanensen, D. M., and Harris, S. R. (2017). Phandango: an Interactive Viewer for Bacterial Population Genomics. *Bioinformatics* 34, 292–293. doi:10.1093/bioinformatics/btx610

Hopfe, M., Dahlmanns, T., and Henrich, B. (2011). In Mycoplasma Hominis the Oppa-Mediated Cytoadhesion Depends on its Atpase Activity. *BMC Microbiol.* 11, 185–189. doi:10.1186/1471-2180-11-185

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and Integrative Analysis of Large Gene Lists Using David Bioinformatics Resources. *Nat. Protoc.* 4, 44–57. doi:10.1038/nprot.2008.211

Iswahyudi, Mukamolova, G. V., Straatman-Iwanowska, A. A., Allcock, N., Ajuh, P., Turapov, O., et al. (2019). Mycobacterial Phosphatase Pstp Regulates Global Serine Threonine Phosphorylation and Cell Division. *Sci. Rep.* 9, 8337–8411. doi:10.1038/s41598-019-44841-9. https://www.nature.com/articles/s41598-019-44841-9#citeas.

Jiao, X., Sherman, B. T., Huang, D. W., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2012). DAVID-WS: a Stateful Web Service to Facilitate Gene/Protein List Analysis. *Bioinformatics* 28, 1805–1806. doi:10.1093/bioinformatics/bts251

Kavvas, E. S., Catoiu, E., Mih, N., Yurkovich, J. T., Seif, Y., Dillon, N., et al. (2018). Machine Learning and Structural Analysis of mycobacterium Tuberculosis Pan-Genome Identifies Genetic Signatures of Antibiotic Resistance. *Nat. Commun.* 9, 4306–4309. doi:10.1038/s41467-018-06634-y

Kohl, M., Wiese, S., and Warscheid, B. (2011). "Cytoscape: Software for Visualization and Analysis of Biological Networks," in *Data Mining in Proteomics* (Springer), 291–303. doi:10.1007/978-1-60761-987-1_18

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: a Fast, Scalable and User-Friendly Tool for Maximum Likelihood Phylogenetic Inference. *Bioinformatics* 35, 4453–4455. doi:10.1093/bioinformatics/btz305

Lerner, T. R., Borel, S., and Gutierrez, M. G. (2015). The Innate Immune Response in Human Tuberculosis. *Cell Microbiol* 17, 1277–1285. doi:10.1111/cmi.12480

Letunic, I., and Bork, P. (2021). Interactive Tree of Life (iTOL) V5: an Online Tool for Phylogenetic Tree Display and Annotation. *Nucleic Acids Res.* 49, W293–W296. doi:10.1093/nar/gkab301

Li, L., Lv, J., He, Y., and Wang, Z. (2020). Gene Network in Pulmonary Tuberculosis Based on Bioinformatic Analysis. *BMC Infect. Dis.* 20, 612–711. doi:10.1186/s12879-020-05335-6

Lu, P., Lill, H., and Bald, D. (2014). Atp Synthase in Mycobacteria: Special Features and Implications for a Function as Drug Target. *Biochim. Biophys. Acta (Bba) - Bioenerg.* 1837, 1208–1218. doi:10.1016/j.bbabio.2014.01.022

Melly, G., and Purdy, G. (2019). Mmpl Proteins in Physiology and Pathogenesis of m. tuberculosis. *Microorganisms* 7, 70. doi:10.3390/microorganisms7030070

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., et al. (2015). Roary: Rapid Large-Scale Prokaryote pan Genome Analysis. *Bioinformatics* 31, 3691–3693. doi:10.1093/bioinformatics/btv421

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., et al. (2016). Snp-sites: Rapid Efficient Extraction of Snps from Multi-Fasta Alignments. *Microb. Genomics* 2, e000056. doi:10.1099/mgen.0.000056

Pattengale, N. D., Alipour, M., Bininda-Emonds, O. R. P., Moret, B. M. E., and Stamatakis, A. (2010). How many Bootstrap Replicates Are Necessary? *J. Comput. Biol.* 17, 337–354. doi:10.1089/cmb.2009.0179

Pensar, J., Puranen, S., Arnold, B., MacAlasdair, N., Kuronen, J., Tonkin-Hill, G., et al. (2019). Genome-Wide Epistasis and Co-Selection Study Using Mutual Information. *Nucleic Acids Res.* 47, e112. doi:10.1093/nar/gkz656

Rempel, S., Gati, C., Nijland, M., Thangaratnarajah, C., Karyolaimos, A., de Gier, J. W., et al. (2020). A Mycobacterial Abc Transporter Mediates the Uptake of Hydrophilic Compounds. *Nature* 580, 409–412. doi:10.1038/s41586-020-2072-8

Sackton, T. B., and Hartl, D. L. (2016). Genotypic Context and Epistasis in Individuals and Populations. *Cell* 166, 279–287. doi:10.1016/j.cell.2016.06.047

Satta, G., Lipman, M., Smith, G. P., Arnold, C., Kon, O. M., and McHugh, T. D. (2018). *Mycobacterium tuberculosis* and Whole-Genome Sequencing: How Close Are We to Unleashing its Full Potential? *Clin. Microbiol. Infect.* 24, 604–609. doi:10.1016/j.cmi.2017.10.030

Seemann, T. (2014). Prokka: Rapid Prokaryotic Genome Annotation. *Bioinformatics* 30, 2068–2069. doi:10.1093/bioinformatics/btu153

Sharma, A. K., Arora, D., Singh, L. K., Gangwal, A., Sajid, A., Molle, V., et al. (2016). Serine/threonine Protein Phosphatase Pstp of mycobacterium Tuberculosis Is Necessary for Accurate Cell Division and Survival of Pathogen. *J. Biol. Chem.* 291, 24215–24230. doi:10.1074/jbc.m116.754531

Skwark, M. J., Croucher, N. J., Puranen, S., Chewapreecha, C., Pesonen, M., Xu, Y. Y., et al. (2017). Interacting Networks of Resistance, Virulence and Core Machinery Genes Identified by Genome-wide Epistasis Analysis. *Plos Genet.* 13, e1006508. doi:10.1371/journal.pgen.1006508

Sulis, G., and Pai, M. (2020). Isoniazid-resistant Tuberculosis: A Problem We Can No Longer Ignore. *Plos Med.* 17, e1003023. doi:10.1371/journal.pmed.1003023

Trauner, A., Borrell, S., Reither, K., and Gagneux, S. (2014). Evolution of Drug Resistance in Tuberculosis: Recent Progress and Implications for Diagnosis and Therapy. *Drugs* 74, 1063–1072. doi:10.1007/s40265-014-0248-y

Tukey, J. (1977). *Exploratory Data Analysis*. Reading, Massachusetts: Addison-Wesley.

WHO (2021). *Catalogue of Mutations in mycobacterium Tuberculosis Complex and Their Association with Drug Resistance*. Geneva: World Health Organization.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Posada-Reyes, Balderas-Martínez, Ávila-Ríos, Vinuesa and Fonseca-Coronado. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 12 July 2022
doi: 10.3389/fmolb.2022.849363

Mapping CAR T-Cell Design Space Using Agent-Based Models

Alexis N. Prybutok[1], *Jessica S. Yu*[1,2], *Joshua N. Leonard*[1,3,4,5]* *and Neda Bagheri*[1,2,3,4,6]*

[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, United States, [2]Department of Biology, University of Washington, Seattle, WA, United States, [3]Center for Synthetic Biology, Northwestern University, Evanston, IL, United States, [4]Chemistry of Life Processes Institute, Northwestern University, Evanston, IL, United States, [5]Robert H. Lurie Comprehensive Cancer Center, Northwestern University, Evanston, IL, United States, [6]Department of Chemical Engineering, University of Washington, Seattle, WA, United States

Chimeric antigen receptor (CAR) T-cell therapy shows promise for treating liquid cancers and increasingly for solid tumors as well. While potential design strategies exist to address translational challenges, including the lack of unique tumor antigens and the presence of an immunosuppressive tumor microenvironment, testing all possible design choices *in vitro* and *in vivo* is prohibitively expensive, time consuming, and laborious. To address this gap, we extended the modeling framework ARCADE (Agent-based Representation of Cells And Dynamic Environments) to include CAR T-cell agents (CAR T-cell ARCADE, or CARCADE). We conducted *in silico* experiments to investigate how clinically relevant design choices and inherent tumor features—CAR T-cell dose, $CD4^+$:$CD8^+$ CAR T-cell ratio, CAR-antigen affinity, cancer and healthy cell antigen expression—individually and collectively impact treatment outcomes. Our analysis revealed that tuning CAR affinity modulates IL-2 production by balancing CAR T-cell proliferation and effector function. It also identified a novel multi-feature tuned treatment strategy for balancing selectivity and efficacy and provided insights into how spatial effects can impact relative treatment performance in different contexts. CARCADE facilitates deeper biological understanding of treatment design and could ultimately enable identification of promising treatment strategies to accelerate solid tumor CAR T-cell design-build-test cycles.

Keywords: agent-based model, CAR T-cell, simulation, cell population dynamics, model-guided design, emergent dynamics

OPEN ACCESS

Edited by:
Ernesto Perez-Rueda, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas. Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Haopeng Wang, ShanghaiTech University, China
Jun Huang, The University of Chicago, United States

***Correspondence:**
Neda Bagheri
nbagheri@uw.edu
Joshua N. Leonard
j-leonard@northwestern.edu

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *06 January 2022*
Accepted: *23 May 2022*
Published: *12 July 2022*

Citation:
Prybutok AN, Yu JS, Leonard JN and Bagheri N (2022) Mapping CAR T-Cell Design Space Using Agent-Based Models. Front. Mol. Biosci. 9:849363. doi: 10.3389/fmolb.2022.849363

1 INTRODUCTION

Chimeric antigen receptor (CAR) T-cell therapy combines advances in cellular engineering and personalized medicine for patient-specific, targeted cancer treatment (Barrett et al., 2014; Jackson et al., 2016). This therapy involves collecting, purifying, and genetically modifying a patient's own T-cells to express a CAR that specifically targets the patient's tumor(s) (Barrett et al., 2014; Jackson et al., 2016). These engineered cells are expanded *ex vivo* and then re-infused into the patient where the CAR T-cells target and kill antigen-expressing tumor cells. The six FDA-approved CAR T-cell therapies and many studies expanding CAR designs exclusively target "liquid" cancers that typically derive from $CD19^+$ B-cells (Jackson et al., 2016; Castellarin et al., 2018; Yanez-Munoz and Grupp, 2018; NCI, 2022). CD19 CAR T-cell therapies have shown great success in the clinic with response rates between 70–90% reported (Lim and June 2017). In contrast, response rates for solid cancers are significantly lower at 4–16% (Hou et al., 2019).

CAR T-cells are currently less effective for treating solid tumors due unique complexities of both the tumor microenvironment (TME) and tumors themselves. First, TME barriers prevent CAR T-cell infiltration (Castellarin et al., 2018). These barriers include the intricate influence of both tumor-suppressing and tumor-promoting cells on the TME (Whiteside, 2008; Galluzzi et al., 2018), immune-evading cell markers promoting tumor escape (Maus and June 2016; Galluzzi et al., 2018), and physical and chemical barriers that impact spatial dynamics and nutrient availability (Whiteside, 2008; Castellarin et al., 2018). Thus, developing CAR T-cells that remodel the immunosuppressive TME has been an active area of research (Cherkassky et al., 2016; Liu et al., 2016; Lim and June 2017; Huang et al., 2018; Rafiq et al., 2018). Second, solid tumors often lack unique tumor antigens for selective targeting (Kakarla and Gottschalk, 2014). Cross-reactivity with healthy tissues present harmful or fatal off-tumor effects (Bonifant et al., 2016; Lim and June 2017). Cellular engineering efforts have focused on increasing CAR specificity by tuning the affinity of receptor-antigen interactions to avoid healthy cells (Caruso et al., 2015; Johnson et al., 2015; Liu et al., 2015; Castellarin et al., 2018). Similarly, creating CAR T-cells that perform Boolean logic can enhance tumor recognition specificity (Wilkie et al., 2012; Lanitis et al., 2013; Wu et al., 2015; Castellarin et al., 2018; Cho et al., 2018). Designing CAR T-cells that target multiple antigens simultaneously can also prevent formation of antigen escape variant tumors (Hegde et al., 2013; Hegde et al., 2016; Cho et al., 2018). Finally, additional factors that have not proven problematic for "liquid" cancers, such as the need for site-specific trafficking of CAR T-cells to solid tumors and tumor antigen heterogeneity, further complicate solid-tumor CAR T-cell therapy design (Lim and June, 2017).

In combination with the array of engineering design choices presented by addressing the constraints above, additional design choices impact CAR T-cell effector functions and long-term persistence regardless of tumor type. These features include $CD4^+$:$CD8^+$ CAR T-cell ratios (Zhao et al., 2015; Sommermeyer et al., 2016; Turtle et al., 2016), choice of intracellular co-stimulatory domain (ICD) in the CAR (Kawalekar et al., 2016; Guedan et al., 2018), and the stage of T-cell differentiation (Sommermeyer et al., 2016). Collectively, the vast number of design choices complicates interpreting and comparing studies of and iteratively tuning CAR T-cell therapies.

Simultaneously tuning multiple features of a CAR T-cell therapy and forecasting their impact on emergent population dynamics remains a grand challenge. Exploring the multidimensional design space becomes prohibitively expensive and laborious *in vitro* and *in vivo*, particularly when considering the time and resources required for mouse experiments. Additionally, some design aspects and emergent properties are difficult to interrogate experimentally, such as cell-level behavioral states that impact treatment efficacy. Employing *in silico* experiments has proven to be a resource-saving and valuable way to understand how underlying biological processes impact CAR treatment outcome and hypothesizing new design features to improve efficacy. Recent CAR T-cell modeling efforts have used ordinary differential equation (ODE) models to understand factors influencing CAR T-cell receptor signaling and downstream activation (Rohrs et al., 2018; Cess and Finley, 2020a; Rohrs et al., 2020). Other CAR T-cell ODE modeling efforts aim to optimize patient pre-conditioning with chemotherapy (Owens and Bozic, 2021). However, these models lack spatial resolution, test a limited set of features, and do not assess emergent cell population dynamics; these important contributions do not yet enable predictions of the sort needed to guide the design of CAR T-cell therapies.

Agent-based models (ABMs) provide ideal *in silico* testbeds for interrogating emergent population dynamics. ABMs are bottom-up computational frameworks that describe the behavior of autonomous agents through defined rules that guide agent actions and interactions within their local environment. The ABM framework provides single-cell spatial and temporal resolution, incorporates quantitative and qualitative experimental observations, and enables tuning and measuring properties of interest through *in silico* experiments (Chavali et al., 2008; Narang et al., 2012; Yu and Bagheri, 2016; Vodovotz et al., 2017). Past ABMs have explored how cell properties influence tumor growth (Zhang et al., 2009; Waclaw et al., 2015; Norton et al., 2017; Yu and Bagheri, 2020), vasculature and microenvironment dynamics (Anderson et al., 2006; Yu and Bagheri, 2021), immune response to infection and tumors (Folcik et al., 2007; Cess and Finley, 2020b), and tumor response to checkpoint inhibitor therapy (Gong et al., 2017). However, to our knowledge, no ABM reported to date has characterized CAR T-cell dynamics in solid tumors or explored how CAR T-cell and tumor features impact outcomes.

In this study, we systematically explore CAR T-cell therapy designs in solid tumor contexts by adding CAR T-cell agents to an established ABM (Agent-based Representation of Cells And Dynamic Environments, or ARCADE) comprising tissue cell agents (Yu and Bagheri, 2020) and dynamic vasculature (Yu and Bagheri, 2021). We use this model—CAR T-cell ARCADE (CARCADE)—to simulate CAR T-cell interactions with tissue cells and analyze a multidimensional design space. We demonstrate that CARCADE recapitulates known observations and predicts responses to new designs for solid tumor CAR T-cell therapies.

2 RESULTS

2.1 CARCADE Characterizes CAR T-Cell Behavior, Metabolism, and Effector Function

CARCADE provides a flexible framework for characterizing and exploring hypothesized dynamics of population-level tumor responses to CAR T-cell treatment by defining individual CAR T-cell, cancer, and healthy cell features and rules.

2.1.1 CAR T-Cell Agents Recapitulate CAR T-Cell Behavior

ARCADE comprises tissue cell agents with individual subcellular metabolism and signaling modules that influence the cell-level

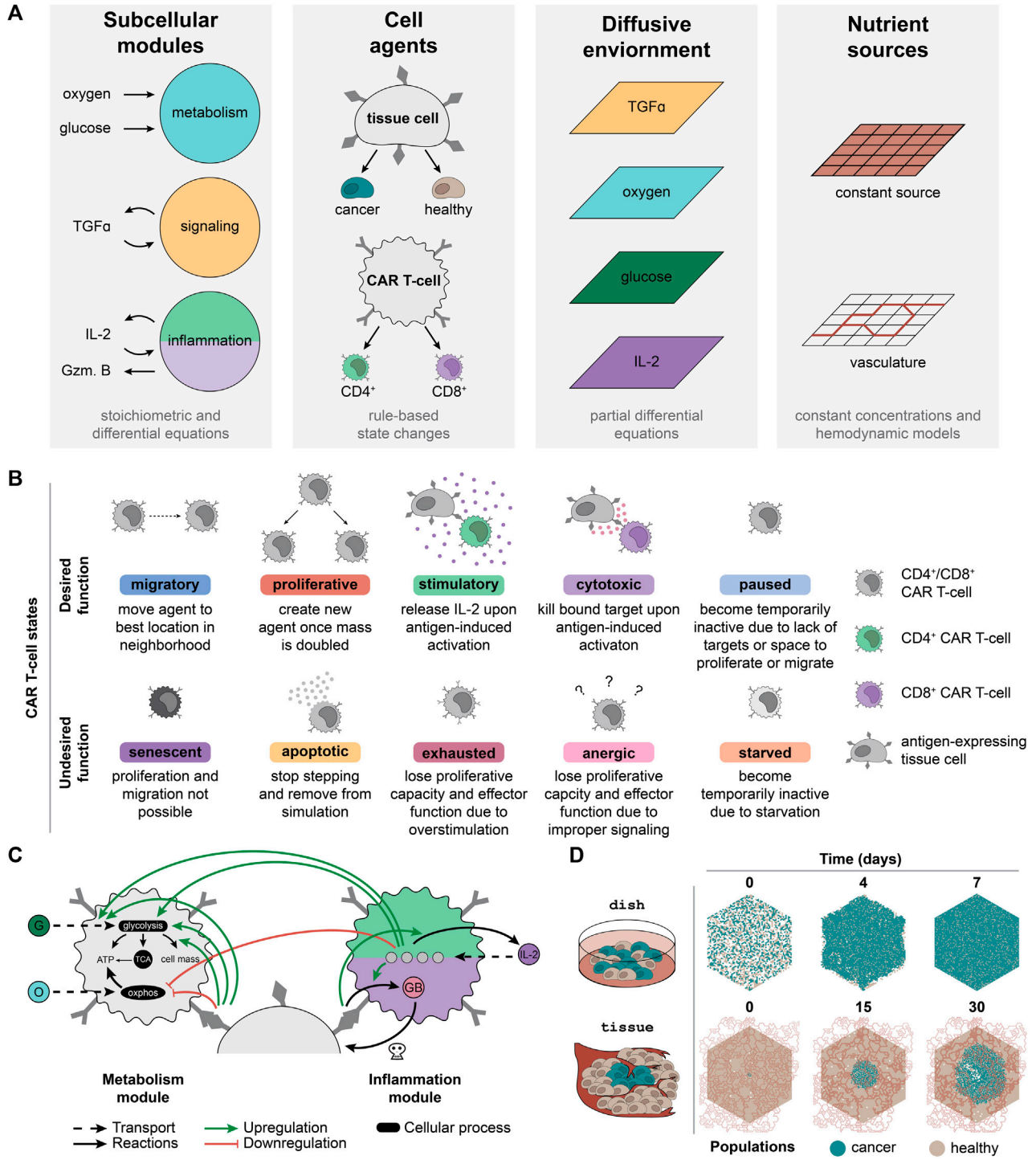


FIGURE 1 | CARCADE structure and CAR T-cell agent design. **(A)** Depiction of CARCADE components. Subcellular modules guide underlying cellular function to influence behavior (Gzm. B, granzyme B). Agents include tissue cell and CAR T-cell agents, each of which has separate rule sets and is depicted with surface ligands and CARs (dark gray). Tissue cells include both healthy cells and cancer cells. Agents exist in an environment where diffusion is controlled by partial differential equations and constant sources or vasculature provide nutrients. **(B)** Descriptions of each CAR T-cell agent state, separated by whether the state is desired or undesired for efficacious treatment. **(C)** Diagram of CAR T-cell metabolism and inflammation module interactions with small molecules, proteins, and regulatory edges. The inflammation module diagram is broken into two parts, showing differences between $CD4^+$ CAR T-cells (light green, top) and $CD8^+$ CAR T-cells (purple, bottom). All CAR T-cells use identical metabolism modules. Regulatory edges (upregulation: green arrow, downregulation: red flathead arrow) result from IL-2 binding and antigen-induced activation. G, glucose; O, oxygen; GB, granzyme B; OXPHOS, oxidative phosphorylation. Legend for cell color is consistent with panel B. **(D)** An example of a single `dish` and `tissue` simulation of untreated cancer and healthy cells shown at select time points. For `tissue`, the dynamic vasculature architecture is overlaid.

decision making rules and drive emergent population- and environment-level dynamics (**Figure 1A**). Tissue cell agent rules and parameters can be tuned to represent either cancer or healthy cells. We introduce a new cell agent representing CAR T-cells into this framework (**Figure 1A**). All cell agents are simulated in a microenvironment that comprises either constant nutrient sources (representing a dish context) or vasculature (representing a vascularized tissue context). To distinguish between simulation and experiment, we denote simulated dish and tissue contexts as `dish` and `tissue`, respectively.

Agents navigate through a set of defined, cell-type specific states and rules derived from experimentally observed states and transitions. Each tissue cell can be in one of six states—migratory, proliferative, quiescent, senescent, necrotic, and apoptotic—at each time step. CAR T-cell agents follow a unique rule set with additional states designed to capture T-cell behaviors (**Figure 1B**). There are two subtypes of CAR T-cell agents: $CD8^+$ T-cells that primarily provide cytotoxic functions and $CD4^+$ T-cells that primarily provide stimulatory functions (Liadi et al., 2015; Golubovskaya and Wu, 2016; Sommermeyer et al., 2016). Although both T-cell subtypes can provide cytotoxic and stimulatory functions, for simplicity, we specified that each of these T-cell subtypes would perform only their primary function. CAR T-cell agents can enter ten different subtype-dependent states, broadly categorized as desirable and undesirable during treatment. Desired states include migratory, proliferative, stimulatory ($CD4^+$ only), cytotoxic ($CD8^+$ only), and paused. Undesired states include apoptotic, senescent, exhausted, anergic, and starved. Cells change state according to the rule set and to their current state (**Supplementary Figures S1, S2**, **Supplementary Methods Details**). All new model parameters are listed in **Supplementary Table S1** (Kuse et al., 1985; Lauffenburger and Linderman, 1993; Robertson et al., 1996; Frauwirth et al., 2002; De Boer et al., 2003; Deenick et al., 2003; Iwashima, 2003; Jacobs et al., 2008; Busse et al., 2010; Yoon et al., 2010; Wang et al., 2011; Altman and Dang, 2012; Robertson-Tessi et al., 2012; Stone et al., 2012; Cheng et al., 2013; Hegde et al., 2013; Heskamp et al., 2015; Kinjyo et al., 2015; Liu et al., 2015; Obst, 2015; Harris and Kranz, 2016; Hegde et al., 2016; Arcangeli et al., 2017; Borghans and Ribeiro, 2017; Gong et al., 2017; Gherbi et al., 2018; Guedan et al., 2018; Salter et al., 2018; Yu and Bagheri, 2020; 2021).

Each agent utilizes subcellular modules to capture underlying metabolic and signaling states. ARCADE tissue agents use two subcellular modules that control metabolism and signaling. The metabolism module uses stoichiometric equations to determine cellular uptake of glucose and oxygen, which is then converted to energy and cell mass. The signaling module uses an ODE model with regulatory nodes to determine the influence of tumor growth factor alpha (TGFα) on a tissue cell's decision to proliferate or migrate. CAR T-cell agents use the tissue cell metabolism module with modifications to capture the influence of IL-2 signaling and antigen-induced activation on T-cell metabolism: 1) increased metabolic preference for glycolysis; 2) increased glucose uptake rate; and 3) increased fraction of glucose used to produce cell mass (**Figure 1C**, **Supplementary Methods Details**) (Frauwirth et al., 2002; Jones and Thompson, 2007; Pearce, 2010; Altman and Dang, 2012; Gerriets and Rathmell, 2012; MacIver et al., 2013; Chang and Pearce, 2016; Mehta et al., 2017). CAR T-cell agents also contain an inflammation module to capture the impact of IL-2 binding and antigen-induced activation on IL-2 production in $CD4^+$ CAR T-cells (Malek and Castro, 2010; Liao et al., 2013; Rosenberg, 2014) and on granzyme production in $CD8^+$ CAR T-cells (Liadi et al., 2015) (**Figure 1C**, **Supplementary Methods Details**). For both CAR T-cell subtypes, the inflammation module uses an ODE model to determine the amount of IL-2 bound to various IL-2 receptor species (Malek and Castro, 2010; Liao et al., 2013; Ross and Cantrell, 2018).

2.1.2 *In Silico* Experiments Mimic *In Vitro* and *In Vivo* Contexts

To provide an *in silico* testbed that can be related to physical experiments, simulations were designed to represent two experimental contexts: `dish` and `tissue` (**Figure 1D**). Each configuration utilizes an environment in which four nutrient and signaling molecules—oxygen, glucose, TGFα, and IL-2—diffuse. Additionally, the environment contains distinct sources from which oxygen and glucose are produced. `Dish` uses a constant nutrient source environment to represent the well-mixed cell media of an *in vitro* experiment. These simulations are initialized with a defined number of tissue cells placed randomly in the environment. CAR T-cells are introduced after 10 min and simulated for 7 d of treatment. `Tissue` uses vasculature to represent realistic hemodynamics of nutrients diffusing through the environment to represent an *in vivo* solid tumor experiment. Vasculature can be degraded and collapse due to cancer cell crowding and movement. These simulations are initialized with a confluent bed of healthy cells and a small colony of cancer cells added to the center of the simulation environment. The cancer cell colony grows for 21 d to form a tumor before CAR T-cells are added and simulated for 9 d of treatment. Untreated `dish` and `tissue` simulations highlight how *in silico* experimental design leads to diverse outcomes (**Figure 1D**).

2.2 Monoculture and Co-Culture Simulations are Consistent With *In Vitro* Observations

CAR T-cell agents were developed *de novo* based on established cell-level observations; resulting emergent dynamics of the simulation were used for model validation. The comparison between *in silico* and *in vitro/in vivo* experiments is a critical and common method for validating ABMs. To confirm that emergent dynamics follow experimental observations, we tested how outcomes vary as a function of four CAR and tumor features—CAR T-cell dose (Sampson et al., 2014), $CD4^+$:$CD8^+$ CAR T-cell ratio (Sommermeyer et al., 2016; Turtle et al., 2016), CAR-antigen affinity (Chmielewski et al., 2004; Hudecek et al., 2013; Caruso et al., 2015; Johnson et al., 2015; Liu et al., 2015; Ghorashian et al., 2019), and antigen density on

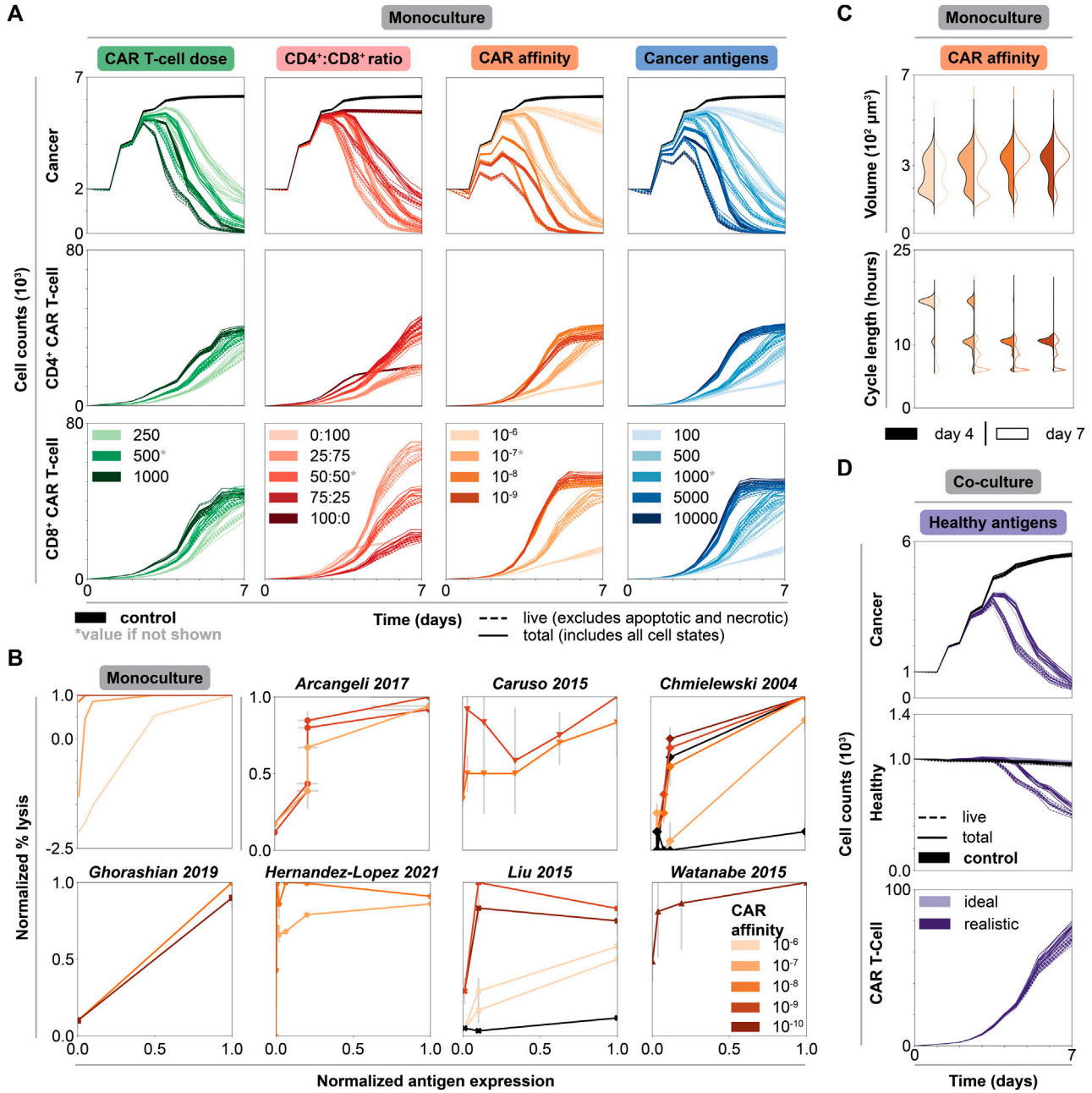


FIGURE 2 | Impact of individual CAR T-cell and tumor features on cytotoxicity and CAR T-cell growth in `dish`. **(A)** Cell counts over time of untreated (black) and treated conditions (graded hues) holding all but one feature constant. Each column shows the axis being changed, where all other features are held constant at indicated intermediate values (indicated by asterisk, CAR T-cell dose = 500 CAR T-cells, $CD4^+$:$CD8^+$ ratio = 50:50, CAR affinity = 10^{-7} M, cancer antigens = 1000 antigens/cell), while rows show the cell type being plotted. **(B)** Normalized percent lysis curves for *in silico* and published experimental *in vitro* data. Plot for simulated data shows percent lysis for each set of CAR affinity values across normalized cancer antigen values. All other axes were held constant, and the data were averaged across replicates. Simulations with negative percent lysis indicate cancer cell growth. Experimental data—representing an array of CAR types, effector to target (E:T) ratios, ICDs, and cancer cell lines (**Supplementary Table S6**, **Supplementary Data S5**)—were normalized to maximum percent lysis and antigen levels with estimated error bars. The plots show percent lysis for each set of CARs tested per paper, each with unique CAR affinity and tested across a range of antigen target values. **(C)** Volume and cell cycle distributions for CAR T-cell populations at t = 4 d (filled) and t = 7 d (outline) holding all but CAR affinity constant at an intermediate value in monoculture. Legend is consistent with panel B. The data for cancer cell populations and for all other features can be found in the **Supplementary Material**. **(D)** Cell counts over time of untreated (black) and treated conditions (graded hues) holding all features constant at an intermediate value. Legend is consistent with panel B for both ideal and realistic co-culture. Solid lines represent total cell counts, dashed lines represent live cell counts.

cancer cells (Stone et al., 2012; Liu et al., 2015; Watanabe et al., 2015; Majzner et al., 2020).

In a clinical setting, CAR T-cells necessarily interact with both healthy and cancer cells, and healthy cell antigen expression can impact off-target effects (Harris and Kranz, 2016). It is critical to consider how these CAR and tumor features impact both cancer and healthy cell populations. We simulated CAR T-cell treatment in three different contexts—1) monoculture with only cancer cells, 2) ideal co-culture with cancer cells and antigen-negative healthy cells, and 3) realistic co-culture with cancer cells and low-level antigen expressing healthy cells—modulating CAR T cells and tumor features in each context to assess how *in silico* dynamics compare to observations *in vitro*. Using `dish` removes confounding effects of nutrient constraints and TME factors. We simulated 10 replicates of each combination of features (**Supplementary Table S2** for monoculture, **Supplementary Table S3** for co-culture). In monoculture, `dish` was randomly plated at t = 0 s with 2×10^3 antigen-expressing cancer cells. At t = 10 min, treatment begins by adding a dose of CAR T-cells, each expressing 5×10^4 CARs with a defined CAR affinity and $CD4^+$:$CD8^+$ ratio. We simulated 7 d of treatment. Co-culture is identical except initial plating uses 1×10^3 cancer cells and 1×10^3 healthy cells. Simulation trajectories—including each cell's location, state, volume, and average cell cycle length—were collected every half day. The input files used to generate `dish` simulations are described in the Supplementary Material (**Supplementary Data S1** and **Supplementary Table S4** for monoculture, **Supplementary Data S2** and **Supplementary Table S5** for co-culture).

2.2.1 Cancer Cell and CAR T-Cell Dynamics Are Independent of Context

We first consider the impact of individual features on cell counts and behavior in `dish` (holding other features constant at intermediate values). In all simulations, cancer cell and CAR T-cell counts follow experimentally observed trends, including conditions with effector-to-target (E:T) ratios less than one where cancer cell killing occurs over several days (**Figure 2A** for monoculture, **Supplementary Figure S3A** for ideal co-culture, **Supplementary Figure S3B** for realistic co-culture) (Chmielewski et al., 2004; Arcangeli et al., 2017). Increasing CAR T-cell dose increases T-cell counts and accelerates cancer cell killing (Sampson et al., 2014; Hamieh et al., 2019). Our simulations mirror this trend; when E:T ratios are increased beyond the initial range explored (i.e., to explore ratios greater than one), substantial cancer cell killing occurred in monoculture in half the time (all other features are held at intermediate values) (**Supplementary Data S3**, **Supplementary Figure S4A**). Increasing the E:T ratio brings closer parity in rate of cancer cell killing between our simulations and experimental analyses, but we acknowledge that there remains a discrepancy based on time to complete elimination of cancer cells. This difference can be attributed to unaccounted for contact-independent mechanisms of killing, potentially including exosomes (Fu et al., 2019); these additional mechanisms were not included in the model for simplicity. Intermediate $CD4^+$:$CD8^+$ ratios maximize cancer killing and increase CAR T-cell proliferation (Sommermeyer et al., 2016; Turtle et al., 2016). Higher fractions of $CD8^+$ CAR T-cell treatments prove less effective because cytotoxic $CD8^+$ cells need the support of the cytokines primarily produced by $CD4^+$ cells (Liadi et al., 2015; Golubovskaya and Wu, 2016). We tested an expanded range of $CD4^+$:$CD8^+$ ratios to include 90:10 and 10:90 in monoculture and co-culture; these extensions further validated observed trends and provided no additional treatment benefit, and thus we do not carry these conditions forward in subsequent analyses (see **Supplementary Note S1**, **Supplementary Data S4**, and **Supplementary Figure S5**). Increasing CAR affinity increases the chances of CAR T-cell antigen binding and subsequent activation, resulting in increased cancer cell killing (Chmielewski et al., 2004; Liu et al., 2015; Hernandez-Lopez et al., 2021). This increased activation also leads to increased proliferation and thus increased T-cell count (Caruso et al., 2015). Increased antigen expression on cancer cells increases cancer cell killing (Chmielewski et al., 2004; Liu et al., 2015; Watanabe et al., 2015; Arcangeli et al., 2017). Similarly, because CAR T-cells are more likely to be activated by high antigen density cancer cells, CAR T-cell proliferation, and thus counts, increase with increasing antigen count (Hernandez-Lopez et al., 2021). $CD8^+$ T-cells counts exceed $CD4^+$ T-cell counts even when cells are delivered at a 50:50 ratio, especially in conditions where cells are more likely to be activated (Sommermeyer et al., 2016; Turtle et al., 2016). The lowest CAR T-cell counts occur when we treat with only one subset of CAR T-cells. Cancer cells cannot be killed off without $CD8^+$ cells. $CD8^+$ cells have limited killing and proliferative capacity without cytokines produced by $CD4^+$ cells, and lack of cancer cell killing presents spatial limitations on CAR T-cell proliferation. Overall, all `dish` simulations, regardless of healthy cell context, support experimental observations of cancer and CAR T-cell dynamics, suggesting that healthy cell presence and antigen expression do not strongly influence cancer and CAR T-cell dynamics or individual feature trends *in vitro*.

2.2.2 Monoculture Data Qualitatively Recapitulate a Range of *In Vitro* CAR T-Cell Studies

Quantifying percent lysis as a function of cancer antigen density is a common experimental analysis. In monoculture, percent lysis increases as a function of both antigen count and CAR affinity. This qualitative trend and the general shape of the data agrees with prior *in vitro* observations (**Figure 2B**) (Chmielewski et al., 2004; Caruso et al., 2015; Liu et al., 2015; Watanabe et al., 2015; Arcangeli et al., 2017; Ghorashian et al., 2019; Hernandez-Lopez et al., 2021). Additionally, for monoculture and most *in vitro* data, higher CAR affinities promote higher percent lysis across all antigen expression values. Our simulations reproduce general trends observed across diverse *in vitro* studies varying in CAR, intracellular co-stimulatory domain, effector to target ratio, and cell lines (**Supplementary Table S6**, **Supplementary Data S5**). Notably, CARCADE captures known experimental trends relevant to many different experimental CAR T-cell scenarios without being trained to any specific CAR T-cell experiment. Consistency in these emergent dynamics provides

baseline validation that supports our use of the model to interrogate CAR T-cell design.

2.2.3 Trends in Cell-Level Features Support Population-Level Observations and Model Validation

Treatment efficacy can be evaluated by volume (Jacobs et al., 2008) and cell cycle length (Yoon et al., 2010) distributions, which serve as proxies for CAR T-cell growth and proliferation resulting from antigen-induced activation and IL-2 binding. As an increasing number of CAR T-cells undergo antigen-induced activation, CAR T-cell volumes increase and cycle lengths decrease both over time and with increasing CAR affinity (**Figure 2C**). In T-cells, antigen-induced activation and IL-2 binding influence metabolism to help T-cells rapidly proliferate by increasing nutrient uptake, metabolic preference for glucose, and flux of nutrients towards producing cell mass (Frauwirth et al., 2002; Jacobs et al., 2008; Pearce, 2010; Altman and Dang, 2012; Gerriets and Rathmell, 2012; Buck et al., 2015; Chang and Pearce, 2016; Mehta et al., 2017). These internal cellular changes increase cell growth rates, increase volumes, and decrease cell cycle lengths (van Stipdonk et al., 2003; Jacobs et al., 2008; Yoon et al., 2010; Altman and Dang, 2012; Kinjyo et al., 2015). The cell cycle length observed *in silico*—an emergent property of the simulations—ranged from around 6–24 h and falls within the range of 2–24 h found *in vitro, in vivo,* and for other *in silico* models (De Boer et al., 2003; van Stipdonk et al., 2003; Yoon et al., 2010; Altman and Dang, 2012; Kinjyo et al., 2015; Gong et al., 2017). Cancer cell volumes increase slightly and cycle lengths decrease slightly with increasing CAR affinity and over time, as cancer cells proliferate to compensate for cell death (**Supplementary Figure S6A**). Similar trends in volume and cell length distributions are observed across all other modulated features, where conditions with more activated CAR T-cells result in increased CAR T-cell volume and decreased cell cycle lengths (**Supplementary Figure S6B–D**). Altogether, the model recapitulates known *in vitro* observations, and furthermore, it enables us to observe single cell-level properties that are non-trivial to measure experimentally.

2.3 Varying Individual Features Highlights Tradeoffs Within Co-culture

Due to the lack of unique tumor antigens, CAR T-cell designs must rely on target antigens that are more highly expressed on cancer cells than healthy cells (Harris and Kranz, 2016). Investigating the difference in treatment outcomes—cancer cell killing, healthy cell sparing, and CAR T-cell growth—between the ideal co-culture (containing antigen-negative healthy cells) and realistic co-culture (containing antigen-expressing healthy cells) is critical for understanding successful CAR T-cell design (Caruso et al., 2015).

2.3.1 Healthy Cell Antigen Expression and Tumor/CAR T-Cell Features Impact Healthy Cell Killing

Healthy cell antigen density does not affect cancer cell killing, CAR T-cell proliferation, or previously noted trends across individual features for these populations (**Supplementary Figure S3**). However, healthy cell antigen density dramatically impacts healthy cell killing (**Figure 2D**) (Arcangeli et al., 2017). The seeming lack of influence that minimal healthy antigen expression has on CAR T-cell proliferation is demonstrated by a lack of clear difference in CAR T-cell volume and cell cycle length distributions (**Supplementary Figure S7**) or fraction of cells in the proliferative state (**Supplementary Figure S8**) between the ideal and realistic co-culture. In general, we hypothesize that the low healthy cell antigen level is too weak to impact these other factors but enables the CAR T-cells to target healthy cells. Thus, healthy cell antigen expression only needs to be considered in avoiding healthy cell death and not in tuning CAR T-cell behavior or cancer cell killing.

To further investigate the impact of healthy cell antigen expression on feature trends, we directly compare cell counts between the ideal and realistic co-culture along the CAR affinity feature axis (**Figure 3A**). Cancer cell killing dynamics are nearly identical in both contexts, increasing with increased CAR affinity. In contrast, healthy cell dynamics differ dramatically between contexts. When healthy cells do not express antigen, increasing CAR affinity leads to increased healthy cell count as healthy cells grow to fill the space left behind by targeted cancer cells. However, when healthy cells do express antigen, healthy cell killing increases with increasing CAR affinity. Additionally, healthy cell counts begin to decrease at increasingly earlier time points with increasing CAR affinity. Comparing cell counts along other features exhibits similar trends: presence of healthy cell antigen generally only impacts healthy cell dynamics, resulting in varying degrees of healthy cell killing (**Supplementary Figure S3**). These data are consistent with experimental studies demonstrating a detrimental effect of high CAR affinity designs on healthy cells (Caruso et al., 2015; Harris and Kranz, 2016). Low affinity CARs successfully target tumors that overexpress the desired antigen and produce minimal off-tumor effects when healthy cells express low antigen levels (Caruso et al., 2015; Johnson et al., 2015; Liu et al., 2015). When healthy cells express antigen, it is not always desirable to have the strongest affinity CAR T-cells.

2.3.2 Cell Dynamics Reveal Potential New Treatment Strategy That Spares Healthy Cells

Comparing trends in cell dynamics between ideal and realistic co-culture provides insight as to why each feature differentially impacts healthy cell killing. In ideal co-culture, increasing CAR affinity and cancer antigen expression level leads to healthy cell growth beyond their original numbers. Increasing CAR T-cell dose and $CD4^+$:$CD8^+$ ratio leads to healthy cell counts similar to those in the untreated control (**Supplementary Figure S3**). Interestingly, increasing CAR affinity results in more healthy cell growth compared to the case in which cancer cell antigen expression is increased. We hypothesize that this difference occurs because cancer cell killing is more strongly impacted by CAR affinity than cancer antigen density, providing healthy cells more opportunity to grow as more cancer cells die. However, in realistic co-culture, increasing cancer antigen level results in more healthy cell growth before being killed off compared to the

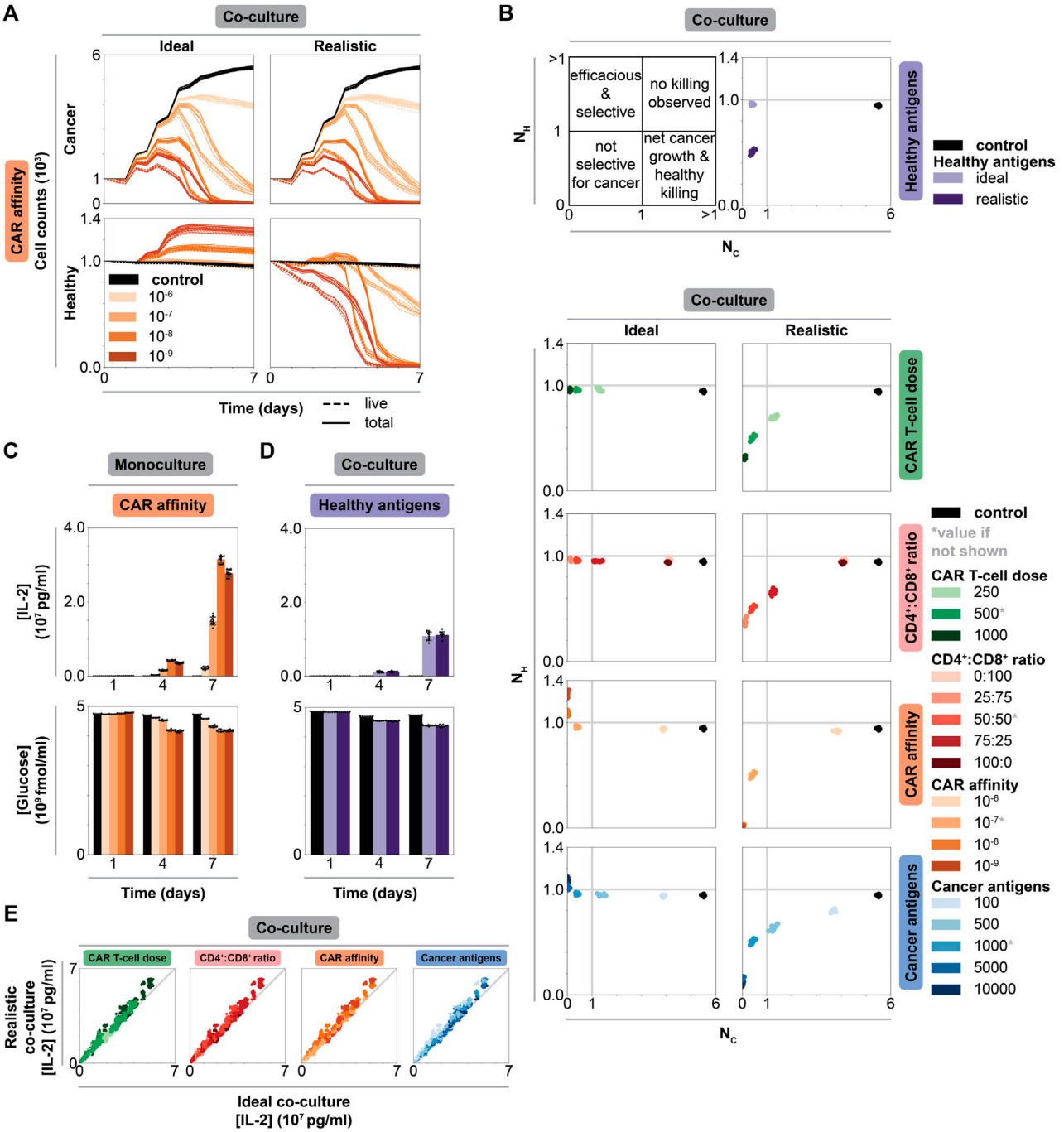


FIGURE 3 | Impact of individual CAR T-cell and tumor features on efficacy, selectivity, and cytokine production in monoculture *vs* co-culture. **(A)** Cancer and healthy cell counts over time of untreated (black) and treated (graded hues) conditions holding all but CAR affinity, which is reported in units of M, constant at an intermediate value and separating data by co-culture context. Column shows co-culture type, row shows cell type. Solid lines represent total cell counts, dashed lines represent live cell counts (live, excludes necrotic and apoptotic states as in **Figure 2**). Intermediate values of other features indicated by asterisk in panel B: CAR T-cell dose = 500 CAR T-cells, $CD4^+$:$CD8^+$ ratio = 50:50, cancer antigens = 1000 antigens/cell. **(B)** Scatter plots of normalized live healthy cell count (N_H) *vs* normalized live cancer cell count (N_C) for untreated (black) and treated conditions (graded hues) holding all but one axis constant at an intermediate value. Upper left plot shows quadrant meanings. Upper right plot shows scatter plot for different co-culture contexts. Columns show co-culture type, and each row indicates which feature is being plotted. **(C)** IL-2 and glucose concentrations over time holding all but CAR affinity constant at an intermediate value in monoculture. Legend is consistent with panel B. **(D)** IL-2 and glucose concentrations over time varying CAR affinity while holding all features constant at an intermediate value in ideal and realistic co-culture. Legend is consistent with panel B. **(E)** Parity plot of IL-2 concentration at final time point (t = 7 d) for all conditions in realistic (*y*-axis) *vs* ideal (*x*-axis) co-culture colored by each feature (column). Legend is consistent with panel B.

scenario in which CAR affinity is increased. Cancer antigen expression primarily impacts cancer cell killing, which gives healthy cells the ability to grow before being targeted after cancer cell populations decline. Meanwhile, CAR affinity impacts both cancer and healthy cell killing, so healthy cells are killed at the same time as cancer cells. These data highlight how each feature differentially impacts the dynamics of this system. A large difference in cancer and healthy cell antigen levels can create a time delay between when cancer killing completes and when healthy cell killing starts, whereas tuning CAR affinity cannot create such a window. This time delay is an emergent phenomenon that occurs in some scenarios—it is not a trained, optimized, or hard-wired parameter in the model. One can design a strategy to take advantage of this time delay in scenarios when it occurs, for example, by deactivating CAR T-cells with an antibody or small-molecule induced off-switch that shuts down effector function after cancer cells are killed but before lower antigen expressing healthy cells are targeted.

2.3.3 Individual Feature Analysis Highlights Tradeoffs in a Pareto Curve

To quantify cancer and healthy cell killing, we use two metrics: normalized live healthy and cancer cell counts. The normalized count for each population (N_P) is calculated as follows:

$$N_P = \frac{n_F}{n_T}$$

where n_F and n_T are the total number of live cancer or healthy cells at the final (t = 7 d) and treatment start (t = 0 d) timepoints, respectively. Values below one indicate net killing, and values above one indicate net growth. Together, these metrics place treatment outcomes within quadrants that can be used as guidelines for classifying efficacy (**Figure 3B**). Ideally, treatment conditions would appear in the upper left quadrant with maximal healthy cell sparing and maximal cancer cell killing. In both contexts, the trends match those of experimental observations—more aggressive treatments with more overall killing result from increasing CAR T-cell dose, intermediate $CD4^+$:$CD8^+$ ratio, increasing CAR-antigen affinity, and increasing cancer antigen density. These conditions allow for healthy cell maintenance or growth in ideal co-culture, nearing or entering the efficacious and selective treatment quadrant. However, in realistic co-culture, there exists a dramatic tradeoff between cancer cell killing and healthy cell killing, presenting a Pareto curve across each feature. Aggressive treatments exist toward the lower left quadrant (not selective for cancer cells). This observation suggests that it is not possible to optimize both efficacy and safety when healthy cells express antigen, and the most useful strategies—typically less aggressive treatments—balance these objectives (Caruso et al., 2015; Johnson et al., 2015; Liu et al., 2015).

2.4 IL-2 Production Is More Strongly Impacted by Tuned Features Than Context

IL-2 production is a standard *in vitro* measurement to quantify T-cell activation (Liu et al., 2015; Sommermeyer et al., 2016; Arcangeli et al., 2017). Similarly, glucose consumption can quantify T-cell activation through nutrient usage and competition (Frauwirth et al., 2002). We compare nutrient consumption and cytokine production across features and contexts to identify strategies for understanding, and potentially controlling, IL-2 production.

2.4.1 Tuning CAR Affinity Modulates IL-2 Production by Balancing CAR T-Cell Proliferation and Effector Function

In `dish` (**Figure 3C** and **Supplementary Figure S9A** for monoculture, **Supplementary Figure S10A** for co-culture), IL-2 increases over time and with increasing values of CAR T-cell dose, $CD4^+$:$CD8^+$ ratio, CAR affinity, and cancer antigen expression level due to increased numbers of activated $CD4^+$ CAR T-cells. Across all contexts and features, glucose decreases as IL-2 increases, indicating that glucose consumption follows CAR T-cell activation and proliferation (**Figure 3D**, **Supplementary Figure S9B**, **Supplementary Figure S10B**).

Unintuitively, IL-2 concentration is not maximized at the highest CAR-antigen affinity in monoculture where CAR T-cell activation is maximized. At the highest CAR affinity, more CAR T-cells spend time in effector, non-proliferative states (**Supplementary Figure S11**), resulting in fewer total $CD4^+$ T-cells producing IL-2 (**Figure 2A**). This decrease is not observed in co-culture where cancer cell numbers are lower, reducing the likelihood that CAR T-cells will be activated. Decreased activation in co-culture produces lower IL-2 concentrations compared to monoculture. Thus, CAR T-cells in co-culture remain outside of the regime at which this tradeoff between activated and proliferating T-cells is observed. We hypothesize that maximum IL-2 production occurs at intermediate CAR affinity where there exists a balance between proliferation and frequent antigen binding. Excessively high CAR affinity leads to frequent target antigen binding, causing CAR T-cells to spend more time in effector rather than proliferating states, leading to fewer total CAR T-cells that can later produce cytokines. On the other hand, very weak affinity CARs drive cells primarily into states other than proliferative and effector states. Maximizing CAR-antigen affinity can therefore prove counterproductive for achieving CAR T-cell proliferation, survival, and cytokine production at the tumor site; moderate CAR-antigen affinities may be more effective.

2.4.2 IL-2 Production Is Independent of Healthy Cell Antigen Expression

In co-culture, healthy antigen expression minimally impacts IL-2 production and glucose consumption over time (**Figure 3D**). We speculate that healthy antigen expression is too low to strongly impact CAR T-cell proliferation and thus IL-2 production. Comparing final IL-2 concentration in all ideal versus realistic co-culture conditions reveal that IL-2 levels are independent of context for a given condition, further supporting this hypothesis (**Figure 3E**). CAR T-cell IL-2 production and overall glucose consumption are more

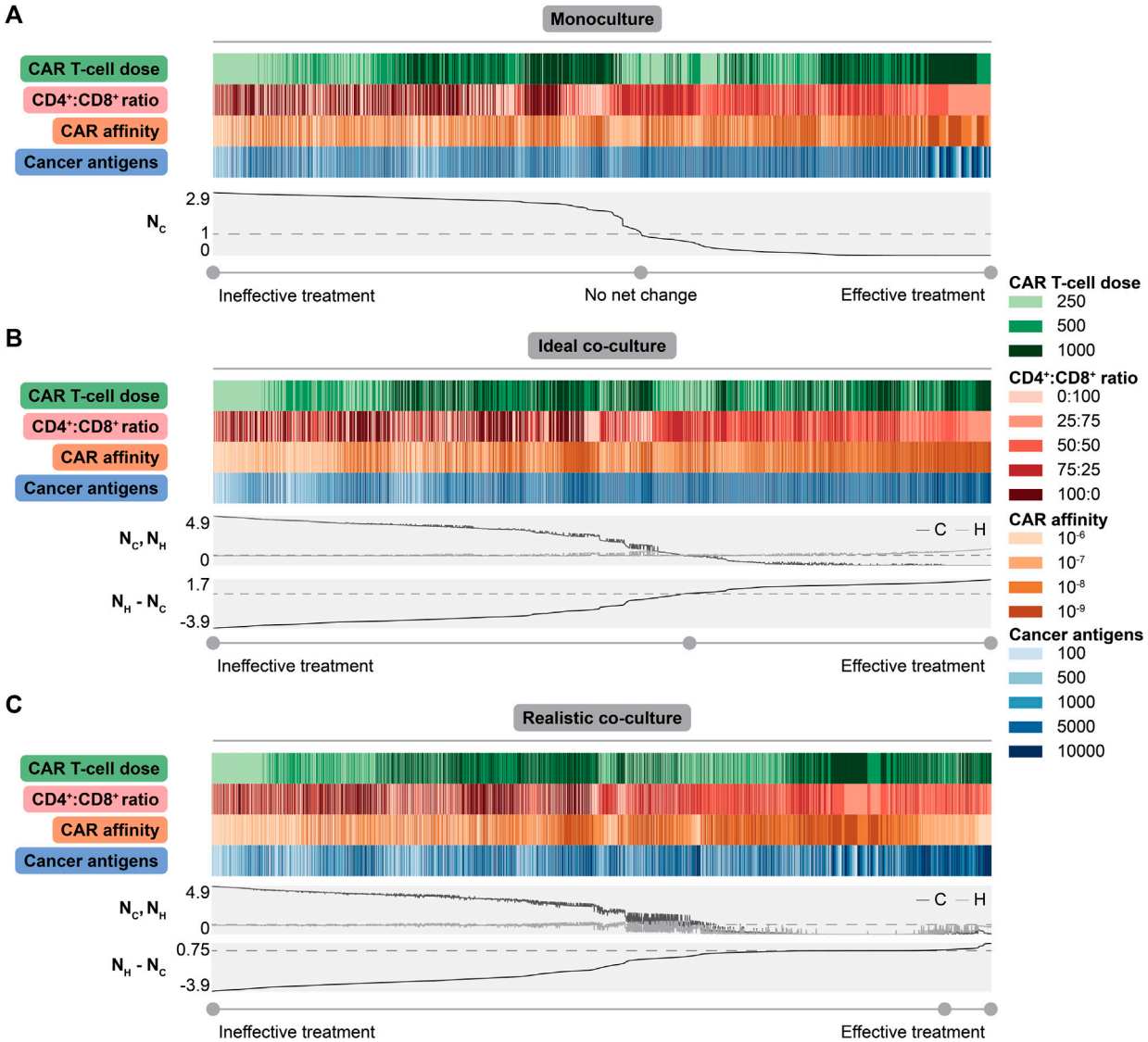


FIGURE 4 | Collective impact of CAR T-cell and tumor features on `dish` outcomes. **(A)** Heatmap showing values for each feature with line plots showing normalized live cancer cell count (N_C) sorted from highest (left) to lowest (right). The dashed line indicates value of $N_C = 1$, meaning no net change due to treatment. Values of $N_C > 1$ indicate net growth and values of $N_C < 1$ indicate net killing. **(B)** Heatmap showing values for each feature with line plots showing normalized live cancer cell count (N_C) and normalized live healthy cell count (N_H) (dashed line indicates normalized live cell count of 1) and the difference in normalized live healthy and cancer cell counts ($N_H - N_C$) for each ideal co-culture simulation individually (dashed line indicates $N_H - N_C = 0$). The heatmap has been sorted from lowest (left) to highest (right) difference. All metrics were calculated at the final time point (t = 7 d). **(C)** Heatmap and normalized cell counts for realistic co-culture. Labels are consistent with panel B. Each feature is reported in the following units: CAR T-cell dose = number of CAR T-cells, $CD4^+$:$CD8^+$ ratio = unitless, CAR affinity = M, cancer antigens = antigens/cell.

strongly impacted by the higher level of antigen expression on the cancer cells than by the low antigen expression on healthy cells. When considering desired IL-2 levels produced by CAR T-cells in patient treatment, IL-2 production can be mostly attributed to and designed around cancer cells in isolation as healthy cell antigen expression does have a significant impact.

2.5 Multidimensional Data Analysis Reveals Context-Specific Treatment Strategies

Since tuning individual features has different impacts on treatment efficacy based on the type of `dish`, we rank-ordered treatment outcomes across all individual simulated conditions, tuning all features simultaneously, within each context. Comparing the strongest treatments between

monoculture and co-culture will enable us to determine how optimal treatments vary between contexts.

2.5.1 Aggressive Feature Choices Additively Benefit Treatment in Monoculture

For monoculture, outcome is sorted by normalized live cancer cell count (**Figure 4A**). The best outcomes typically occur at the highest CAR T-cell doses, at a 25:75 $CD4^+$:$CD8^+$ ratio, at moderate to strong CAR affinity, and with high cancer cell antigen density. These trends are consistent with individual feature analyses in monoculture, and the same trends are observed in scenarios in which we considered expanded CAR T-cell doses (**Supplementary Figures S4A,B**) and $CD4^+$:$CD8^+$ ratios (**Supplementary Figure S5F**). Choosing aggressive values for all features and using large E:T ratios yield cancer cell killing rates that are comparable with those observed in most experimental studies that use E:T ratios greater than one (i.e., killing most cancer cells occurs within hours), (**Supplementary Figure S4A**). Worse outcomes, in which cancer cells grow beyond their initial plated count, occur at low CAR T-cell doses, at 100:0 and 0:100 $CD4^+$:$CD8^+$ ratios, with the weakest CAR affinity, or with lower cancer cell antigen expression. Overall, combining aggressive choices for individual features additively benefits treatment outcome in monoculture. Effective CAR T-cell designs in the absence of healthy cells combine design choices from individually optimized features.

2.5.2 Addressing Off-Target Effects Requires Tuning Multiple Parameters

To identify general conclusions across diverse co-culture conditions, we considered treatment outcomes across all individual simulated conditions, sorted by the difference in the normalized live healthy and cancer cell count at the endpoint (**Figure 4B** for ideal co-culture, **Figure 4C** for realistic co-culture). This difference is maximized when healthy cells are spared and cancer cells are killed. We expect aggressive treatments to be most effective in the ideal cases, as healthy cells that do not express antigen cannot be killed. Trends in ideal co-culture match those in monoculture, supporting the idea that "invisible" healthy cells do not change observed trends.

However, the realistic co-culture where healthy cells express antigen, and can therefore be targeted by CAR T-cells, is more clinically relevant. In this context, there is a distinct tradeoff between cancer cell killing and healthy cell sparing. Conditions with the lowest normalized live cancer cell counts also show the lowest normalized live healthy cell counts (**Figure 4C**). Treatments with a positive difference all have some amount of healthy cell killing, but this killing is minimal compared to other conditions. Effective treatments have the highest doses of CAR T-cells, weaker CARs, $CD4^+$:$CD8^+$ ratios of 25:75 or 50:50, and higher cancer cell antigen count (**Figure 4C**). These observations agree with experimental findings that optimization of CAR T-cell therapy design yields different conclusions when balancing cancer cell killing and healthy cell sparing, versus focusing on the former objective alone (Caruso et al., 2015; Johnson et al., 2015; Liu et al., 2015). Though choosing high doses of weak CAR T-cells might seem unintuitive, using weak CARs minimizes the probability of targeting healthy cells while the high dose maximizes the probability that these weaker CARs successfully interact with high antigen density cancer cells. These results suggest that delivering higher doses of weaker CAR T-cells with $CD4^+$:$CD8^+$ ratios of 25:75 or 50:50 kill more cancer cells and spare more healthy cells for tumors where on-target off-tumor killing is undesired or detrimental.

2.6 Spatial Dynamics Drive Vascularized `Tissue` Treatment Efficacy

CAR T-cell therapy has great potential for use in solid tumor contexts, which include a complex tumor microenvironment, vasculature, spatial dynamics, and potentially antigen-expressing healthy cells. Predicting how the *in vitro* behavior conferred by various CAR T-cell designs corresponds to *in vivo* performance is not straightforward. We investigate the translation and efficacy of select treatment strategies in vascularized `tissue` where a solid tumor exists in a bed of antigen-expressing healthy cells within a dynamic microenvironment. We chose a subset of simulations—the realistic co-culture conditions deemed effective after averaging across replicates (**Supplementary Table S7**)—to analyze in `tissue`. Effective treatments were those that met the following two conditions: 1) cancer cells did not grow beyond the initial number, and 2) no more than 50% of the initial healthy cells were killed off.

A `tissue` is initialized with a bed of healthy cells in vascularized tissue that was inoculated with cancer cells and grown for 30 d. At t = 21 d, treatment began by adding a specified total dose of CAR T-cells, each expressing 5×10^4 CARs with the given CAR affinity, and $CD4^+$:$CD8^+$ ratio. CAR T-cells were spawned at locations adjacent to vasculature to mimic intravenous trafficking to the tumor; they were not spawned adjacent to vessels that are too small in diameter for CAR T-cells to pass through. Files used to generate `tissue` simulations are described in **Supplementary Data S6** and **Supplementary Table S8**.

2.6.1 Tested Treatments Are Effective in `Tissue` but Differ in Healthy Cell Killing

All treated tumors resulted in far fewer cancer cells and somewhat fewer healthy cells compared to untreated conditions, indicating that all strategies identified as effective in realistic co-culture proved effective in `tissue` (**Figure 5A**). As in `dish`, healthy cell killing occurred primarily after most cancer cells were removed. This again motivates treatment strategies in which CAR T-cells include an inducible off-switch that shuts down effector function after cancer cells are killed but before lower antigen expressing healthy cells are targeted.

Comparing normalized live cancer and healthy cell counts at treatment endpoint enables direct comparison of treatment efficacy (**Figure 5B**). Notably, the primarily difference between treatment strategies is in degree of healthy cell killing. CAR affinity and cancer antigen expression, but not CAR T-cell dose or $CD4^+$:$CD8^+$ ratio, dictate this difference. In general, increasing CAR T-cell dose and using higher $CD4^+$:$CD8^+$ ratio treatments results in increased CAR T-cell counts, but it has little

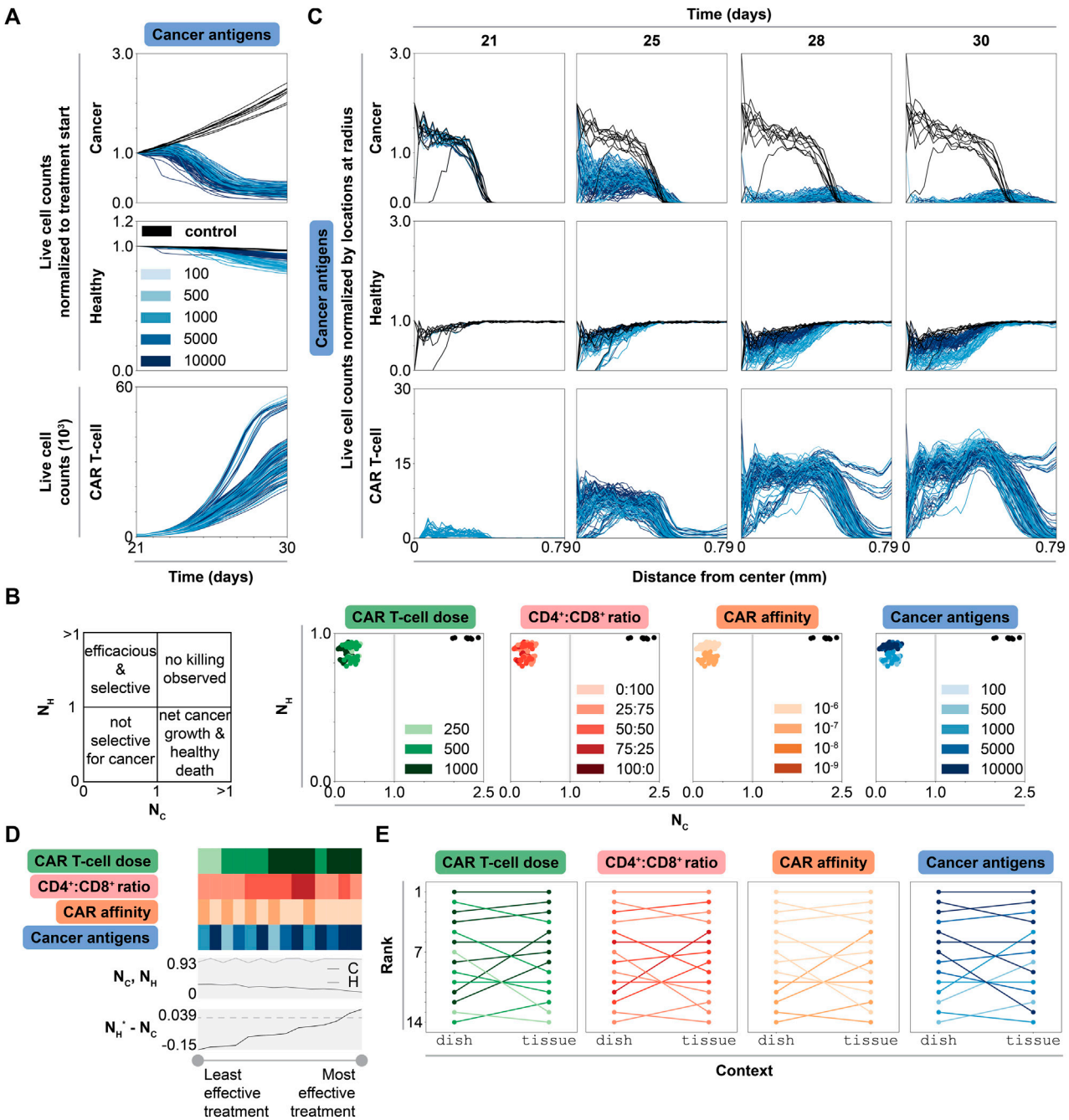


FIGURE 5 | Dynamic, spatial, and ranked outcomes for selected promising treatment combinations in `tissue`. **(A)** Live cell counts over time of untreated (black) and treated conditions (graded hues) normalized to cell count at start of treatment (t = 21 d), for all simulations, colored by cancer antigens (other features may be changing as well). Cancer antigens reported in antigens/cell. The same data colored by other features are shown in **Supplementary Figure S12 (B)** Scatter plots of normalized live cancer cell count (N_C, *x*-axis) vs normalized live healthy cell count (N_H, *y*-axis), each normalized to initial value at start of treatment (t = 21 d), for untreated (black) and treated conditions (graded hues) for all simulations, colored by one feature at a time. Each feature is reported in the following units: CAR T-cell dose = number of CAR T-cells, $CD4^+$:$CD8^+$ ratio = unitless, CAR affinity = M, cancer antigens = antigens/cell. **(C)** Normalized live cell counts over time (t = 21, 25, 28, and 30 d shown) for untreated (black) and treated conditions (graded hues), normalized to locations per radius, for all simulations, colored by cancer antigens. The columns indicate the timepoint in the simulation (day), while the rows indicate cell type plotted, and the *x*-axis for each plot shows the distance from the center. Legend is consistent with panel B. **(D)** Heatmap showing values for each feature with line plots showing normalized live cancer and healthy cell counts and difference in normalized live healthy and cancer cell counts ($N_H^* - N_C$, where healthy cell value is multiplied by the ratio of cancer to healthy cells at the start of treatment to ensure equal weighting since initial cell population sizes are not equal; dashed line indicates value of 0) at final time point averaged across replicates. The heatmap is sorted from lowest (left) to highest (right) difference. Feature legends are consistent with panels A and C. **(E)** Ladder plots of condition rankings in both `dish` and `tissue`, where condition outcome (averaged across replicates) is colored by each corresponding feature.

effect on cancer and healthy cell killing. Interestingly, most of the simulations that show the highest CAR T-cell production use the highest CAR T-cell doses and the weakest CAR affinity (**Supplementary Figure S12**, **Figure 5A**), indicating that designs with high doses of weak CAR T-cells result in the highest CAR T-cell growth rate *in vivo*. Overall, these observations reinforce the previously identified treatment strategy: use weaker CARs and select antigens with the highest differential between cancer and healthy cell expression. With this strategy, even though the CAR is weaker, the cancer antigen density is high enough to result in effective, selective treatment.

2.6.2 Cancer Cells With Higher Antigen Density Shield Healthy Cells From CAR T-Cell Killing

Though changing multiple features simultaneously complicates analysis, we noted an interesting pattern in which increasing cancer cell antigen density spares more healthy cells in `tissue`, representing a stark contrast to our `dish` findings. We thus investigated the spatial dynamics of each cell type to probe whether the mechanism by which CAR T-cells navigate within the solid tumor gives rise to this observation. At t = 21 d, cancer cells exist primarily in the center of the simulation, between the center and a radius of about 0.39 mm, while healthy cells are evenly spread across the simulation (**Figure 5C**). In untreated conditions, cancer cells grow to cover a radius of 0.58 mm by t = 30 d and healthy cell count remains unchanged over time. In treated conditions, cancer and healthy cell counts decrease over time, primarily starting from the center where most CAR T-cells are initially spawned and moving outward. Cancer cell counts decrease with increasing cancer antigen density. CAR T-cell counts increase as a function of time and cancer cell count, but not as a function of cancer antigen density. Meanwhile, higher cancer antigen levels result in decreased healthy cell killing. We hypothesize that this phenomenon occurs when high antigen density cancer cells effectively outcompete healthy cells for CAR T-cell effector function due to large differences in the probability of CAR-antigen binding between these two potential target cell types. In such scenarios, CAR T cells that successfully traffic to a tumor core are more likely to selectively target cancer cells even if healthy cells are present.

2.6.3 Spatial Differences Between `Dish` and `Tissue` Explain Treatment Performance

Comparing simulation rankings between `dish` and `tissue` reveals how context impacts treatment efficacy. To rank treatment strategies in `tissue`, we consider treatment outcomes across simulations (averaged across replicates) sorted from best to worst outcome in terms of difference in healthy and cancer cell counts normalized to start of treatment (**Figure 5D**). Nearly all highest ranked simulations use the highest CAR T-cell dose, a $CD4^+$:$CD8^+$ ratio of 25:75, the lowest CAR affinity, and the highest cancer antigen level. The four highest ranked treatment conditions in `dish` remain the four highest ranked treatment conditions in `tissue` (**Figure 5E**, **Supplementary Table S9**). The rankings for the mid and lower tier ranked simulations (5th-14th in `tissue`) are shuffled from their original rankings in `dish`. One of the worst ranked treatments in `dish` (11th) jumped to 5th in `tissue`, while a middle-ranked simulation (7th) fell to 13th in `tissue`. These data predict that the most effective treatment conditions in `dish` will perform similarly in `tissue` assuming perfect CAR T-cell trafficking. Even with perfect trafficking, performance in `dish` does not exactly correlate with performance in `tissue`. Certain conditions may outperform *in vivo* conditions compared to their performance *in vitro*.

Trends in how each feature impacts relative rank reveal which features most strongly dictate performance in `tissue` (**Figure 5E**). There are no distinct trends as a function of $CD4^+$:$CD8^+$ ratio. Most conditions that improve in rank use the relatively higher (though still objectively moderate) CAR affinity and higher CAR T-cell dose, and nearly all conditions that decrease in rank (from `dish` to `tissue`) have higher cancer antigen expression level. This finding is surprising given earlier observations that lowest CAR affinity with highest cancer antigen expression level combinations were most effective.

We hypothesize the differences in `dish` and `tissue` trends/rank result from differences in spatial dynamics. In `dish`, both healthy and cancer cells are well-mixed across the simulation, even after treatment, which results in an even spatial distribution of CAR T-cells (**Supplementary Figure S13**). In `tissue`, cancer cells sit in the center of the simulation surrounded by a large bed of healthy cells, with few healthy cells in the tumor core. When CAR T-cells are spawned with bias towards locations with more cancer cells to mimic perfect trafficking, the probability that spawn locations are adjacent to that of a healthy cell is higher in `dish` compared to `tissue`. Analyzing CAR T-cell state dynamics in both realistic co-culture `dish` and `tissue` for the selected promising treatment strategies further informs this spatial analysis. When we examine the distribution of CAR T-cell states only considering T-cells that are adjacent to a cancer cell (i.e., somewhat controlling for the local environment that a T-cell experiences), we find similar distributions of cells in effector states across `dish` and `tissue` simulations (**Supplementary Figure S14**). Exhausted and anergic states are rare in both contexts, which is unsurprising as they are expected to accumulate over longer time courses than were used in these experiments. Thus, CAR T-cells in proximity to cancer cells exhibit similar behavior independent of experimental setup, and differences in overall trends/rank between contexts result from differences in collective cancer and healthy cell spatial distributions. Overall, this spatial difference in cancer and healthy cell distribution parallels comparisons between physical *in vitro* and *in vivo* experiments, even if CAR T-cell trafficking deviates from the perfect mechanism employed in our simulations, reinforcing the key role that spatial dynamics play in treatment outcome.

3 DISCUSSION

We developed CARCADE as an open-access *in silico* testbed that enables systematic interrogation of the multidimensional design landscape of cellular engineering strategies, therapeutic optimization, and hypothesis generation. After verifying that

the developed model recapitulates known trends *in vitro*, we explored design strategies in both `dish` and `tissue` contexts to gain insight into CAR T-cell design.

Tuning individual features in `dish` revealed key insights as to how these features impact CAR T-cell design. For example, we determined that healthy cell antigen expression results in healthy cell killing but has no impact on CAR T-cell or cancer cell dynamics. Modulating individual features recapitulated known tradeoffs between cancer cell killing and healthy cell sparing in realistic co-cultures. A new observation uniquely enabled by our model's high resolution is that maximizing CAR affinity not only increases healthy cell killing but can also be counterproductive to CAR T-cell proliferation and cytokine production. In a related finding, we observed that IL-2 production is influenced more by tunable CAR T-cell design features than by healthy cell-related context.

Multidimensional analysis revealed that the relative performance of various treatment strategies is context dependent. Aggressive treatments are more effective in monoculture and ideal co-culture experiments, but effective treatment in realistic co-culture requires balancing all tuned features. We identified a particularly effective treatment strategy that balances cancer cell killing and healthy cell sparing when healthy cells express antigen. Specifically, we identified that the use of high doses of weak CAR T-cells with intermediate $CD4^+$:$CD8^+$ ratio and a maximized difference between cancer and healthy cell antigen expression produces the most effective treatments. By investigating these effective treatments in `tissue` context, we determined that differences in spatial distributions of cancer and healthy cells in `dish` and `tissue` contexts explain differences in treatment performance between contexts.

CARCADE is a first pass toward demonstrating the utility of models for generating hypotheses and informing design strategies for this class of problem, and it is important to consider that this model makes several assumptions and simplifications. First, the model is not tuned to a specific context. Results are general and might not hold in specific tumor contexts. A major strength of the model is that it can be easily tuned to a specific CAR and/or tumor type, and to interrogate specific design questions of interest. For example, CARCADE does not currently specify the CAR construct's intracellular co-stimulatory domain (ICD), which is known to be an important factor in dictating CAR T-cell efficacy, persistence, and dynamics; rather, we approximate CAR behavior independent of ICD and find that broad trends hold despite not accounting for this factor explicitly. The model could be tuned to capture the effect of different ICD choices on CAR T-cell function. Similarly, the analysis can be tuned to change the definition of effective treatment outcomes to further penalize healthy cell killing (e.g., when considering treatments in which damage to CAR target antigen-expressing healthy cells is less tolerable from a safety standpoint). When treating B-cell cancers, off-tumor effects like B-cell aplasia are manageable with treatment, and healthy cell killing is less of a concern. In glioblastomas, EGFR is expressed on cancer cells, healthy brain cells, and other tissues, making healthy cell killing a greater risk of morbidity and mortality (Caruso et al., 2015). Another assumption made in the current CARCADE model is that there is no T-cell-mediated killing of bystander cells unless those bystander cells express the target antigen, which represents an ideal case. This assumption could easily be relaxed to interrogate the consequences of various forms of non-ideal T-cell killing. Additionally, the process by which CAR T-cells traffic to the tumor has been simplified and idealized, as CAR T-cells spawn at sites closest to cancer cells. The model could be adjusted to contemplate other scenarios, such as spawning CAR T-cells at the simulation edge while including CAR T-cell and environmental features that influence CAR T-cell trafficking to the tumor. An important limitation is that CAR T-cell exhaustion and anergy are longer-term phenomena for which our understanding is continually evolving; the current formulation of CARCADE reflects an abstraction of the state of this knowledge. Future development and use of CARCADE will benefit by incorporating new insights from experiments or clinical studies, and through corresponding simulations focused on longer time scale phenomena. Such refinements will improve our ability to address important properties including CAR T-cell efficacy and persistence.

Expanding the agents, environment, or subcellular functions included in CARCADE offers opportunities for future model development and use in the field of CAR T-cell engineering. The present model comprises CAR T-cells and cancerous and healthy tissue cells; addition of macrophages, regulatory T-cells, natural killer cells, and other regulatory or supporting cell types or environmental factors could enable investigation of CAR T-cell therapy in a more complete and complex immune environment. In future studies, it may be particularly important to include the cell and environmental factors that contribute to immunosuppressive environments, as this is a common issue faced with *in vivo* CAR T-cell therapy. Additionally, while the current model was designed to facilitate analysis of treatments for solid tumors, particularly through the use of the `tissue` simulations, the `dish` simulations could be adapted to investigate liquid cancer treatment strategies. Future expansions could also incorporate trogocytosis, a processes by which CAR T-cells pick up tumor antigens from cancer cells and then experience fratricidal killing by other CAR T-cells, to investigate how this phenomena affects CAR T-cell persistence (Hamieh et al., 2019).

Overall, we believe that CARCADE will prove valuable for CAR T-cell designers and enable cross-cutting collaborations to facilitate further model development or tuning to specific contexts and questions of interest. By further refining the model using experimental data, CARCADE could help suggest potential promising strategies for experimental pursuit by testing strategies in `dish` and/or `tissue` contexts. CARCADE is designed to enable interrogation of questions and phenomena that are beyond the scope of the current study. For example, future studies using the current model could include a more granular consideration of CAR T-cell trafficking within the tumor by removing the assumption of perfect trafficking. Tumor immune escape could be investigated by creating multiple tumor subpopulations with variable antigen expression levels or

susceptibly to killing. Additionally, inter- and intra-tumor heterogeneity could be integrated by simulating tumors that comprise multiple populations with different parameters and/or differing levels of heterogeneity. Ultimately, integrating CARCADE into the CAR T-cell design process could accelerate the design-build-test cycle, saving resources and time associated with new therapeutic development.

4 MATERIALS AND METHODS

CARCADE was developed by extending ARCADE, an existing multi-scale, multi-class agent-based model that includes tissue cells and hemodynamic environments. We used CARCADE to generate *in silico* experiments where we treated monoculture `dish`, ideal and realistic co-culture `dish`, and `tissue` contexts with CAR T-cells. All model details, including adaptation of tissue cell agents, development of CAR T-cell agents, development and adaptation of subcellular modules, development of `dish` plating, and all simulation setups and analyses are described in detail in the Supplementary Methods Details section of the Supplementary Material (Kuse et al., 1985; Lauffenburger and Linderman, 1993; Robertson et al., 1996; Schwartz, 1996; Frauwirth et al., 2002; De Boer et al., 2003; Deenick et al., 2003; Iwashima, 2003; Schwartz, 2003; Chmielewski et al., 2004; Macian et al., 2004; Janas et al., 2005; Jacobs et al., 2008; Busse et al., 2010; Malek and Castro, 2010; Pearce, 2010; Yoon et al., 2010; Akbar and Henson, 2011; Wang et al., 2011; Wherry, 2011; Altman and Dang, 2012; Gerriets and Rathmell, 2012; Robertson-Tessi et al., 2012; Stone et al., 2012; Cheng et al., 2013; Crespo et al., 2013; Hegde et al., 2013; Liao et al., 2013; MacIver et al., 2013; Rosenberg, 2014; Buck et al., 2015; Heskamp et al., 2015; Kinjyo et al., 2015; Liadi et al., 2015; Liu et al., 2015; Long et al., 2015; Obst, 2015; Wherry and Kurachi, 2015; Chang and Pearce, 2016; Cherkassky et al., 2016; Golubovskaya and Wu, 2016; Harris and Kranz, 2016; Hegde et al., 2016; Liu et al., 2016; Maus and June 2016; Sommermeyer et al., 2016; Verbist et al., 2016; Arcangeli et al., 2017; Borghans and Ribeiro, 2017; Gong et al., 2017; Mehta et al., 2017; Gherbi et al., 2018; Guedan et al., 2018; Huang et al., 2018; Kasakovski et al., 2018; Rafiq et al., 2018; Ross and Cantrell, 2018; Salter et al., 2018; Watanabe et al., 2018; Yost et al., 2019; Yu and Bagheri, 2020; Hernandez-Lopez et al., 2021; Yu and Bagheri, 2021).

All source code for CARCADE is available on GitHub at https://github.com/bagherilab/CARCADE. Scripts used to process and analyze data are available on GitHub at https://github.com/bagherilab/carcade_mapping_design_space.

DATA AVAILABILITY STATEMENT

The *in silico* datasets generated and analyzed for this study can be provided upon request.

AUTHOR CONTRIBUTIONS

AP, JL, and NB conceived the project. AP developed CARCADE by extending the ARCADE model, originally created by JY. JY provided mentorship on the model development process. AP designed and conducted *in silico* experiments; processed data; and analyzed and plotted results. AP prepared the initial draft of the manuscript; AP, JY, JL, and NB wrote, refined, and edited the manuscript. JL and NB supervised the project.

FUNDING

This work was supported by the National Science Foundation Graduate Research Fellowship Program award DGE-1842165 (AP); the National Science Foundation CAREER award CBET-1653315 (NB); and by the Washington Research Foundation (NB).

ACKNOWLEDGMENTS

The authors acknowledge computational resources and staff contributions by the Quest high performance computing facility at Northwestern University, which is jointly supported by the Office of the Provost, the Office for Research, and the Northwestern University Information Technology. The authors also thank members of the Leonard and Bagheri Labs for invaluable project discussions and feedback.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.849363/full#supplementary-material

REFERENCES

Akbar, A. N., and Henson, S. M. (2011). Are Senescence and Exhaustion Intertwined or Unrelated Processes that Compromise Immunity? *Nat. Rev. Immunol.* 11 (4), 289–295. doi:10.1038/nri2959

Altman, B. J., and Dang, C. V. (2012). Normal and Cancer Cell Metabolism: Lymphocytes and Lymphoma. *FEBS J.* 279 (15), 2598–2609. doi:10.1111/j.1742-4658.2012.08651.x

Anderson, A. R. A., Weaver, A. M., Cummings, P. T., and Quaranta, V. (2006). Tumor Morphology and Phenotypic Evolution Driven by Selective Pressure from the Microenvironment. *Cell* 127 (5), 905–915. doi:10.1016/j.cell.2006.09.042

Arcangeli, S., Rotiroti, M. C., Bardelli, M., Simonelli, L., Magnani, C. F., Biondi, A., et al. (2017). Balance of Anti-cd123 Chimeric Antigen Receptor Binding Affinity and Density for the Targeting of Acute Myeloid Leukemia. *Mol. Ther.* 25 (8), 1933–1945. doi:10.1016/j.ymthe.2017.04.017

Barrett, D. M., Singh, N., Porter, D. L., Grupp, S. A., and June, C. H. (2014). Chimeric Antigen Receptor Therapy for Cancer. *Annu. Rev. Med.* 65, 333–347. doi:10.1146/annurev-med-060512-150254

Bonifant, C. L., Jackson, H. J., Brentjens, R. J., and Curran, K. J. (2016). Toxicity and Management in CAR T-Cell Therapy. *Mol. Ther. - Oncolytics* 3, 16011. doi:10.1038/mto.2016.11

Borghans, J., and Ribeiro, R. M. (2017). The Maths of Memory. *eLife* 6. doi:10.7554/eLife.26754

Buck, M. D., O'Sullivan, D., and Pearce, E. L. (2015). T Cell Metabolism Drives Immunity. *J. Exp. Med.* 212 (9), 1345–1360. doi:10.1084/jem.20151159

Busse, D., de la Rosa, M., Hobiger, K., Thurley, K., Flossdorf, M., Scheffold, A., et al. (2010). Competing Feedback Loops Shape IL-2 Signaling Between Helper and Regulatory T Lymphocytes in Cellular Microenvironments. *Proc. Natl. Acad. Sci. U.S.A.* 107 (7), 3058–3063. doi:10.1073/pnas.0812851107

Caruso, H. G., Hurton, L. V., Najjar, A., Rushworth, D., Ang, S., Olivares, S., et al. (2015). Tuning Sensitivity of CAR to EGFR Density Limits Recognition of Normal Tissue While Maintaining Potent Antitumor Activity. *Cancer Res.* 75 (17), 3505–3518. doi:10.1158/0008-5472.CAN-15-0139

Castellarin, M., Watanabe, K., June, C. H., Kloss, C. C., and Posey, A. D., Jr. (2018). Driving Cars to the Clinic for Solid Tumors. *Gene Ther.* 25 (3), 165–175. doi:10.1038/s41434-018-0007-x

Cess, C. G., and Finley, S. D. (2020a). Data-Driven Analysis of a Mechanistic Model of CAR T Cell Signaling Predicts Effects of Cell-To-Cell Heterogeneity. *J. Theor. Biol.* 489, 110125. doi:10.1016/j.jtbi.2019.110125

Cess, C. G., and Finley, S. D. (2020b). Multi-Scale Modeling of Macrophage-T Cell Interactions within the Tumor Microenvironment. *PLoS Comput. Biol.* 16 (12), e1008519. doi:10.1371/journal.pcbi.1008519

Chang, C.-H., and Pearce, E. L. (2016). Emerging Concepts of T Cell Metabolism as a Target of Immunotherapy. *Nat. Immunol.* 17 (4), 364–368. doi:10.1038/ni.3415

Chavali, A. K., Gianchandani, E. P., Tung, K. S., Lawrence, M. B., Peirce, S. M., and Papin, J. A. (2008). Characterizing Emergent Properties of Immunological Systems with Multi-Cellular Rule-Based Computational Modeling. *Trends Immunol.* 29 (12), 589–599. doi:10.1016/j.it.2008.08.006

Cheng, X., Veverka, V., Radhakrishnan, A., Waters, L. C., Muskett, F. W., Morgan, S. H., et al. (2013). Structure and Interactions of the Human Programmed Cell Death 1 Receptor. *J. Biol. Chem.* 288 (17), 11771–11785. doi:10.1074/jbc.M112.448126

Cherkassky, L., Morello, A., Villena-Vargas, J., Feng, Y., Dimitrov, D. S., Jones, D. R., et al. (2016). Human CAR T Cells with Cell-Intrinsic PD-1 Checkpoint Blockade Resist Tumor-Mediated Inhibition. *J. Clin. Invest.* 126 (8), 3130–3144. doi:10.1172/JCI83092

Chmielewski, M., Hombach, A., Heuser, C., Adams, G. P., and Abken, H. (2004). T Cell Activation by Antibody-Like Immunoreceptors: Increase in Affinity of the Single-Chain Fragment Domain Above Threshold Does Not Increase T Cell Activation Against Antigen-Positive Target Cells But Decreases Selectivity. *J. Immunol.* 173 (12), 7647–7653. doi:10.4049/jimmunol.173.12.7647

Cho, J. H., Collins, J. J., and Wong, W. W. (2018). Universal Chimeric Antigen Receptors for Multiplexed and Logical Control of T Cell Responses. *Cell* 173 (6), 1426–1438. e1411. doi:10.1016/j.cell.2018.03.038

Crespo, J., Sun, H., Welling, T. H., Tian, Z., and Zou, W. (2013). T Cell Anergy, Exhaustion, Senescence, and Stemness in the Tumor Microenvironment. *Curr. Opin. Immunol.* 25 (2), 214–221. doi:10.1016/j.coi.2012.12.003

De Boer, R. J., Homann, D., and Perelson, A. S. (2003). Different Dynamics of CD4+ and CD8+ T Cell Responses During and After Acute Lymphocytic Choriomeningitis Virus Infection. *J. Immunol.* 171 (8), 3928–3935. doi:10.4049/jimmunol.171.8.3928

Deenick, E. K., Gett, A. V., and Hodgkin, P. D. (2003). Stochastic Model of T Cell Proliferation: A Calculus Revealing IL-2 Regulation of Precursor Frequencies, Cell Cycle Time, and Survival. *J. Immunol.* 170 (10), 4963–4972. doi:10.4049/jimmunol.170.10.4963

Folcik, V. A., An, G. C., and Orosz, C. G. (2007). The Basic Immune Simulator: An Agent-Based Model to Study the Interactions Between Innate and Adaptive Immunity. *Theor. Biol. Med. Model* 4, 39. doi:10.1186/1742-4682-4-39

Frauwirth, K. A., Riley, J. L., Harris, M. H., Parry, R. V., Rathmell, J. C., Plas, D. R., et al. (2002). The CD28 Signaling Pathway Regulates Glucose Metabolism. *Immunity* 16 (6), 769–777. doi:10.1016/s1074-7613(02)00323-0

Fu, W., Lei, C., Liu, S., Cui, Y., Wang, C., Qian, K., et al. (2019). CAR Exosomes Derived from Effector CAR-T Cells Have Potent Antitumour Effects and Low Toxicity. *Nat. Commun.* 10 (1), 1–12. doi:10.1038/s41467-019-12321-3

Galluzzi, L., Chan, T. A., Kroemer, G., Wolchok, J. D., and López-Soto, A. (2018). The Hallmarks of Successful Anticancer Immunotherapy. *Sci. Transl. Med.* 10 (459), 1–14. doi:10.1126/scitranslmed.aat7807

Gerriets, V. A., and Rathmell, J. C. (2012). Metabolic Pathways in T Cell Fate and Function. *Trends Immunol.* 33 (4), 168–173. doi:10.1016/j.it.2012.01.010

Gherbi, K., Briddon, S. J., and Charlton, S. J. (2018). Micro-Pharmacokinetics: Quantifying Local Drug Concentration at Live Cell Membranes. *Sci. Rep.* 8 (1), 3479. doi:10.1038/s41598-018-21100-x

Ghorashian, S., Kramer, A. M., Onuoha, S., Wright, G., Bartram, J., Richardson, R., et al. (2019). Enhanced CAR T Cell Expansion and Prolonged Persistence in Pediatric Patients with ALL Treated with a Low-Affinity CD19 CAR. *Nat. Med.* 25 (9), 1408–1414. doi:10.1038/s41591-019-0549-5

Golubovskaya, V., and Wu, L. (2016). Different Subsets of T Cells, Memory, Effector Functions, and CAR-T Immunotherapy. *Cancers* 8 (3), 36. doi:10.3390/cancers8030036

Gong, C., Milberg, O., Wang, B., Vicini, P., Narwal, R., Roskos, L., et al. (2017). A Computational Multiscale Agent-Based Model for Simulating Spatio-Temporal Tumour Immune Response to PD1 and PDL1 Inhibition. *J. R. Soc. Interface.* 14 (134), 20170320. doi:10.1098/rsif.2017.0320

Guedan, S., Posey, A. D., Jr., Shaw, C., Wing, A., Da, T., Patel, P. R., et al. (2018). Enhancing CAR T Cell Persistence Through ICOS and 4-1BB Costimulation. *JCI Insight* 3 (1). doi:10.1172/jci.insight.96976

Hamieh, M., Dobrin, A., Cabriolu, A., van der Stegen, S. J. C., Giavridis, T., Mansilla-Soto, J., et al. (2019). CAR T Cell Trogocytosis and Cooperative Killing Regulate Tumour Antigen Escape. *Nature* 568 (7750), 112–116. doi:10.1038/s41586-019-1054-1

Harris, D. T., and Kranz, D. M. (2016). Adoptive T Cell Therapies: A Comparison of T Cell Receptors and Chimeric Antigen Receptors. *Trends Pharmacol. Sci.* 37 (3), 220–230. doi:10.1016/j.tips.2015.11.004

Hegde, M., Corder, A., Chow, K. K., Mukherjee, M., Ashoori, A., Kew, Y., et al. (2013). Combinational Targeting Offsets Antigen Escape and Enhances Effector Functions of Adoptively Transferred T Cells in Glioblastoma. *Mol. Ther.* 21 (11), 2087–2101. doi:10.1038/mt.2013.185

Hegde, M., Mukherjee, M., Grada, Z., Pignata, A., Landi, D., Navai, S. A., et al. (2016). Tandem CAR T Cells Targeting HER2 and IL13Rα2 Mitigate Tumor Antigen Escape. *J. Clin. Invest.* 126 (8), 3036–3052. doi:10.1172/JCI83416

Hernandez-Lopez, R. A., Yu, W., Cabral, K. A., Creasey, O. A., del Pilar Lopez Pazmino, M., Tonai, Y., et al. (2021). T Cell Circuits That Sense Antigen Density with an Ultrasensitive Threshold. *Science* 1855. doi:10.1126/science.abc1855

Heskamp, S., Hobo, W., Molkenboer-Kuenen, J. D. M., Olive, D., Oyen, W. J. G., Dolstra, H., et al. (2015). Noninvasive Imaging of Tumor PD-L1 Expression Using Radiolabeled Anti-PD-L1 Antibodies. *Cancer Res.* 75 (14), 2928–2936. doi:10.1158/0008-5472.CAN-14-3477

Hou, B., Tang, Y., Li, W., Zeng, Q., and Chang, D. (2019). Efficiency of CAR-T Therapy for Treatment of Solid Tumor in Clinical Trials: A Meta-Analysis. *Dis. Markers* 2019, 1–11. doi:10.1155/2019/3425291

Huang, Y., Li, D., Qin, D.-Y., Gou, H.-F., Wei, W., Wang, Y.-S., et al. (2018). Interleukin-Armed Chimeric Antigen Receptor-Modified T Cells for Cancer Immunotherapy. *Gene Ther.* 25 (3), 192–197. doi:10.1038/gt.2017.81

Hudecek, M., Lupo-Stanghellini, M.-T., Kosasih, P. L., Sommermeyer, D., Jensen, M. C., Rader, C., et al. (2013). Receptor Affinity and Extracellular Domain Modifications Affect Tumor Recognition by ROR1-Specific Chimeric Antigen Receptor T Cells. *Clin. Cancer Res.* 19 (12), 3153–3164. doi:10.1158/1078-0432.CCR-13-0330

Iwashima, M. (2003). Kinetic Perspectives of T Cell Antigen Receptor Signaling. *Immunol. Rev.* 191 (1), 196–210. doi:10.1034/j.1600-065x.2003.00024.x

Jackson, H. J., Rafiq, S., and Brentjens, R. J. (2016). Driving CAR T-Cells Forward. *Nat. Rev. Clin. Oncol.* 13 (6), 370–383. doi:10.1038/nrclinonc.2016.36

Jacobs, S. R., Herman, C. E., Maciver, N. J., Wofford, J. A., Wieman, H. L., Hammen, J. J., et al. (2008). Glucose Uptake Is Limiting in T Cell Activation and Requires CD28-Mediated Akt-Dependent and Independent Pathways. *J. Immunol.* 180 (7), 4476–4486. doi:10.4049/jimmunol.180.7.4476

Janas, M. L., Groves, P., Kienzle, N., and Kelso, A. (2005). IL-2 Regulates Perforin and Granzyme Gene Expression in CD8+T Cells Independently of its Effects on

Survival and Proliferation. *J. Immunol.* 175 (12), 8003–8010. doi:10.4049/jimmunol.175.12.8003

Johnson, L. A., Scholler, J., Ohkuri, T., Kosaka, A., Patel, P. R., McGettigan, S. E., et al. (2015). Rational Development and Characterization of Humanized Anti-EGFR Variant III Chimeric Antigen Receptor T Cells for Glioblastoma. *Sci. Transl. Med.* 7 (275), 275ra222. doi:10.1126/scitranslmed.aaa4963

Jones, R. G., and Thompson, C. B. (2007). Revving the Engine: Signal Transduction Fuels T Cell Activation. *Immunity* 27 (2), 173–178. doi:10.1016/j.immuni.2007.07.008

Kakarla, S., and Gottschalk, S. (2014). CAR T Cells for Solid Tumors. *Cancer J.* 20 (2), 151–155. doi:10.1097/PPO.0000000000000032

Kasakovski, D., Xu, L., and Li, Y. (2018). T Cell Senescence and CAR-T Cell Exhaustion in Hematological Malignancies. *J. Hematol. Oncol.* 11 (1), 91. doi:10.1186/s13045-018-0629-x

Kawalekar, O. U., O'Connor, R. S., Fraietta, J. A., Guo, L., McGettigan, S. E., Posey, A. D., Jr., et al. (2016). Distinct Signaling of Coreceptors Regulates Specific Metabolism Pathways and Impacts Memory Development in CAR T Cells. *Immunity* 44 (2), 380–390. doi:10.1016/j.immuni.2016.01.021

Kinjyo, I., Qin, J., Tan, S.-Y., Wellard, C. J., Mrass, P., Ritchie, W., et al. (2015). Real-Time Tracking of Cell Cycle Progression During CD8+ Effector and Memory T-Cell Differentiation. *Nat. Commun.* 6, 6301. doi:10.1038/ncomms7301

Kuse, R., Schuster, S., Schubbe, H., Dix, S., and Hausmann, K. (1985). Blood Lymphocyte Volumes and Diameters in Patients with Chronic Lymphocytic Leukemia and Normal Controls. *Blut* 50 (4), 243–248. doi:10.1007/BF00320301

Lanitis, E., Poussin, M., Klattenhoff, A. W., Song, D., Sandaltzopoulos, R., June, C. H., et al. (2013). Chimeric Antigen Receptor T Cells with Dissociated Signaling Domains Exhibit Focused Antitumor Activity with Reduced Potential for Toxicity *In Vivo*. *Cancer Immunol. Res.* 1 (1), 43–53. doi:10.1158/2326-6066.CIR-13-0008

Lauffenburger, D. A., and Linderman, J. J. (1993). *Receptors: Models for Binding, Trafficking, and Signaling*. Oxford University Press.

Liadi, I., Singh, H., Romain, G., Rey-Villamizar, N., Merouane, A., Adolacion, J. R. T., et al. (2015). Individual Motile CD4+ T Cells Can Participate in Efficient Multikilling Through Conjugation to Multiple Tumor Cells. *Cancer Immunol. Res.* 3 (5), 473–482. doi:10.1158/2326-6066.CIR-14-0195

Liao, W., Lin, J.-X., and Leonard, W. J. (2013). Interleukin-2 at the Crossroads of Effector Responses, Tolerance, and Immunotherapy. *Immunity* 38 (1), 13–25. doi:10.1016/j.immuni.2013.01.004

Lim, W. A., and June, C. H. (2017). The Principles of Engineering Immune Cells to Treat Cancer. *Cell* 168 (4), 724–740. doi:10.1016/j.cell.2017.01.016

Liu, X., Jiang, S., Fang, C., Yang, S., Olalere, D., Pequignot, E. C., et al. (2015). Affinity-Tuned ErbB2 or EGFR Chimeric Antigen Receptor T Cells Exhibit an Increased Therapeutic Index Against Tumors in Mice. *Cancer Res.* 75 (17), 3596–3607. doi:10.1158/0008-5472.CAN-15-0159

Liu, X., Ranganathan, R., Jiang, S., Fang, C., Sun, J., Kim, S., et al. (2016). A Chimeric Switch-Receptor Targeting PD1 Augments the Efficacy of Second-Generation CAR T Cells in Advanced Solid Tumors. *Cancer Res.* 76 (6), 1578–1590. doi:10.1158/0008-5472.CAN-15-2524

Long, A. H., Haso, W. M., Shern, J. F., Wanhainen, K. M., Murgai, M., Ingaramo, M., et al. (2015). 4-1BB Costimulation Ameliorates T Cell Exhaustion Induced by Tonic Signaling of Chimeric Antigen Receptors. *Nat. Med.* 21 (6), 581–590. doi:10.1038/nm.3838

Macián, F., Im, S.-H., Garcı´a-Cozar, F. J., and Rao, A. (2004). T-cell Anergy. *Curr. Opin. Immunol.* 16 (2), 209–216. doi:10.1016/j.coi.2004.01.013

MacIver, N. J., Michalek, R. D., and Rathmell, J. C. (2013). Metabolic Regulation of T Lymphocytes. *Annu. Rev. Immunol.* 31, 259–283. doi:10.1146/annurev-immunol-032712-095956

Majzner, R. G., Rietberg, S. P., Sotillo, E., Dong, R., Vachharajani, V. T., Labanieh, L., et al. (2020). Tuning the Antigen Density Requirement for CAR T-Cell Activity. *Cancer Discov.* 10 (5), 702–723. doi:10.1158/2159-8290.CD-19-0945

Malek, T. R., and Castro, I. (2010). Interleukin-2 Receptor Signaling: at the Interface between Tolerance and Immunity. *Immunity* 33 (2), 153–165. doi:10.1016/j.immuni.2010.08.004

Maus, M. V., and June, C. H. (2016). Making Better Chimeric Antigen Receptors for Adoptive T-Cell Therapy. *Clin. Cancer Res.* 22 (8), 1875–1884. doi:10.1158/1078-0432.CCR-15-1433

Mehta, M. M., Weinberg, S. E., and Chandel, N. S. (2017). Mitochondrial Control of Immunity: Beyond ATP. *Nat. Rev. Immunol.* 17 (10), 608–620. doi:10.1038/nri.2017.66

Narang, V., Decraene, J., Wong, S.-Y., Aiswarya, B. S., Wasem, A. R., Leong, S. R., et al. (2012). Systems Immunology: A Survey of Modeling Formalisms, Applications and Simulation Tools. *Immunol. Res.* 53 (1-3), 251–265. doi:10.1007/s12026-012-8305-7

NCI (2022). *CAR T Cells: Engineering Patients' Immune Cells To Treat Their Cancers* [Online]. Available: https://www.cancer.gov/about-cancer/treatment/research/car-t-cells.

Norton, K.-A., Wallace, T., Pandey, N. B., and Popel, A. S. (2017). An Agent-Based Model of Triple-Negative Breast Cancer: the Interplay Between Chemokine Receptor CCR5 Expression, Cancer Stem Cells, and Hypoxia. *BMC Syst. Biol.* 11 (1), 68. doi:10.1186/s12918-017-0445-x

Obst, R. (2015). The Timing of T Cell Priming and Cycling. *Front. Immunol.* 6, 563. doi:10.3389/fimmu.2015.00563

Owens, K., and Bozic, I. (2021). Modeling CAR T-Cell Therapy with Patient Preconditioning. *Bull. Math. Biol.* 83 (5), 42. doi:10.1007/s11538-021-00869-5

Pearce, E. L. (2010). Metabolism in T Cell Activation and Differentiation. *Curr. Opin. Immunol.* 22 (3), 314–320. doi:10.1016/j.coi.2010.01.018

Rafiq, S., Yeku, O. O., Jackson, H. J., Purdon, T. J., van Leeuwen, D. G., Drakes, D. J., et al. (2018). Targeted Delivery of a PD-1-Blocking scFv by CAR-T Cells Enhances Anti-Tumor Efficacy *In Vivo*. *Nat. Biotechnol.* 36 (9), 847–856. doi:10.1038/nbt.4195

Robertson, K. W., Reeves, J. R., Smith, G., Keith, W. N., Ozanne, B. W., Cooke, T. G., et al. (1996). Quantitative Estimation of Epidermal Growth Factor Receptor and C-erbB-2 in Human Breast Cancer. *Cancer Res.* 56 (16), 3823–3830.

Robertson-Tessi, M., El-Kareh, A., and Goriely, A. (2012). A Mathematical Model of Tumor-Immune Interactions. *J. Theor. Biol.* 294, 56–73. doi:10.1016/j.jtbi.2011.10.027

Rohrs, J. A., Siegler, E. L., Wang, P., and Finley, S. D. (2020). ERK Activation in CAR T Cells Is Amplified by CD28-Mediated Increase in CD3ζ Phosphorylation. *iScience* 23 (4), 101023. doi:10.1016/j.isci.2020.101023

Rohrs, J. A., Zheng, D., Graham, N. A., Wang, P., and Finley, S. D. (2018). Computational Model of Chimeric Antigen Receptors Explains Site-Specific Phosphorylation Kinetics. *Biophysical J.* 115 (6), 1116–1129. doi:10.1016/j.bpj.2018.08.018

Rosenberg, S. A. (2014). IL-2: The First Effective Immunotherapy for Human Cancer. *J. I.* 192 (12), 5451–5458. doi:10.4049/jimmunol.1490019

Ross, S. H., and Cantrell, D. A. (2018). Signaling and Function of Interleukin-2 in T Lymphocytes. *Annu. Rev. Immunol.* 36, 411–433. doi:10.1146/annurev-immunol-042617-053352

Salter, A. I., Ivey, R. G., Kennedy, J. J., Voillet, V., Rajan, A., Alderman, E. J., et al. (2018). Phosphoproteomic Analysis of Chimeric Antigen Receptor Signaling Reveals Kinetic and Quantitative Differences that Affect Cell Function. *Sci. Signal* 11. doi:10.1126/scisignal.aat6753

Sampson, J. H., Choi, B. D., Sanchez-Perez, L., Suryadevara, C. M., Snyder, D. J., Flores, C. T., et al. (2014). EGFRvIII mCAR-Modified T-Cell Therapy Cures Mice with Established Intracerebral Glioma and Generates Host Immunity Against Tumor-Antigen Loss. *Clin. Cancer Res.* 20 (4), 972–984. doi:10.1158/1078-0432.CCR-13-0709

Schwartz, R. H. (1996). Models of T Cell Anergy: Is There a Common Molecular Mechanism? *J. Exp. Med.* 184, 1–8. doi:10.1084/jem.184.1.1

Schwartz, R. H. (2003). T Cell Anergy. *Annu. Rev. Immunol.* 21, 305–334. doi:10.1146/annurev.immunol.21.120601.141110

Sommermeyer, D., Hudecek, M., Kosasih, P. L., Gogishvili, T., Maloney, D. G., Turtle, C. J., et al. (2016). Chimeric Antigen Receptor-Modified T Cells Derived from Defined CD8+ and CD4+ Subsets Confer Superior Antitumor Reactivity *In Vivo*. *Leukemia* 30 (2), 492–500. doi:10.1038/leu.2015.247

Stone, J. D., Aggen, D. H., Schietinger, A., Schreiber, H., and Kranz, D. M. (2012). A Sensitivity Scale for Targeting T Cells with Chimeric Antigen Receptors (CARs) and Bispecific T-Cell Engagers (BiTEs). *Oncoimmunology* 1 (6), 863–873. doi:10.4161/onci.20592

Turtle, C. J., Hanafi, L.-A., Berger, C., Gooley, T. A., Cherian, S., Hudecek, M., et al. (2016). CD19 CAR-T Cells of Defined CD4+:CD8+ Composition in Adult B Cell ALL Patients. *J. Clin. Invest.* 126 (6), 2123–2138. doi:10.1172/JCI85309

van Stipdonk, M. J. B., Hardenberg, G., Bijker, M. S., Lemmens, E. E., Droin, N. M., Green, D. R., et al. (2003). Dynamic Programming of CD8+ T Lymphocyte Responses. *Nat. Immunol.* 4 (4), 361–365. doi:10.1038/ni912

Verbist, K. C., Guy, C. S., Milasta, S., Liedmann, S., Kamiński, M. M., Wang, R., et al. (2016). Metabolic Maintenance of Cell Asymmetry Following Division in Activated T Lymphocytes. *Nature* 532 (7599), 389–393. doi:10.1038/nature17442

Vodovotz, Y., Xia, A., Read, E. L., Bassaganya-Riera, J., Hafler, D. A., Sontag, E., et al. (2017). Solving Immunology? *Trends Immunol.* 38 (2), 116–127. doi:10.1016/j.it.2016.11.006

Waclaw, B., Bozic, I., Pittman, M. E., Hruban, R. H., Vogelstein, B., and Nowak, M. A. (2015). A Spatial Model Predicts That Dispersal and Cell Turnover Limit Intratumour Heterogeneity. *Nature* 525 (7568), 261–264. doi:10.1038/nature14971

Wang, R., Dillon, C. P., Shi, L. Z., Milasta, S., Carter, R., Finkelstein, D., et al. (2011). The Transcription Factor Myc Controls Metabolic Reprogramming Upon T Lymphocyte Activation. *Immunity* 35 (6), 871–882. doi:10.1016/j.immuni.2011.09.021

Watanabe, K., Kuramitsu, S., Posey, A. D., Jr., and June, C. H. (2018). Expanding the Therapeutic Window for CAR T Cell Therapy in Solid Tumors: The Knowns and Unknowns of CAR T Cell Biology. *Front. Immunol.* 9, 2486. doi:10.3389/fimmu.2018.02486

Watanabe, K., Terakura, S., Martens, A. C., van Meerten, T., Uchiyama, S., Imai, M., et al. (2015). Target Antigen Density Governs the Efficacy of Anti-CD20-CD28-CD3 ζ Chimeric Antigen Receptor-Modified Effector CD8+ T Cells. *J. I.* 194 (3), 911–920. doi:10.4049/jimmunol.1402346

Wherry, E. J., and Kurachi, M. (2015). Molecular and Cellular Insights into T Cell Exhaustion. *Nat. Rev. Immunol.* 15 (8), 486–499. doi:10.1038/nri3862

Wherry, E. J. (2011). T Cell Exhaustion. *Nat. Immunol.* 12 (6), 492–499. doi:10.1038/ni.2035

Whiteside, T. L. (2008). The Tumor Microenvironment and its Role in Promoting Tumor Growth. *Oncogene* 27 (45), 5904–5912. doi:10.1038/onc.2008.271

Wilkie, S., van Schalkwyk, M. C. I., Hobbs, S., Davies, D. M., van der Stegen, S. J. C., Pereira, A. C. P., et al. (2012). Dual Targeting of ErbB2 and MUC1 in Breast Cancer Using Chimeric Antigen Receptors Engineered to Provide Complementary Signaling. *J. Clin. Immunol.* 32 (5), 1059–1070. doi:10.1007/s10875-012-9689-9

Wu, C.-Y., Roybal, K. T., Puchner, E. M., Onuffer, J., and Lim, W. A. (2015). Remote Control of Therapeutic T Cells Through a Small Molecule-Gated Chimeric Receptor. *Science* 350 (6258), aab4077. doi:10.1126/science.aab4077

Yáñez-Muñoz, R. J., and Grupp, S. A. (2018). CAR-T in the Clinic: Drive with Care. *Gene Ther.* 25 (3), 157–161. doi:10.1038/s41434-018-0023-x

Yoon, H., Kim, T. S., and Braciale, T. J. (2010). The Cell Cycle Time of CD8+ T Cells Responding *In Vivo* Is Controlled by the Type of Antigenic Stimulus. *PLoS One* 5 (11), e15423. doi:10.1371/journal.pone.0015423

Yost, K. E., Satpathy, A. T., Wells, D. K., Qi, Y., Wang, C., Kageyama, R., et al. (2019). Clonal Replacement of Tumor-Specific T Cells Following PD-1 Blockade. *Nat. Med.* 25 (8), 1251–1259. doi:10.1038/s41591-019-0522-3

Yu, J. S., and Bagheri, N. (2020). Agent-Based Models Predict Emergent Behavior of Heterogeneous Cell Populations in Dynamic Microenvironments. *Front. Bioeng. Biotechnol.* 8, 249. doi:10.3389/fbioe.2020.00249

Yu, J. S., and Bagheri, N. (2021). Modular Microenvironment Components Reproduce Vascular Dynamics *De Novo* in a Multi-Scale Agent-Based Model. *Cell Syst.* 12 (8), 795–809. doi:10.1016/j.cels.2021.05.007

Yu, J. S., and Bagheri, N. (2016). Multi-Class and Multi-Scale Models of Complex Biological Phenomena. *Curr. Opin. Biotechnol.* 39, 167–173. doi:10.1016/j.copbio.2016.04.002

Zhang, L., Wang, Z., Sagotsky, J. A., and Deisboeck, T. S. (2009). Multiscale Agent-Based Cancer Modeling. *J. Math. Biol.* 58 (4-5), 545–559. doi:10.1007/s00285-008-0211-1

Zhao, Z., Condomines, M., van der Stegen, S. J. C., Perna, F., Kloss, C. C., Gunset, G., et al. (2015). Structural Design of Engineered Costimulation Determines Tumor Rejection Kinetics and Persistence of CAR T Cells. *Cancer Cell* 28 (4), 415–428. doi:10.1016/j.ccell.2015.09.004

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Prybutok, Yu, Leonard and Bagheri. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Molecular Biosciences

BRIEF RESEARCH REPORT
published: 14 July 2022
doi: 10.3389/fmolb.2022.898627

Generalized Property-Based Encoders and Digital Signal Processing Facilitate Predictive Tasks in Protein Engineering

David Medina-Ortiz[1,2], *Sebastian Contreras*[3]*, *Juan Amado-Hinojosa*[1,4], *Jorge Torres-Almonacid*[2], *Juan A. Asenjo*[1,4], *Marcelo Navarrete*[5] and *Álvaro Olivera-Nappa*[1,4]*

[1]Centre for Biotechnology and Bioengineering, Universidad de Chile, Santiago, Chile, [2]Departamento de Ingeniería en Computación, Universidad de Magallanes, Punta Arenas, Chile, [3]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany, [4]Departamento de Ingeniería Química, Biotecnología y Materiales, Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile, Santiago, Chile, [5]Escuela de Medicina, Universidad de Magallanes, Punta Arenas, Chile

OPEN ACCESS

Edited by:
Daniel Garrido,
Pontificia Universidad Católica de Chile, Chile

Reviewed by:
Chris Kieslich,
Auburn University, United States
Md. Mehedi Hasan,
Kyushu Institute of Technology, Japan

***Correspondence:**
Sebastian Contreras
sebastian.contreras@ds.mpg.de
Álvaro Olivera-Nappa
aolivera@ing.uchile.cl

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: 17 March 2022
Accepted: 23 June 2022
Published: 14 July 2022

Citation:
Medina-Ortiz D, Contreras S, Amado-Hinojosa J, Torres-Almonacid J, Asenjo JA, Navarrete M and Olivera-Nappa Á (2022) Generalized Property-Based Encoders and Digital Signal Processing Facilitate Predictive Tasks in Protein Engineering. Front. Mol. Biosci. 9:898627. doi: 10.3389/fmolb.2022.898627

Computational methods in protein engineering often require encoding amino acid sequences, i.e., converting them into numeric arrays. Physicochemical properties are a typical choice to define encoders, where we replace each amino acid by its value for a given property. However, what property (or group thereof) is best for a given predictive task remains an open problem. In this work, we generalize property-based encoding strategies to maximize the performance of predictive models in protein engineering. First, combining text mining and unsupervised learning, we partitioned the AAIndex database into eight semantically-consistent groups of properties. We then applied a non-linear PCA within each group to define a single encoder to represent it. Then, in several case studies, we assess the performance of predictive models for protein and peptide function, folding, and biological activity, trained using the proposed encoders and classical methods (One Hot Encoder and TAPE embeddings). Models trained on datasets encoded with our encoders and converted to signals through the Fast Fourier Transform (FFT) increased their precision and reduced their overfitting substantially, outperforming classical approaches in most cases. Finally, we propose a preliminary methodology to create *de novo* sequences with desired properties. All these results offer simple ways to increase the performance of general and complex predictive tasks in protein engineering without increasing their complexity.

Keywords: protein engineering, predictive models, machine learning, digital signal processing, fourier transform, numerical representation strategies

1 INTRODUCTION

Protein Engineering is one of the main research areas of biotechnology. It focuses on designing and implementing strategies that allow or optimize the production of proteins with desired properties. The main strategies used to achieve this objective are directed evolution and rational design. The first focuses on emulating and accelerating the evolution process, evaluating mutations and selecting those that show the desired trait, iterating the process until reaching an economically feasible optimum. The second consists of applying existing knowledge about a protein system—both

empirical and theoretical—to propose mutations or variants that are likelier to exhibit the desired property.

However, predicting the outcome of replacing one (or more) amino acids in a protein sequence is a central task in Protein Engineering because it is unclear how the individual sequences relate to higher-order properties, e.g., folding Khoury et al. (2014). As protein function and properties are closely related to their constitutive amino acids (Sadowski and Jones, 2009), it is possible to design variants with enhanced functions by changing the constitutive amino acids of a sequence. However, amino acid sequences need to be *encoded* into numeric arrays in order to facilitate their computational processing. In other words, every amino acid has to be turned into a number.

Unsurprisingly, encoders play a fundamental role in the quality of the outcome of predictive models (Yang et al., 2019; Wittmann et al., 2021). However, while there is a wide variety of encoding techniques, there is no general agreement on which one to select for a specific task (Yang et al., 2018; Siedhoff et al., 2021). The first encoding approaches represented amino acid sequences in discrete manner (numeric-wise), using techniques such as One Hot or Ordinal Encoder (Winter 1998; Pavelka et al., 2009; Brownlee, 2020). However, these techniques struggle to handle high-dimensional datasets and often lack biological interpretation (Yang et al., 2018). Therefore, efficient encoding strategies allowing handling high dimensional datasets while capturing biological and physicochemical properties of the sequences are required.

Researchers have intensively used physicochemical properties of the constitutive amino acids to encode sequences (Potapov et al., 2009; Broom et al., 2017; Ancien et al., 2018). One of the open datasets summarizing these properties is the AAIndex database (Kawashima and Kanehisa, 2000), with (to date) 566 different entries for the 20 canonical amino acids. Property selection based on unsupervised machine learning (ML) algorithms (Saha et al., 2012; Forghani and Khani, 2017) often generates groups with mixed properties (Georgiev, 2009), in the sense that they are not semantically or physically coherent. Various studies have combined physicochemical properties and digital signal processing in protein engineering (Cosic and Nesic, 1987; Hejase de Trad et al., 2000). An example of such digital signal processing is the use of (Fast) Fourier Transforms (FFT) to analyze encoded sequences' spectra. Integral transforms (as the FFT) have some interesting properties, as facilitating the convolution of signals and, eventually, capturing the interaction between amino acids. Consistently, in the context of protein engineering, transforming encoded sequences and training models in the frequency space instead allows capturing the interactions between amino acids in the whole range of the sequence (Siedhoff et al., 2020). Veljkovic et al. (1985) were pioneers in the application of discrete Fourier Transforms to analyze DNA and protein sequences. Other remarkable examples find applications in cancer studies (Cosic et al., 2016), analysis of conserved motif regions (Hejase de Trad et al., 2000), evaluation of bioactivity (Cosic, 1994), and the prediction of secondary structure and protein-protein interactions (Cosic et al., 2016). Recently, researchers have combined with great success digital signal processing with machine learning to develop predictive models to evaluate—among other variables—enantioselectivity and protein thermostability (Cadet et al., 2018a,b; Siedhoff et al., 2021).

However, the use of integral transforms (as Fourier transforms) dates further back on time (see, e.g., Eisenberg et al. (1984); Rackovsky (1998); Cosic and Nesic, 1987). Recently, Kieslich et al. aimed to generalize the property-based encoding of sequences by applying a PCA to the AAIndex. They select the 18 most explanatory principal components to define encoders and use them for training Support Vector Machine (SVM) models to predict antiviral activity on peptides, reaching outstanding performance metrics (Kieslich et al., 2021). Thereby, the authors showcase the benefits of extracting the full potential of the AAIndex dataset by proper data preprocessing. Could it then be possible to extract even more information from the AAIndex database so that we could reach higher performance metrics by employing even fewer independent encoders?

In this work, we aim to go one step beyond generalizing property-based encoding strategies and improve the numerical representation of amino acid sequences for predictive tasks in protein engineering in a way that is both explainable and consistent with previous findings. First, we applied text mining techniques to the AAIndex database to define eight semantically consistent groups of properties (i.e., groups of properties with compatible physical meaning, which naturally arise). Then, using the first component of a Kernel PCA (which is less restrictive than classic PCA), we define eight encoders that we use to represent the same protein sequence. After applying FFT, we train predictive models using the complex modulus of the Fourier spectra as input. Thereby, we facilitate the development of predictive models using ML algorithms, outperforming classical encoding strategies in the studied cases. Finally, we demonstrate the usability of the proposed approach to enhance performance in predictive tasks and to design proteins with desirable properties.

2 METHODS

2.1 Semantic Clustering of Properties in the AAIndex Database

We sought to identify groups of physicochemical properties in the AAIndex database (Kawashima and Kanehisa, 2000), maximizing the separation between groups while conserving semantic consistency within them (in the sense of all properties of the same group having compatible descriptions). Our methodology combined doc2vec strategies as document representation (Kim et al., 2019) and several unsupervised learning algorithms. Below, we describe each of the four stages involved in the proposed methodology for semantic clustering and the derivation of generalized property-based encoders.

2.1.1 Data Pre-Processing

We retrieved the AAIndex database records from its official site https://www.genome.jp/ftp/db/community/aaindex/aaindex1. Then, we processed the dataset generating two **.csv* files to facilitate its handling, one containing numeric values for the properties and the other containing their description.

2.1.2 Unsupervised Learning

As classical clustering methods based on unsupervised learning using the values of the properties cannot ensure semantic consistency within the partition generated, we designed a staged process, combining them with doc2vec techniques to generate autoencoders. We first train a doc2vec autoencoder on the descriptions of the physicochemical properties in the AAIndex database and apply it over the same dataset to obtain embedding representations. We then explore different classical unsupervised learning algorithms and combinations of their hyperparameters (as described in the exploration stage in Medina-Ortiz et al. (2020b,c)) to obtain several candidate partitions of the dataset.

2.1.3 Selection of the Best Partition

We assessed the quality of each partition by obtaining their Calinsky-Harabasz indexes and selecting the one with the highest. Finally, we retrieve the original descriptions applying the inverse encoder (decoder) and review whether the condition of semantic consistency is met within the groups generated.

2.1.4 Encoder Creation

Using the partition generated in the previous step, we studied how property values are distributed for the different amino acids. We created a $20 \times N_i$ matrix containing the values of the N_i properties contained in the $i-$ th group for each amino acid. We then applied a kernel principal component analysis (kernel-PCA, radial basis function RBF-kernel with default settings) to the matrix representing each group. Noteworthy, a kernel-PCA expands the traditional PCA's limitations, such as requiring the components (namely, columns of the matrix) to distribute normally, and prevents the information loss that would cause removing those properties that do not meet this condition. Finally, we define encoders as the first component of each intra-group kernel-PCA.

2.2 Numerical Representation of Protein Sequences and Fast Fourier Applications

The general principle behind encodings is to map a categorical variable into a numeric value. In the context of protein engineering, encoding sequences of amino acids translates them into vectors. However, distance-based algorithms cannot capture the interactions between residues when comparing different sequences (or variants of the same sequence when replacing one or more of the constitutive amino acids). As we expect changes in one residue to impact the protein's function depending on who the neighbor residues were, we need a method to account for the impact that each amino acid has on the whole sequence. One way to capture this broad range of interactions is to use Fourier transforms (Sneddon, 1995).

Alongside other integral transforms, Fourier transforms search to represent functions (or vectors) as a superposition of other functions or vectors that form a basis of the correspondent space. For Fourier transforms, such a basis is all possible sinusoidal functions. Although it was originally thought for a continuously valued function, it is possible to define the Fourier transform and its inverse for discrete distributions. In this case, only a finite sample segment of the continuous data set is required to reconstruct the frequency spectrum (Rao and Yip, 2014).

The Fast Fourier transform (FFT) algorithm enables the efficient computation of the Fourier transform; Solving the problem directly from the discrete Fourier Transform (DFT) yields a complexity of $O(N^2)$, while using the FFT generates a complexity of $O(N \log N)$ (Welch, 1967). In the context of the present work, we apply FFT to each encoded sequence according to the following steps: 1) As required to apply FFT, we complete every vector with zeros (zero padding) such that the resulting dimension is $(2^n) - 1$. 2) We apply FFT to each resized vector independently, obtaining a $n \times \frac{m}{2}$ matrix of frequencies, where n is the number of sequences and m is the number of points in the vector. We then use the obtained frequencies as input to train predictive models.

2.3 Predictive Models Training

Throughout the different case studies presented in this work, we use Random Forest predictive models due to their easy implementation and interpretation. Hyperparameters are those of the default configuration of DMAKit (Medina-Ortiz et al., 2020b): n_estimatorsint = 100, criterion = gini, min_samples_split = 2, min_samples_leaf = 1, and n_jobs = −1 so that all available cores are used. After preprocessing, each input dataset was divided into training and testing datasets in an 80:20 ratio. For the performance assessment experiment, we repeated the 80:20 split of the dataset 1.000 times using different random seeds, aiming to compensate for any potential selection bias. Thus, instead of reporting a single value for model precision, we report, in this case, a distribution. Model training involves a k cross-validation stage, with k = 10. We also put forward a metric to assess overfitting, the overfitting ratio, defined as the ratio between model precision in the training and validation stages.

2.4 Testing Datasets and Case Studies

Here we describe the different datasets we evaluated in the case studies to assess the proposed encoders and methodology.

2.4.1 DNA-Binding Protein

DNA-binding protein (DBP) classification is one of the most exciting problems in biotechnology, mainly because of its implications in protein engineering, synthetic biology, molecular biology, and genetic engineering (Rahman et al., 2018). Furthermore, it finds direct application in the improvement of commercial DNA polymerases and restriction enzymes (Wei et al., 2017). Different computational methods to develop classification models for DNA-Binding protein have been proposed, involving various sequence coding and characterization strategies. Despite the enormous efforts aiming to solve this problem, it remains open. The dataset for this task was built using different previously reported datasets (Wei et al., 2017; Rahman et al., 2018; Adilina et al., 2019). We also removed all sequences without classification, generating a balanced dataset with 504 examples of DNA binding protein and 523 non-DNA binding protein.

2.4.2 Folding and Function Recognition

Two of the most common tasks in protein engineering are the prediction of the folding of secondary structures and the classification of protein function (Marchler-Bauer et al., 2017). Based on this, our approach was based on solving two questions of interest. 1) Given a set of proteins with the same folding, is it possible to recognize or predict the functions of these proteins? 2) Given a set of proteins with the same function, is it feasible to classify protein folding? First, we used the Protein Data Base (PDB) to build these data sets. Then, for each search, we applied the following filters: 1) *Homo sapiens* organism, 2) X-Ray diffraction as the experimental method, 3) Protein type as Polymer entity type, and 4) a resolution lower than 3 Å. Next, we implemented a bash script to download the protein sequences and save them in csv files for the different applications. Remarkably, we developed balanced datasets to reduce the possible problems in the training process in all cases.

2.4.3 Biological Activity Prediction for Peptide Sequences

Antimicrobial peptides (AMPs) are known as host-defense peptides (Sitaram and Nagaraj, 2002). These molecules play an essential role in the innate immune response, thus having direct application in the pharmaceutical, biotechnological, and industrial areas (Papagianni, 2003; Ma et al., 2018). Different computational methods based on ML have been developed to classify antimicrobial peptides (Xiao et al., 2013; Chen et al., 2016; Zimmer et al., 2018; Yi et al., 2019; Yi et al., 2019). In this case study, we used the peptide sequences reported in PeptipediaDB (Quiroz et al., 2021) to develop classification models of AMPs peptides, generating a dataset with six types of biological activities.

2.5 Implementation Strategies and Library Developing

Scripts to develop, assess, and exemplify the usage of the proposed encoders are written in *Python* v3.9, powered by libraries as Pandas (McKinney, 2010), Numpy, Gensim (Řehřek and Sojka, 2011), and DMAKit (Medina-Ortiz et al., 2020b), among others. The encoding library proposed in this work was designed under the Object-Oriented Programming paradigm (Wegner, 1990), which is advantageous for its modularity.

3 RESULTS AND DISCUSSION

3.1 Combining Text Mining and Unsupervised Learning Reveal Semantic Groups of Physicochemical Properties in the AAIndex Database

Using a combination of doc2vec strategies and unsupervised learning algorithms, we identified eight groups semantically-consistent groups of physicochemical properties within the AAIndex database (Kawashima and Kanehisa, 2000). By semantic consistency, we refer to these groups representing the same physical aspect of amino acids, such as general structural and thermodynamic properties and indices. To determine them, we explored about one million possible partitions of the dataset, changing the way of generating embeddings of property descriptions, the clustering algorithms, and their hyperparameters. We performed this using the model exploration tools presented in Medina-Ortiz et al. (2020b).

The resulting eight groups of properties were obtained by training autoencoders with hyperparameters of 500 epochs, a value of α = 0.025, and an embedding size of 2, and partitioning the dataset by applying the k − means algorithm with k = 8. This was the best performing algorithm found in the exploration stage, reaching a Calinski-Harabasz index of **1,532.36** and a silhouette coefficient of **0.43**. Finally, we assessed the semantic consistency of each group, evaluating whether the properties within the group presented the same contexts or specific words. As a result, only 17 descriptions were reclassified from the group of *Other indexes* to the groups of *α structure* and *β structure*.

One of the advantages of implementing a strategy based on doc2vec is the semanticity generated by separating the properties by their descriptions, which facilitates a simple visualization of the existing contexts or topics in each group. On the other hand, applying unsupervised learning algorithms to property values will generate partitions that do not ensure semantic consistency within groups, as clustering criteria will be numeric. In this way, the semantic clustering methodology proposed in this work ensures that the random selection of any member of a particular group will have the same physical meaning, otherwise not possible.

We analyzed the AAIndex database from a numeric perspective to test the statement above. We explored different combinations of unsupervised learning algorithms and hyperparameters to partition the dataset. The best performing algorithm was k − means (k = 2), yielding a Calinski-Harabasz index of **1,527.81** and a silhouette coefficient of **0.87**. Although these results hint at an excellent separation between the groups in the partition, not only is there no relationship between the descriptions within the groups, but also unbalanced divisions of properties between groups. Forcing the k − means algorithm to produce **eight** groups generates a partition with a Calinski-Harabasz index of **614.25** and a silhouette coefficient of **0.50**. However, and as expected, no semantic consistency within the groups was achieved.

Once the groups of descriptions were generated and corrected, these were used to generate eight data sets with the property values for each amino acid. We applied a kernel-PCA (Radial Basis Function–RBF–kernel) to the numeric values of each group and assessed how much of the variance was explained by the first component. In all groups, the variance explained by the first component of the kernel-PCA was higher than **85%**. Furthermore, the different groups resulted in being linearly separable in the PCA1/PCA2 space, as their convex hulls are disjoint (cf. **Figure 1A**). Therefore, we proposed to use the first component of each semantic group of properties generated as an encoder. These encoders are listed in **Table 1**.

3.2 Semantically-Consistent Encoders and Fourier Transform Facilitate Predictive Tasks in Protein Engineering

We used the proposed semantic encoders to tackle four different predictive tasks in protein engineering (DNA-

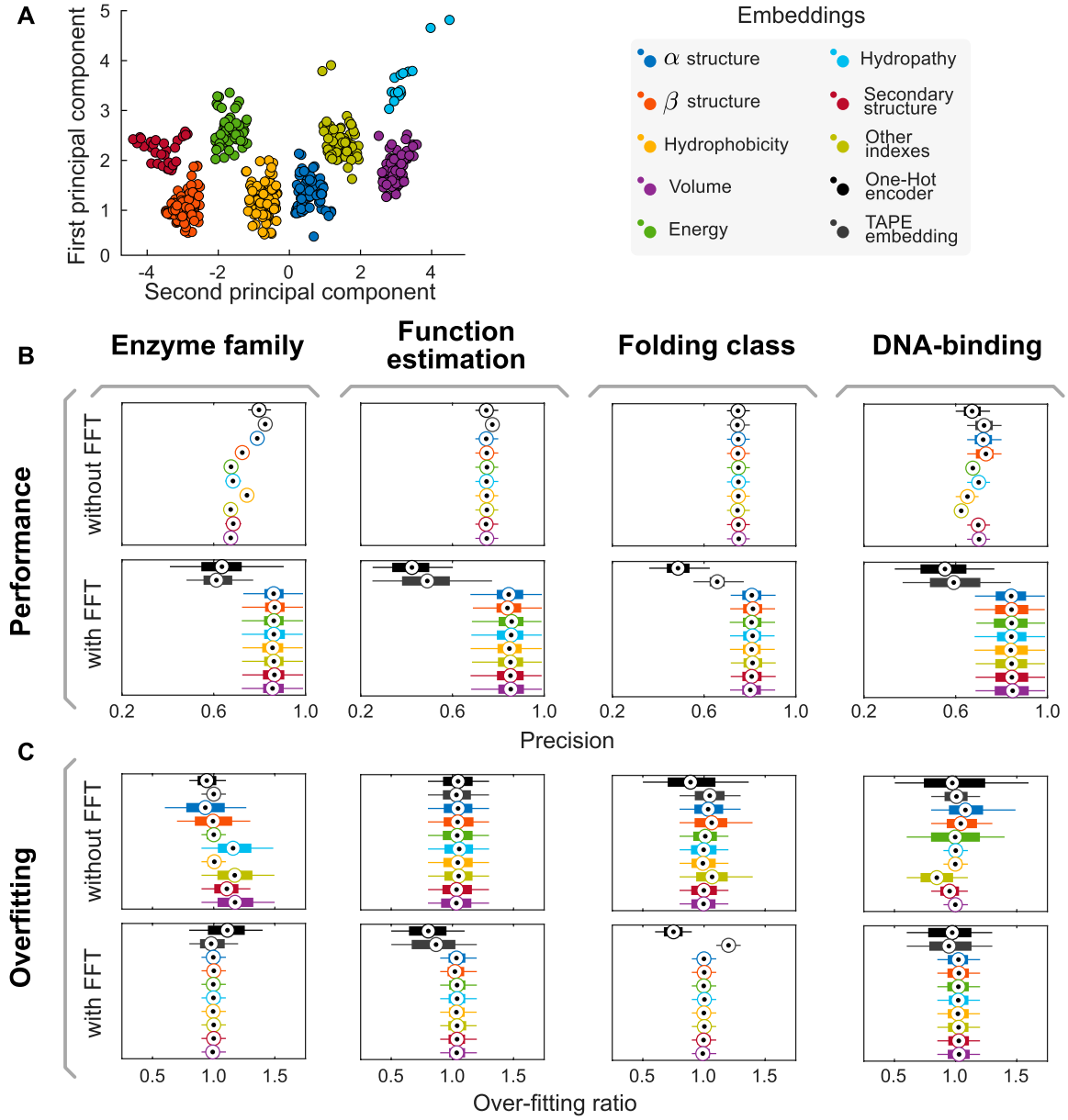


FIGURE 1 | The AAIndex database of amino acid physicochemical properties can be split into eight semantically-consistent groups. **(A)** Combining doc2vec strategies with unsupervised learning algorithms, we proposed a methodology to generate groups that preserve semantic consistency within the partition. Applying an RBF kernel PCA on the whole dataset, we observe that the groups are linearly separable in the PCA1/PCA2 space, as their convex hulls are disjoint. **(B,C)** Combining our encoders with FFT improves model performance and helps reducing overfitting in several predictive tasks. Here, boxplots summarize the distribution of performances reached in each experiment across the 1,000 independent realizations of the 80/20 split of the input dataset for the task. Central circles represent medians, bars the interquartile range, and whiskers the 95% CI. Complementary analyses of model performance, including other metrics (such as recall, F-Score, and area under the receiver operating curves AUC), are presented in **Supplementary Section S3** and summarized in **Supplementary Tables S3–S6**.

binding protein classification, protein folding, protein function, and enzyme family determination) using Random Forest algorithms. The datasets and hyperparameters are described in the Methods section. First, input datasets were split into training and validation datasets in an 80: 20 proportion. Then, aiming to prevent any stochastic artifact induced by a favorable/unfavorable partition of the dataset, we repeated this stage 1,000 times using different random seeds. Thus, instead of obtaining a single value for the performance of a model, we obtained a distribution of performances (cf. **Figures 1B,C**). When comparing model performance achieved using our encoders with that of models trained with classical methods (e.g., One Hot Encoder (Broom et al., 2017) and TAPE embeddings (Rao et al., 2019)), there is

TABLE 1 | Generalized property-based encoders for amino acids.

Amino acid	α structure	β structure	Hydrophobicity	Volume	Energy	Hydropathy	Secondary structure	Other indexes
A	290.41	71.85	6.25	44.65	−107.79	15.33	56.16	92.92
R	172.57	−6.96	84.09	200.15	51.15	172.36	1.44	−37.39
N	−38.37	−90.14	−21.73	−191.18	73.94	−259.13	−54.69	−77.74
D	159.43	−56.58	−28.96	−232.26	55.36	−216.01	−29.38	−7.42
C	−4.24	15.67	−34.88	−156.21	−54.19	−242.01	10.07	40.04
Q	−268.55	−32.61	38.46	179.88	31.44	145.73	-15.43	−45.52
E	−0.02	21.03	−21.48	−170.44	−49.97	8.11	20.20	50.74
G	−104.49	−62.33	53.16	250.66	92.25	256.52	-39.89	−95.41
H	−159.87	31.27	−69.67	194.47	−39.54	455.61	34.12	43.37
I	−34.08	164.64	−54.85	−88.56	−48.44	−274.76	25.05	52.40
L	−91.11	−16.38	−64.98	−201.08	7.56	−257.27	−10.20	4.27
K	195.59	54.45	−52.92	−118.84	−109.99	−136.28	55.31	85.66
M	21.94	−18.77	−26.70	−227.61	−7.39	−139.71	−19.45	16.04
F	88.02	21.61	−21.46	−78.96	−56.97	80.68	30.31	46.42
P	317.10	115.37	−22.23	−44.80	−157.63	−126.45	95.69	136.09
S	−314.20	−106.56	61.31	221.12	174.08	248.05	−85.57	−122.66
T	−252.51	−23.99	13.72	−3.30	17.50	−153.13	−25.56	−31.46
W	−118.15	−76.02	88.28	34.80	105.47	19.24	−59.91	−124.49
Y	−10.20	−15.49	40.85	203.07	36.61	171.61	−4.25	−33.07
V	150.75	9.929	33.77	184.45	−13.45	231.50	15.99	7.21

no major difference (cf. **Figure 1B**). However, our models reached over-fitting ratios (defined as the performance of training divided by the performance in validation) closer to one than classical approaches, suggesting that our encoders are better suited for these predictive tasks (cf. **Figure 1C**).

We repeated the experiments above but applied the Fast Fourier Transform (FFT) to the encoded sequences before training predictive models. We then use the complex modulus of the discrete Fourier transform as a feature to train our models. By doing so, we aim to capture the influence of the position of each amino acid within the sequence, which affects other amino acids in different ranges of influence. We will provide further details on the interpretation of the FFT-related variables in the next section. While there is a drop in performance when combining One Hot and embedding-based encoders and FFT, the use of FFT increased the precision of predictive models trained using the semantic encoders herein proposed. Moreover, the over-fitting ratio decreases even further in this case, suggesting a synergistic effect on the predictive performance of trained models.

A possible interpretation of this effect relates to the Fourier transform's properties, which capture the influence of each component of the input on the others, thereby incorporating more information into the predictive systems. Amino acids within a protein sequence influence each other. Thus, by applying Fourier Transforms, we can capture, to some extent, this spatial dependency. Furthermore, this property results beneficial for any property-based encoding strategy, as previously reported in Siedhoff et al. (2021), Cadet et al. (2018b), and Cosic (1994). Based on the above, we propose the combination of our encoders together with the application of Fourier transforms in order to improve the performance of predictive models.

We performed a complementary model evaluation analyzing the whole spectra of performance metrics, including recall, F-Score, and area under the receiver operating curves AUC. We found a marked consistency between the precision and recall obtained by trained models, and these metrics were further increased when training models in the frequency space. Furthermore, the high values reached for the AUC across predictive tasks highlight the predictive power of our approach. The reader is referred to **Supplementary Section S3**, and **Supplementary Tables S3–S6**.

We compared the performance of our encoders and similar approaches to assess whether we reached a sweet spot regarding the number of proposed encoders and information contained therein. In particular, we compare our results against 1) using all properties in the AAIndex as independent encoders and 2) applying a linear PCA directly on the AAIndex database and using the most informative components as independent encoders. The reader is referred to **Supplementary Section S4**.

3.3 The Combination of Our Encoders With FFT Allow Detecting Profiles Related to Folding and Protein Functions

Combining the encoders proposed in this work, the interpretation of protein sequences as signals, and processing them after applying FFT, facilitates the identification of profiles at the folding and functional levels. To demonstrate this, we propose the following case study. We encode the *protein function* dataset employing the *secondary structure*-related encoder and apply FFT on it to analyse its spectra. In particular, we sought to find relations for the mean complex modulus of the transformed signals of different families of enzymes (in this case, hydrolase and ligase), and used their length to x-scale the frequency, zoomed to the active site, and excluded extremes that could be affected by either zero-padding or border conditions. We found a clear difference between the mean complex modulus of the Fourier spectra of ligases and hydrolases (cf. **Figure 2A**).

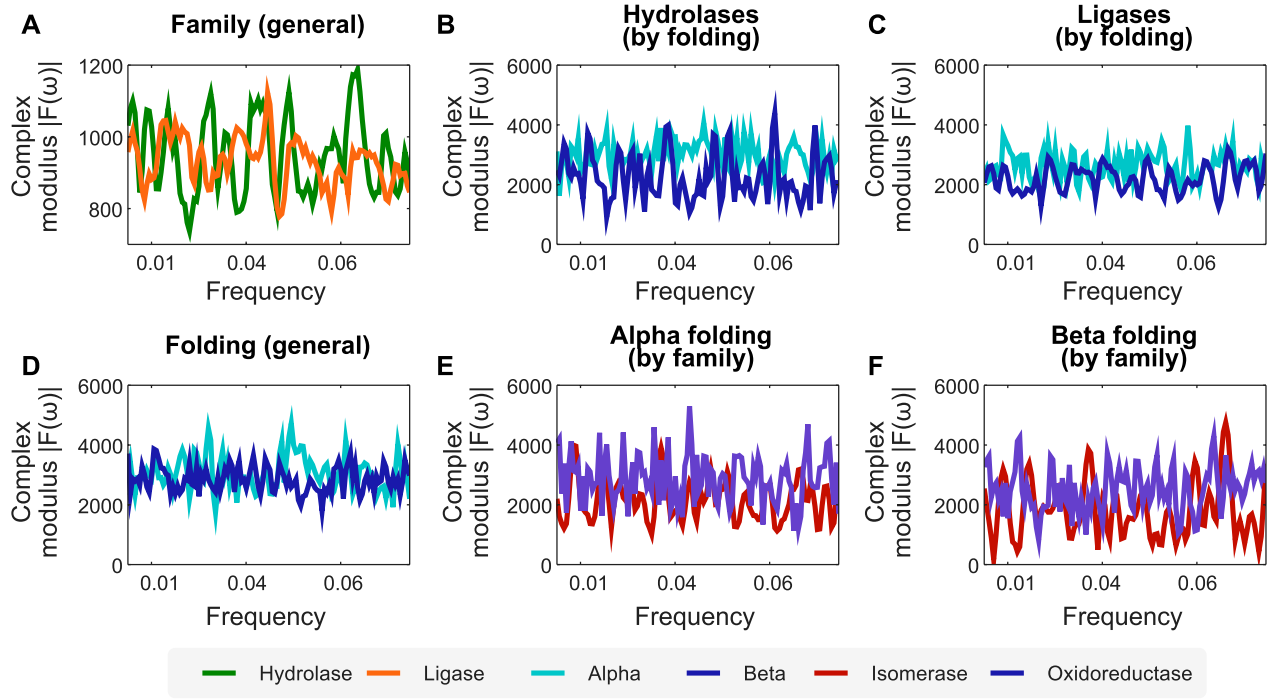


FIGURE 2 | The combination of our encoders with FFT unveils frequency profiles associated to specific protein folding and functions. We used the encoder of secondary structure combined with FFT to create profiles related to folding and protein functions. **(A)** Fourier spectra for two family enzymes (hydrolases and ligases) in a dataset of enzyme families. **(B,C)** Fourier spectra of the same family separated by folding, showing that our methodology is sensitive to apparent differences between alpha and beta folding types. **(D)** Fourier spectra for alpha and beta folding in a dataset of different protein families. **(E,F)** Fourier spectra of the same folding separated by protein family, showing that our methodology is sensitive to proteins with the same folding but belonging to different families. *N* for frequency normalization = 1,024.

Furthermore, we also found differences in the mean complex modulus of the Fourier spectra when analyzing two folding classes within the same family (cf. **Figures 2B,C**).

We employ the same approach (*secondary structure*-related encoder combined with FFT) to identify protein folding profiles and sub-profiles related to protein function. **Figure 2D** shows the average spectra for the α and β folds of enzymes with different functions. Similarly, we found that isomerases and oxidoreductases have slightly different mean Fourier spectra, although sharing the α and β folding properties (cf. **Figures 2E,F**).

3.4 Towards a New Design Strategy for Protein Sequences With Desirable Properties

One of the most challenging problems in protein engineering is protein design (Yang et al., 2019). Considering the advantages of combining our semantic encoders and FFT, we put forward a prospective methodology to design peptide sequences with desired properties. In this case study, we illustrate the use of this methodology to design peptides with antimicrobial activity. Using the antimicrobial peptide dataset described in Methods, we apply our encoders and FFT to the dataset and trained two random forest predictive models. The first model is a binary classification model for antimicrobial activity, while the second corresponds to a multi-class model of various biological activities for antimicrobial peptides. The latter include peptide classes such as antibacterial, anti-viral, anti-cancer, anti-HIV, and anti-fungal. The models had an accuracy of respectively **95.3%** and **89.41%**. On the one hand, the clear separation between the spectra of antimicrobial and non-antimicrobial peptides explains the high performance reached by the binary classifier. On the other hand, marked patterns for each biological activity facilitate the generalization in the multi-class model (cf. **Figure 3**, where panels A–I represent the different encoders proposed herein). Altogether, when analyzing the distribution of values for each position, we can define a *latent space* where, theoretically, encoded signals with the same complex modulus would have the same activity.

After characterizing the classification mechanisms of the models described above, we put forward the following methodology to generate new sequences that would be classified as "having an activity" by them. First, we collect different peptide sequences with antimicrobial activity from the Peptipedia database (Quiroz et al., 2021). These sequences are new examples for the classifiers, as they were not used during the model training step. Alternatively, another way to generate new sequences for this stage is through deep generative models (Wu et al., 2021). Note that we already know that these sequences do have antimicrobial activity. These build up a $m \times n$ matrix,

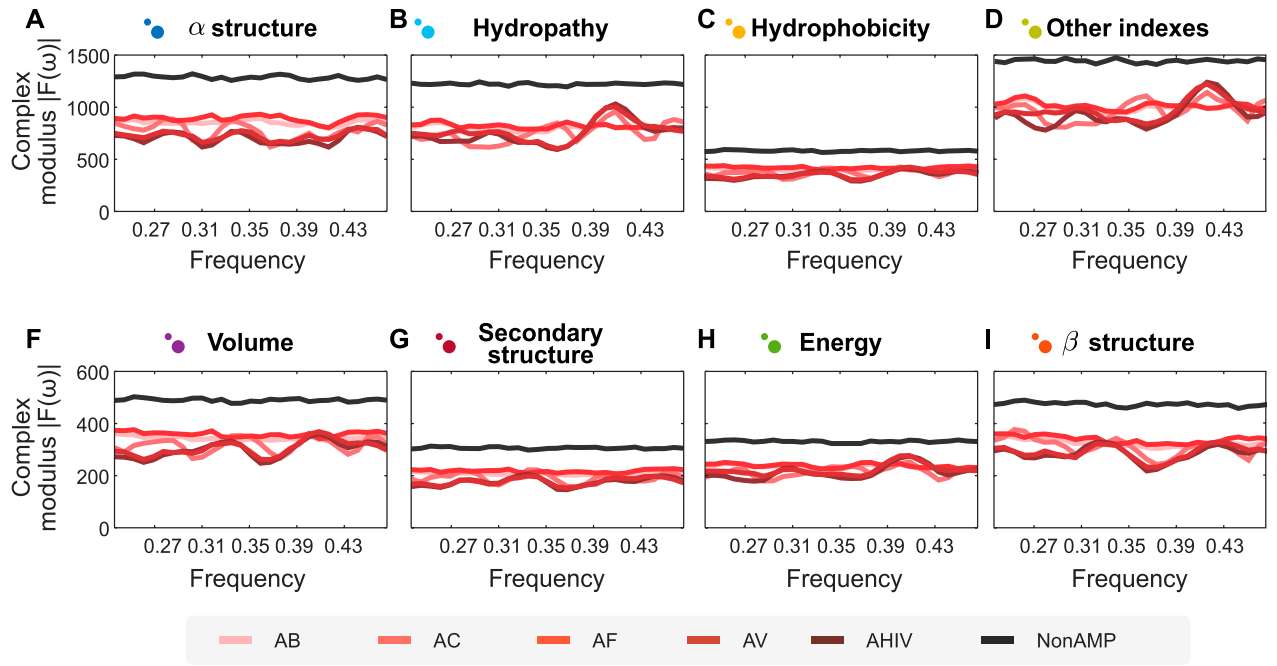


FIGURE 3 | Fourier spectra of encoded amino acid sequences with different activities are visually separated. Sub figures show the Fourier spectrum of different sequences of peptides, encoded according to the groups of properties proposed in this article, represented in panels **(A–I)**. We analyse two types of peptides: Antimicrobial (AMPs) and non-Antimicrobial (nonAMP). AMPs are subsequently divided into five categories: Antibacterial Peptides (AB), Anticancer Peptides (AC), Antifungal Peptides (AF), Anti-HIV Peptides (AHIV), and Antiviral Peptides (AV). The signals analyzed show a clear differentiation for AMPs concerning nonAMPs. *N* for frequency normalization = 128.

where m are the number of sequences and n the length of the longest of those (all others are completed with zeros). Second, we encode and transform the sequences using all the proposed encoders in this work separately to obtain 8 $m \times n$ matrices. Third, we characterize the distribution of values column-wise for each matrix, so we obtain confidence intervals for the encoded values of each position. Fourth, we calculate the likelihood of new sequences belonging to each category's latent space for each encoder. Precisely, for each residue in the sequence, we calculate a probability. Assuming that all these are independent, the t probability of belonging to the latent space is the multiplication of the individual probabilities obtained for each position. In this way, we have eight statistical tests where belonging to a latent space could predict a unique biological activity. Fifth, we used the trained model to predict the category of the proposal sequence. Finally, we evaluate the predictions and check if the proposed sequences are classified in the class of interest. A step-by-step, in-depth explanation of the proposed methodology and a summary flowchart can be found in **Supplementary Section S4**.

Using the proposed strategy, we randomly explored 10,000 sequences. We defined a selection criterion of 90% probability of existing within the latent space of desirable biological activity, in this case, antimicrobial peptides and their different subcategories. Of the 10,000 sequences explored, only **3,513** met the established probability criteria, and their activity was predicted using the previously trained models. Remarkably, because the biological activities of the sequences were known in advance, the performance of the screening methodology could be evaluated by comparing the predicted rankings with the biological activities reported by each sequence. Performance metrics are reported in **Supplementary Table S3**. Notably, the sequences recognized as antimicrobial peptides showed performance similar to the training result. However, the rest of the biological activities evaluated showed a decrease concerning the predictive model. This is not surprising since the models were trained using sequences that only presented a specific activity, while the evaluated sequences showed primarily moonlight activity (which is why the sub-activities of antimicrobial peptides do not add up to the total number of antimicrobial peptides evaluated sequences). Despite these results, the proposed methodology facilitates the exploration of new sequences from a probabilistic point of view, being enormously efficient for antimicrobial peptides and promising for future applications.

CONCLUSION

The results presented in this work can be summarized as three main contributions. First, we extend the traditional property-based encoding strategy and propose eight new encoders that represent semantically-consistent groups of physicochemical properties of the AAIndex database. Second, we illustrate how

using these encoders together with Fourier transforms can substantially improve the performance of machine learning models in general predictive tasks in protein engineering. Furthermore, we found a synergistic interaction between the proposed encoders and the FFT that simultaneously increases the precision of the trained models while reducing their overfitting to the data. Finally, we put forward a simple and preliminary statistically-based methodology to create *de novo* peptide and protein sequences with desirable properties. We will extend the modeling framework to simultaneously use the eight encoders to tackle more complex predictive tasks in protein engineering in future work. We expect these independent descriptions of a sequence to interact synergistically and increase model performance.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

Conceptualization, DM-O, SC, AO-N; methodology, DM-O, JT-A; validation AO-N, DM-O, JA-H, MN, SC; investigation, DM-O, JA-H, SC; writing, review and editing, DM-O, SC, JA-H, JA, MN, AO-N; supervision, AO-N, DM-O; project administration, AO-N, DM-O, SC; funding resources, AO-N, SC.

FUNDING

Open Access (OA) publication of this work has been enabled by the Max-Planck-Society. The authors gratefully acknowledge support from the Centre for Biotechnology and Bioengineering - CeBiB (PIA project FB0001, Conicyt, Chile). DM-O gratefully acknowledges ANID, Chile, for Ph.D. fellowship 21181435. DM-O received support from Universidad de Magallanes (project MAG-2095). SC received support from the Max-Planck-Society. JA-H gratefully acknowledges ANID, for Ph.D. fellowship 21182109. MN and JT-A gratefully acknowledge ANID, Chile for Fondecyt 1180882 project and Universidad de Magallanes for 2095 project. Powered@NLHPC; This research was partially supported by the supercomputing infrastructure of the National Laboratory for High-Performance Computing, NLHPC (ECM-02), Chile.

ACKNOWLEDGMENTS

Part of the content of this manuscript was released as a preprint at Medina-Ortiz et al. (2020a)

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.898627/full#supplementary-material

REFERENCES

Adilina, S., Farid, D. M., and Shatabda, S. (2019). Effective DNA Binding Protein Prediction by Using Key Features via Chou's General PseAAC. *J. Theor. Biol.* 460, 64–78. doi:10.1016/j.jtbi.2018.10.027

Ancien, F., Pucci, F., Godfroid, M., and Rooman, M. (2018). Prediction and Interpretation of Deleterious Coding Variants in Terms of Protein Structural Stability. *Sci. Rep.* 8, 4480. doi:10.1038/s41598-018-22531-2

Broom, A., Jacobi, Z., Trainor, K., and Meiering, E. M. (2017). Computational Tools Help Improve Protein Stability but with a Solubility Tradeoff. *J. Biol. Chem.* 292, 14349–14361. doi:10.1074/jbc.m117.784165

Brownlee, J. (2020). Ordinal and One-Hot Encodings for Categorical Data. Machine Learning Mastery. Available at: https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/.

Cadet, F., Fontaine, N., Li, G., Sanchis, J., Ng Fuk Chong, M., Pandjaitan, R., et al. (2018a). A Machine Learning Approach for Reliable Prediction of Amino Acid Interactions and its Application in the Directed Evolution of Enantioselective Enzymes. *Sci. Rep.* 8, 16757–16815. doi:10.1038/s41598-018-35033-y

Cadet, F., Fontaine, N., Vetrivel, I., Ng Fuk Chong, M., Savriama, O., Cadet, X., et al. (2018b). Application of Fourier Transform and Proteochemometrics Principles to Protein Engineering. *BMC Bioinform.* 19, 382. doi:10.1186/s12859-018-2407-8

Chen, W., Ding, H., Feng, P., Lin, H., and Chou, K.-C. (2016). Iacp: a Sequence-Based Tool for Identifying Anticancer Peptides. *Oncotarget* 7, 16895–16909. doi:10.18632/oncotarget.7815

Cosic, I., Cosic, D., and Lazar, K. (2016). Analysis of Tumor Necrosis Factor Function Using the Resonant Recognition Model. *Cell Biochem. Biophys.* 74, 175–180. doi:10.1007/s12013-015-0716-3

Cosic, I. (1994). Macromolecular Bioactivity: Is it Resonant Interaction between Macromolecules?-Theory and Applications. *IEEE Trans. Biomed. Eng.* 41, 1101–1114. doi:10.1109/10.335859

Cosic, I., and Nesic, D. (1987). Prediction of 'hot Spots' in SV40 Enhancer and Relation with Experimental Data. *Eur. J. Biochem.* 170, 247–252. doi:10.1111/j.1432-1033.1987.tb13692.x

Eisenberg, D., Weiss, R. M., and Terwilliger, T. C. (1984). The Hydrophobic Moment Detects Periodicity in Protein Hydrophobicity. *Proc. Natl. Acad. Sci. U.S.A.* 81, 140–144. doi:10.1073/pnas.81.1.140

Forghani, M., and Khani, R. (2017). "A Multivariate Clustering of Aaindex Database for Protein Numerical Representation," in *2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS)* (IEEE), 1–4. doi:10.1109/icspis.2017.8311579

Georgiev, A. G. (2009). Interpretable Numerical Descriptors of Amino Acid Space. *J. Comput. Biol.* 16, 703–723. doi:10.1089/cmb.2008.0173

Hejase de Trad, C., Fang, Q., and Cosic, I. (2000). The Resonant Recognition Model (Rrm) Predicts Amino Acid Residues in Highly Conserved Regions of the Hormone Prolactin (Prl). *Biophys. Chem.* 84, 149–157. doi:10.1016/S0301-4622(00)00109-5

Kawashima, S., and Kanehisa, M. (2000). Aaindex: Amino Acid Index Database. *Nucleic Acids Res.* 28, 374. doi:10.1093/nar/28.1.374

Khoury, G. A., Smadbeck, J., Kieslich, C. A., and Floudas, C. A. (2014). Protein Folding and De Novo Protein Design for Biotechnological Applications. *Trends Biotechnol.* 32, 99–109. doi:10.1016/j.tibtech.2013.10.008

Kieslich, C. A., Alimirzaei, F., Song, H., Do, M., and Hall, P. (2021). Data-driven Prediction of Antiviral Peptides Based on Periodicities of Amino Acid Properties. *Comput. Aided Chem. Eng.* 50, 2019–2024. doi:10.1016/b978-0-323-88506-5.50312-0

Kim, D., Seo, D., Cho, S., and Kang, P. (2019). Multi-co-training for Document Classification Using Various Document Representations: TF-IDF, LDA, and Doc2Vec. *Inf. Sci.* 477, 15–29. doi:10.1016/j.ins.2018.10.006

Ma, C., Ren, Y., Yang, J., Ren, Z., Yang, H., and Liu, S. (2018). Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Anal. Chem.* 90, 10881–10888. doi:10.1021/acs.analchem.8b02386

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., et al. (2017). Cdd/sparcle: Functional Classification of Proteins via Subfamily Domain Architectures. *Nucleic Acids Res.* 45, D200–D203. doi:10.1093/nar/gkw1129

McKinney, W. (2010). "Data Structures for Statistical Computing in python," in *Proceedings of the 9th Python in Science Conference*. Editors S. van der Walt and J. Millman, 51–56. doi:10.25080/majora-92bf1922-00a

Medina-Ortiz, D., Contreras, S., Amado-Hinojosa, J., Torres-Almonacid, J., Asenjo, J. A., Navarrete, M., et al. (2020a). Combination of Digital Signal Processing and Assembled Predictive Models Facilitates the Rational Design of Proteins. *arXiv* [Preprint]. Available at: https://arxiv.org/abs/2010.03516.

Medina-Ortiz, D., Contreras, S., Quiroz, C., Asenjo, J. A., and Olivera-Nappa, Á. (2020b). Dmakit: A User-Friendly Web Platform for Bringing State-Of-The-Art Data Analysis Techniques to Non-specific Users. *Inf. Syst.* 93, 101557. doi:10.1016/j.is.2020.101557

Medina-Ortiz, D., Contreras, S., Quiroz, C., and Olivera-Nappa, Á. (2020c). Development of Supervised Learning Predictive Models for Highly Non-linear Biological, Biomedical, and General Datasets. *Front. Mol. Biosci.* 7, 13. doi:10.3389/fmolb.2020.00013

Papagianni, M. (2003). Ribosomally Synthesized Peptides with Antimicrobial Properties: Biosynthesis, Structure, Function, and Applications. *Biotechnol. Adv.* 21, 465–499. doi:10.1016/s0734-9750(03)00077-6

Pavelka, A., Chovancova, E., and Damborsky, J. (2009). Hotspot Wizard: a Web Server for Identification of Hot Spots in Protein Engineering. *Nucleic acids Res.* 37, W376–W383. doi:10.1093/nar/gkp410

Potapov, V., Cohen, M., and Schreiber, G. (2009). Assessing Computational Methods for Predicting Protein Stability upon Mutation: Good on Average but Not in the Details. *Protein Eng. Des. Sel.* 22, 553–560. doi:10.1093/protein/gzp030

Quiroz, C., Saavedra, Y. B., Armijo-Galdames, B., Amado-Hinojosa, J., Olivera-Nappa, Á., Sanchez-Daza, A., et al. (2021). Peptipedia: a User-Friendly Web Application and a Comprehensive Database for Peptide Research Supported by Machine Learning Approach. *Database (Oxford)* 2021. doi:10.1093/database/baab055

Rackovsky, S. (1998). "Hidden" Sequence Periodicities and Protein Architecture. *Proc. Natl. Acad. Sci. U.S.A.* 95, 8580–8584. doi:10.1073/pnas.95.15.8580

Rahman, M. S., Shatabda, S., Saha, S., Kaykobad, M., and Rahman, M. S. (2018). DPP-PseAAC: A DNA-Binding Protein Prediction Model Using Chou's General PseAAC. *J. Theor. Biol.* 452, 22–34. doi:10.1016/j.jtbi.2018.05.006

Rao, K. R., and Yip, P. (2014). *Discrete Cosine Transform: Algorithms, Advantages, Applications*. Academic Press.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., et al. (2019). "Evaluating Protein Transfer Learning with Tape," in *Advances in Neural Information Processing Systems*.

Řehřek, R., and Sojka, P. (2011). *Gensim—Statistical Semantics in Python*. Retrieved from genism.org.

Sadowski, M. I., and Jones, D. T. (2009). The Sequence-Structure Relationship and Protein Function Prediction. *Curr. Opin. Struct. Biol.* 19, 357–362. doi:10.1016/j.sbi.2009.03.008

Saha, I., Maulik, U., Bandyopadhyay, S., and Plewczynski, D. (2012). Fuzzy Clustering of Physicochemical and Biochemical Properties of Amino Acids. *Amino acids* 43, 583–594. doi:10.1007/s00726-011-1106-9

Siedhoff, N. E., Illig, A.-M., Schwaneberg, U., and Davari, M. D. (2021). PyPEF-An Integrated Framework for Data-Driven Protein Engineering. *J. Chem. Inf. Model.* 61, 3463–3476. doi:10.1021/acs.jcim.1c00099

Siedhoff, N. E., Schwaneberg, U., and Davari, M. D. (2020). Machine Learning-Assisted Enzyme Engineering. *Methods Enzym.* 643, 281–315. doi:10.1016/bs.mie.2020.05.005

Sitaram, N., and Nagaraj, R. (2002). Host-defense Antimicrobial Peptides: Importance of Structure for Activity. *Curr. Pharm. Des.* 8, 727–742. doi:10.2174/1381612023395358

Sneddon, I. N. (1995). *Fourier Transforms*. Courier Corporation. New York: Dover Publications.

Veljkovic, V., Cosic, I., Dimitrijevicand Lalovic, D. (1985). Is it Possible to Analyze Dna and Protein Sequences by the Methods of Digital Signal Processing? *IEEE Trans. Biomed. Eng.* 32, 337–341. doi:10.1109/tbme.1985.325549

Wegner, P. (1990). Concepts and Paradigms of Object-Oriented Programming. *SIGPLAN OOPS Mess.* 1, 7–87. doi:10.1145/382192.383004

Wei, L., Tang, J., and Zou, Q. (2017). Local-dpp: An Improved Dna-Binding Protein Prediction Method by Exploring Local Evolutionary Information. *Inf. Sci.* 384, 135–144. doi:10.1016/j.ins.2016.06.026

Welch, P. (1967). The Use of Fast Fourier Transform for the Estimation of Power Spectra: a Method Based on Time Averaging over Short, Modified Periodograms. *IEEE Trans. Audio Electroacoust.* 15, 70–73. doi:10.1109/tau.1967.1161901

Winter, G. (1998). Synthetic Human Antibodies and a Strategy for Protein Engineering. *FEBS Lett.* 430, 92–94. doi:10.1016/s0014-5793(98)00628-0

Wittmann, B. J., Johnston, K. E., Wu, Z., and Arnold, F. H. (2021). Advances in Machine Learning for Directed Evolution. *Curr. Opin. Struct. Biol.* 69, 11–18. doi:10.1016/j.sbi.2021.01.008

Wu, Z., Johnston, K. E., Arnold, F. H., and Yang, K. K. (2021). Protein Sequence Design with Deep Generative Models. *Curr. Opin. Chem. Biol.* 65, 18–27. doi:10.1016/j.cbpa.2021.04.004

Xiao, X., Wang, P., Lin, W.-Z., Jia, J.-H., and Chou, K.-C. (2013). iamp-2l: a Two-Level Multi-Label Classifier for Identifying Antimicrobial Peptides and Their Functional Types. *Anal. Biochem.* 436, 168–177. doi:10.1016/j.ab.2013.01.019

Yang, K. K., Wu, Z., and Arnold, F. H. (2019). Machine-learning-guided Directed Evolution for Protein Engineering. *Nat. Methods* 16, 687–694. doi:10.1038/s41592-019-0496-6

Yang, K. K., Wu, Z., Bedbrook, C. N., and Arnold, F. H. (2018). Learned Protein Embeddings for Machine Learning. *Bioinformatics* 34, 2642–2648. doi:10.1093/bioinformatics/bty178

Yi, H.-C., You, Z.-H., Zhou, X., Cheng, L., Li, X., Jiang, T.-H., et al. (2019). Acp-dl: a Deep Learning Long Short-Term Memory Model to Predict Anticancer Peptides Using High-Efficiency Feature Representation. *Mol. Ther. - Nucleic Acids* 17, 1–9. doi:10.1016/j.omtn.2019.04.025

Zimmer, D., Schneider, K., Sommer, F., Schroda, M., and Mühlhaus, T. (2018). Artificial Intelligence Understands Peptide Observability and Assists with Absolute Protein Quantification. *Front. Plant Sci.* 9, 1559. doi:10.3389/fpls.2018.01559

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Medina-Ortiz, Contreras, Amado-Hinojosa, Torres-Almonacid, Asenjo, Navarrete and Olivera-Nappa. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 14 July 2022
doi: 10.3389/fmolb.2022.866676

Time-Optimal Adaptation in Metabolic Network Models

Markus A. Köbis[1]*, *Alexander Bockmayr*[2] and *Ralf Steuer*[3]*

[1]Research Group Dynamical Systems and Numerical Analysis, Department of Mathematics, Norwegian University of Science and Technology, Trondheim, Norway, [2]Mathematics in Life Science Group, Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany, [3]Humboldt-University of Berlin, Institute for Biology, Institute for Theoretical Biology (ITB), Berlin, Germany

Analysis of metabolic models using constraint-based optimization has emerged as an important computational technique to elucidate and eventually predict cellular metabolism and growth. In this work, we introduce time-optimal adaptation (TOA), a new constraint-based modeling approach that allows us to evaluate the fastest possible adaptation to a pre-defined cellular state while fulfilling a given set of dynamic and static constraints. TOA falls into the mathematical problem class of time-optimal control problems, and, in its general form, can be broadly applied and thereby extends most existing constraint-based modeling frameworks. Specifically, we introduce a general mathematical framework that captures many existing constraint-based methods and define TOA within this framework. We then exemplify TOA using a coarse-grained self-replicator model and demonstrate that TOA allows us to explain several well-known experimental phenomena that are difficult to explore using existing constraint-based analysis methods. We show that TOA predicts accumulation of storage compounds in constant environments, as well as overshoot uptake metabolism after periods of nutrient scarcity. TOA shows that organisms with internal temporal degrees of freedom, such as storage, can in most environments outperform organisms with a static intracellular composition. Furthermore, TOA reveals that organisms adapted to better growth conditions than present in the environment ("optimists") typically outperform organisms adapted to poorer growth conditions ("pessimists").

Keywords: constraint-based modeling, cellular metabolism, flux balance analysis, resource balance analysis, dynamic enzyme-cost flux balance analysis, optimal control, overshoot metabolism, luxury uptake

OPEN ACCESS

Edited by:
Alberto Jesus Martin, Universidad Mayor, Chile

Reviewed by:
Jean-Luc Gouzé, Research Centre Inria Sophia Antipolis Méditerranée, France
Pedro Andres Saa, Pontificia Universidad Católica de Chile, Chile

***Correspondence:**
Markus A. Köbis
markus.kobis@ntnu.no
Ralf Steuer
ralf.steuer@hu-berlin.de

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: *31 January 2022*
Accepted: *31 May 2022*
Published: *14 July 2022*

Citation:
Köbis MA, Bockmayr A and Steuer R (2022) Time-Optimal Adaptation in Metabolic Network Models. Front. Mol. Biosci. 9:866676. doi: 10.3389/fmolb.2022.866676

1 INTRODUCTION

Over the past decades, various modeling frameworks have been proposed to understand the organization and functioning of cellular metabolism and growth. Among the most popular approaches are constraint-based methods, in particular flux balance analysis (FBA) (Orth et al., 2010). Constraint-based methods typically make use of optimality principles that are motivated by evolutionary arguments. That is, instead of requiring a detailed mechanistic understanding of the underlying regulatory machinery, properties of cellular metabolism, such as exchange fluxes or biomass accumulation, are predicted based on the assumption that metabolism has evolved according to certain evolutionary optimality principles.

More recently, constraint-based methods have been extended to quantitatively account for the synthesis costs of the biological macromolecules that are required for cellular metabolism and

growth, giving rise to resource balance analysis (RBA) (Goelzer et al., 2011) and integrated reconstructions of Metabolism and macromolecular Expression (ME) (Lerman et al., 2012). While the initial approaches were restricted to time-invariant environments and subject to steady-state conditions, various dynamic extensions have also been proposed, such as dynamic FBA (dFBA) Mahadevan et al. (2002), dynamic enzyme-cost FBA (deFBA) (Waldherr et al., 2015), conditional FBA (cFBA) (Rügen et al., 2015; Reimers et al., 2017), dynamic RBA (dRBA) (Jeanne et al., 2018), dynamic ME (Yang et al., 2019), and regulatory dynamic enzyme-cost FBA (r-deFBA) (Liu and Bockmayr, 2020). These dynamic frameworks are computationally more expensive and allow predicting time courses over a given time interval, such that the variables fulfil a given (linear) optimality principle. Typically, within these frameworks, the time intervals over which the solutions are considered are predefined.

In this work, we extend these existing approaches and propose time-adaptation (TOA) as a new constraint-based modeling framework that allows us to evaluate the fastest possible adaptation to a pre-defined cellular state while fulfilling a given set of dynamic and static constraints. If the underlying dynamics of the biological system are governed by ordinary differential equations (ODEs) subject to algebraic constraints such as positivity, that is, so-called differential-algebraic equations (DAEs), time-optimal adaptation falls into the mathematical problem class of time-optimal control problems, which are optimal control problems where the time-interval is part of the objective (Hermes and Lasalle, 1969). In its general form, TOA can be applied in a very broad sense and thereby extends most of the existing constraint-based modeling frameworks.

Our approach allows us to compute feasible time courses to simulate or predict adaptations of cellular metabolism to environmental shifts. Potential applications include an analysis of cellular doubling, i.e., to analyze the optimal metabolic trajectory that results in a doubling of all cellular components in the shortest time, as well as an analysis of the temporal adaptation to changing nutrient availability.

We exemplify TOA using a coarse-grained self-replicator model (Molenaar et al., 2009; Giordano et al., 2016; Yegorov et al., 2018; Yabo et al., 2022) and demonstrate that TOA allows us to explain several known experimental phenomena that are difficult to investigate using existing static or dynamic constraint-based analysis methods. In particular, we demonstrate that TOA can explain the accumulation of storage compounds also in time-invariant environments–a counterintuitive fact that cannot be predicted using RBA and related methods. Likewise, we demonstrate that "luxury uptake" of nutrients, i.e., the fact that microorganisms may take up more of a limiting resource than strictly required for steady-state growth, can be explained by TOA and does not necessarily require competition within a microbial community. Furthermore, our analysis shows that organisms with internal temporal degrees of freedom, such as storage, can in most environments outperform organisms with a static intracellular composition. Finally, TOA shows that in constant (or slowly changing) environments, organisms adapted to better growth conditions ("optimists") outperfom organisms adapted to poorer growth conditions ("pessimists") when placed in the same environment.

The manuscript is organized as follows: Within **Sections 2.1** and **2.2** we introduce notation and define a general constraint-based framework to describe cellular metabolism and growth. This framework captures most current examples of dynamic constraint-based modeling, in particular dynamic FBA (Mahadevan et al., 2002), dynamic enzyme-cost FBA (Waldherr et al., 2015) and conditional FBA (Reimers et al., 2017). In **Section 2.3**, we formally introduce time-optimal adaptation (TOA) and discuss two relevant applications in **Section 2.4**: cell doubling in minimal time, as well as transition after a nutrient shift. The latter is formulated as a two-objective optimization problem (in the sense of Pareto) that considers a minimal time for the transition versus a total increase in biomass. In **Sections 2.5–2.7**, we discuss numerical aspects, variability analysis, and implementation, respectively.

Readers not interested in the mathematical details may skip most of *Materials and Methods* and focus on *Results*. In **Sections 3.1** and **3.2**, we describe the coarse-grained self-replicator model and its properties using RBA. In **Section 3.3**, we then apply TOA to describe cell doubling in minimal time in a constant environment. In **Section 3.4**, we discuss the role of "expectation", i.e., the consequences of being mis-adapted to a given environment. In **Section 3.5**, we apply TOA to simulate the metabolic response after a nutrient shift. In the final **Sections 4** and **5**, we discuss the biological implications of our results, and provide conclusions.

2 MATERIALS AND METHODS

2.1 Introduction and Notation

The dynamic simulation of metabolic networks by means of a fully parameterized ODE/DAE model is an ideal scenario that, in most cases, cannot be met due to the inherent incompleteness and uncertainty of the description and the involved parameters. Constraint-based modeling (Bordbar et al., 2014) has therefore become an important paradigm for the computational description of cellular metabolism and growth. The general idea can be framed as follows: instead of making use of a fully mechanistic description of biochemical dependencies by means of reaction rate equations, the system is characterized by a set of constraints/inclusions, typically defined by (in-)equalities that constrain the dynamics over a time interval $[t_0, t_{\mathrm{end}}]$ of interest.

Before capturing our approach in mathematical terms in **Section 2.2**, we introduce some notation, see also **Supplementary Appendix S1**. The function $\boldsymbol{y}\colon [t_0, t_{\mathrm{end}}] \to \mathbb{R}^{n_y}$ is used to describe the cellular dynamics by the total amounts $y(t)$ of intracellular compounds at time t (typically measured in number of molecules, mol), with $\dot{\boldsymbol{y}}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{y}(t)$ denoting the time-derivative. For simplicity, we focus on the dynamics of intracellular compounds only, extracellular compounds (e.g., nutrient or waste product concentrations) are not included in $\boldsymbol{y}$. Our framework, however, can be readily adapted to include the dynamics of extracellular compounds (see the **Supplementary Appendix S2.3** for details). Furthermore, our description is based on the assumption of a well-stirred metabolism, i.e., the spatial distribution of compounds is not considered.

We distinguish the total amounts of molecules $y(t)$ from their concentrations $\boldsymbol{c}(t)$, defined by

$$\boldsymbol{c}(t) = \frac{\boldsymbol{y}(t)}{bio(t)}, \tag{2.1}$$

where the term $bio(t) := \boldsymbol{w}^\top \cdot \boldsymbol{y}(t)$ denotes the total biomass of the system (measured in Gram cellular dry mass). The vector $\boldsymbol{w} \in \mathbb{R}^{n_y}$ denoting the molar masses of the entities of $\boldsymbol{y}$ (measured in gram cellular dry mass per mol).

The time evolution of the state vector $\boldsymbol{y}(t)$ can be described by means of ordinary differential equations.

$$\dot{\boldsymbol{y}}(t) = \mathbf{S} \cdot \boldsymbol{v}(t), \tag{2.2a}$$

where $\mathbf{S} \in \mathbb{R}^{n_y \times n_v}$ denotes the stoichiometric matrix and $\boldsymbol{v}\colon [t_0, t_{\text{end}}] \to \mathbb{R}^{n_v}$ the flux rates of the reactions. The flux rates $\boldsymbol{v}(t)$ may in general also depend on the environment the cells are exposed to. Typically, and specifically for large networks, the stoichiometric matrix $\mathbf{S} = \begin{pmatrix} \mathbf{S}_y \\ \mathbf{S}_x \end{pmatrix}$ is split up such that "fast" and "slow" intracellular compounds, usually metabolites resp. macromolecules, are described separately and (**2.2a**) is replaced by.

$$\begin{aligned} \dot{\boldsymbol{y}}(t) &= \mathbf{S}_y \cdot \boldsymbol{v}(t), \\ \mathbf{0} &= \mathbf{S}_x \cdot \boldsymbol{v}(t), \end{aligned} \tag{2.2b}$$

where the fast compounds, corresponding to the rows of $\mathbf{S}_x$, are subject to a quasi steady-state approximation (QSSA) (Segel and Slemrod, 1989). In this case, for simplicity of notation, the fast components will be removed from the vector $\boldsymbol{y}(t)$. We note that the splitting into "slow" and "fast" compounds is not a necessary step and its validity has to be verified in any particular application.

2.2 Constraint-Based Modeling

To capture the broad range of simulation frameworks that time-optimal adaptation is able to cover, we abstractly denote the constraints defining the specific constraint-based description of a cell *via*.

$$\text{for (almost) all } t\colon \ \left(\dot{\boldsymbol{y}}(t), \boldsymbol{y}(t), \boldsymbol{u}(t)\right) \in \mathcal{A}(t), \tag{2.3a}$$

where the set $\mathcal{A}(t) \subseteq \mathbb{R}^{n_y} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_u}$ is typically defined through (in-)equalities such as steady-state assumptions and/or positivity requirements. The particular form of the set $\mathcal{A}(t)$ usually depends on the chosen modeling framework and its granularity. For the present work, we model the influence of the external conditions via the explicit time-dependence of $\mathcal{A}(t)$. The vector-valued function $\boldsymbol{u} = \boldsymbol{u}(t)\colon [t_0, t_{\text{end}}] \to \mathbb{R}^{n_u}$ signifies the degrees-of-freedom of the cell, i.e., quantities that are not uniquely determined from the current state of the cell and its environment. In the context of control theory, $\boldsymbol{u}(t)$ defines the controls; on the biochemical level, it can for example stand for flux rates $\boldsymbol{v}(t)$ but also for parameters within the model.

The formal statement (**2.3a**) is usually not enough to sufficiently constrain the solutions, because the feasible region is too large to obtain biochemical insight. To get biochemically meaningful results, (**2.3a**) is therefore often accompanied by boundary conditions and an optimality principle, i.e., a global objective function f to be optimized:

$$\boldsymbol{\phi}_{\text{bndry}}\left(\boldsymbol{y}(t_0), \boldsymbol{y}(t_{\text{end}}), \boldsymbol{u}(t_0), \boldsymbol{u}(t_{\text{end}})\right) \le \mathbf{0}, \tag{2.3b}$$

$$\min_{\boldsymbol{y}(\cdot), \boldsymbol{u}(\cdot)} \ f(\boldsymbol{y}, \boldsymbol{u}) \tag{2.3c}$$

The boundary conditions **(2.3b)** are defined by means of inequalities to allow for more generality of this description. Usually, the boundary conditions will only contain initial values, provided by equality constraints, i.e., two inequalities. In some cases, optimality principles are already incorporated into the constraint set $\mathcal{A}(t)$, see the following examples.

In the context of optimal control-based methods with ODE/DAE constraints, the flux rates at any fixed point in time cannot (mathematically) be determined as they enter the problem as control variables (Gerdts, 2011). This is why (**2.3a**) technically can only be enforced for almost all times. Numerically or with respect to the biochemical reasoning, however, this has no further implications. In the following, we illustrate how (**2.3**) provides an abstract framework to describe established examples of constraint-based modeling.

EXAMPLE **2.1.** (Dynamic FBA, dFBA). Dynamic (or iterative) flux balance analysis (Varma and Palsson, 1994; Mahadevan et al., 2002), although one of the most commonly used dynamic frameworks within constraint-based modeling, is not consistently defined in the literature. Here, we refer to the formulation in (Höffner et al., 2016), see also (Höffner et al., 2012), for the characterization of dynamic FBA as a "dynamical system with a linear program embedded."

The control quantities $\boldsymbol{u}(t)$ can in this case be directly identified with the flux rates in the metabolic network model, i.e., $\boldsymbol{v}(t) = \boldsymbol{u}(t)$. The overall dynamics are governed by (**2.2a**), positivity requirements on $\boldsymbol{y}(t)$ and flux bounds $\boldsymbol{lb}, \boldsymbol{ub} \in \mathbb{R}^{n_v}$, which might be dependent on the time t:

$$\begin{aligned} \dot{\boldsymbol{y}}(t) &= \mathbf{S}_y \cdot \boldsymbol{v}(t), && \text{(dynamics, often just biomass)} \\ \mathbf{0} &= \mathbf{S}_x \cdot \boldsymbol{v}(t), && \text{(quasi steady - state)} \\ \mathbf{0} &\le \boldsymbol{y}(t), && \text{(positivity)} \\ \boldsymbol{lb}(t) &\le \boldsymbol{v}(t) \le \boldsymbol{ub}(t), && \text{(flux bounds, dependent on environment)} \end{aligned}$$

with given initial conditions

$$\boldsymbol{y}(t_0) = \boldsymbol{y}_0 \in \mathbb{R}^{n_y}.$$

The flux rates are determined through optimization of a linear functional (often the flux through the biomass reaction, assembled in a vector $\boldsymbol{w}_{\text{obj}} \in \mathbb{R}^{n_v}$)

$$\boldsymbol{v}(t) \in \operatorname*{argmin}_{\boldsymbol{v}} \boldsymbol{w}_{\text{obj}}^\top \cdot \boldsymbol{v}.$$

The quantities in (**2.3**) can be identified as:

$$\mathcal{A}(t) = \left\{ (\dot{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{u})\colon \begin{array}{rcl} \dot{\boldsymbol{y}} &=& \mathbf{S}_y \cdot \boldsymbol{u}, \\ \mathbf{0} &=& \mathbf{S}_x \cdot \boldsymbol{u}, \\ \mathbf{0} &\le& \boldsymbol{y}, \\ \boldsymbol{lb}(t) &\le& \boldsymbol{u} \le \boldsymbol{ub}(t), \\ \boldsymbol{u} &\in& \operatorname{argmin}_{\boldsymbol{v}} \boldsymbol{w}_{\text{obj}}^\top \cdot \boldsymbol{v} \end{array} \right\},$$

$$\boldsymbol{\phi}_{\text{bndry}} = \begin{pmatrix} \boldsymbol{y}(t_0) - \boldsymbol{y}_0 \\ \boldsymbol{y}_0 - \boldsymbol{y}(t_0) \end{pmatrix},$$

while typically no additional (global) objective function is present. Note that the defining condition on the fluxes $\boldsymbol{u} \in \underset{\boldsymbol{v}}{\operatorname{argmin}} \boldsymbol{w}_{\text{obj}}^{\top} \cdot \boldsymbol{v}$ is an inclusion, such that the solutions to dynamic FBA problems are, in general, not unique. To remedy this, flux variability analysis (FVA) (Mahadevan and Schilling, 2003) was introduced as a computational tool to explore the range of possible solutions of the static sub-problems.

EXAMPLE 2.2. (Dynamic enzyme-cost FBA, deFBA). Dynamic enzyme-cost FBA (Waldherr et al., 2015) is a dynamic extension of FBA that takes into account the temporal development and function of the enzymes. This is modeled by a system of linear inequalities

$$\mathbf{H}_{\boldsymbol{y}}(t) \cdot \boldsymbol{y}(t) + \mathbf{H}_{\boldsymbol{v}}(t) \cdot \boldsymbol{v}(t) \leq \boldsymbol{h}(t), \tag{2.4}$$

with

$$\mathbf{H}_{\boldsymbol{y}} \in \mathbb{R}^{n_{\boldsymbol{h}} \times n_{\boldsymbol{y}}},\ \mathbf{H}_{\boldsymbol{v}} \in \mathbb{R}^{n_{\boldsymbol{h}} \times n_{\boldsymbol{v}}}. \tag{2.5}$$

The model is usually formulated as an initial-value problem

$$\boldsymbol{y}(t_0) = \boldsymbol{y}_0 \in \mathbb{R}^{n_{\boldsymbol{y}}}.$$

Similar to FBA, deFBA assumes that a certain objective function is to be optimized. Since the framework entails a fully dynamic model over the whole time range of interest, the objective function contains "global" information, expressed as an optimal control objective of Boltza-type (Gerdts, 2011),

$$\min_{\boldsymbol{y},\boldsymbol{u}} \int_{t_0}^{t\text{end}} \boldsymbol{\phi}_{\boldsymbol{y}}^{\top}(t) \cdot \boldsymbol{y}(t) dt + \boldsymbol{\phi}_{\text{end}}^{\top} \cdot \boldsymbol{y}(t_{\text{end}}), \quad \text{with } \boldsymbol{\phi}_{\boldsymbol{y}}, \boldsymbol{\phi}_{\text{end}} \in \mathbb{R}^{n_{\boldsymbol{y}}}.$$

Like in dFBA, the control variables in deFBA can be identified with the flux rates and the description in terms of (**2.3**) is given by

$$\begin{aligned}
\boldsymbol{u}(t) &= \boldsymbol{v}(t), \\
\mathcal{A}(t) &= \left\{ \begin{array}{rrcl} & \dot{\boldsymbol{y}} & = & \mathbf{S}_{\boldsymbol{y}} \cdot \boldsymbol{u}, \\ & \mathbf{0} & = & \mathbf{S}_{\boldsymbol{x}} \cdot \boldsymbol{u}, \\ (\dot{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{u}): & \mathbf{0} & \leq & \boldsymbol{y} \\ & \boldsymbol{lb}(t) & \leq & \boldsymbol{u} \leq \boldsymbol{ub}(t), \\ & \boldsymbol{h}(t) & \geq & \mathbf{H}_{\boldsymbol{y}}(t) \cdot \boldsymbol{y} + \mathbf{H}_{\boldsymbol{u}}(t) \cdot \boldsymbol{u} \end{array} \right\} \\
\boldsymbol{\phi}_{\text{bndry}} &= \begin{pmatrix} \boldsymbol{y}(t_0) - \boldsymbol{y}_0 \\ \boldsymbol{y}_0 - \boldsymbol{y}(t_0) \end{pmatrix}, \\
f(\boldsymbol{y}(\cdot), \boldsymbol{u}(\cdot)) &= \int_{t_0}^{t_{\text{end}}} \boldsymbol{\phi}_{\boldsymbol{y}}^{\top}(t) \cdot \boldsymbol{y}(t) dt + \boldsymbol{\phi}_{\text{end}}^{\top} \cdot \boldsymbol{y}(t_{\text{end}}).
\end{aligned}$$

EXAMPLE 2.3. (Conditional FBA, cFBA). This framework (Rügen et al., 2015; Reimers et al., 2017) is again a dynamic extension of resource balance analysis (RBA) (Goelzer et al., 2011). Like in deFBA, enzymatic constraints (potentially alongside further constraints, e.g., on the cell's density) are included via (**2.4**). The boundary values in cFBA, however, are defined through a periodicity condition that accounts for the growth of the cell:

$$\boldsymbol{c}(t_0) = \frac{1}{bio(t_0)} \cdot \boldsymbol{y}(t_0) = \frac{1}{bio(t_{\text{end}})} \cdot \boldsymbol{y}(t_{\text{end}}) = \boldsymbol{c}(t_{\text{end}}). \tag{2.6}$$

Instead of using the biomass production on all time points, the objective in cFBA is the total growth of the cell until t_{end}. In terms of (**2.3**), cFBA can be summarized as

$$\begin{aligned}
\boldsymbol{u}(t) &= \begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_{2:} \end{pmatrix} = \begin{pmatrix} \alpha \\ \boldsymbol{v}(t) \end{pmatrix}, \quad \alpha \in \mathbb{R}, \\
\mathcal{A}(t) &= \left\{ \begin{array}{rrcl} & \dot{\boldsymbol{y}} & = & \mathbf{S}_{\boldsymbol{y}} \cdot \boldsymbol{u}_{2:}, \\ & \mathbf{0} & = & \mathbf{S}_{\boldsymbol{x}} \cdot \boldsymbol{u}_{2:}, \\ (\dot{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{u}): & \mathbf{0} & \leq & \boldsymbol{y}, \\ & \boldsymbol{lb}(t) & \leq & \boldsymbol{u}_{2:} \leq \boldsymbol{ub}(t), \\ & \boldsymbol{h}(t) & \geq & \mathbf{H}_{\boldsymbol{y}}(t) \cdot \boldsymbol{y} + \mathbf{H}_{\boldsymbol{v}}(t) \cdot \boldsymbol{u}_{2:} \end{array} \right\}, \\
\boldsymbol{\phi}_{\text{bndry}} &= \begin{pmatrix} \boldsymbol{u}_1 \cdot \boldsymbol{y}(t_0) - \boldsymbol{y}(t_{\text{end}}) \\ \boldsymbol{y}(t_{\text{end}}) - \boldsymbol{u}_1 \cdot \boldsymbol{y}(t_0) \end{pmatrix}, \\
f(\boldsymbol{y}(\cdot), \boldsymbol{u}(\cdot)) &= -\boldsymbol{u}_1,
\end{aligned}$$

where $\boldsymbol{u}_1$ refers to the first component of the vector $\boldsymbol{u}$ and $\boldsymbol{u}_{2:}$ to the vector of the remaining entries. If no constraints on the cell density are included in (**2.4**), the inequalities defining cFBA are often scale-invariant in the sense that for each solution $\boldsymbol{y}(t)$ and each number $\beta \geq 0$, the function $\beta \cdot \boldsymbol{y}(t)$ is also a solution. To exclude trivial solutions, the boundary conditions are therefore often extended such that the biomass at t_0 is equal to one. Note that cFBA is inherently nonlinear as the products $\boldsymbol{u}_1 \cdot \boldsymbol{y}$ in the boundary value constraints contribute quadratically in the unknowns $\boldsymbol{y}$ and $\boldsymbol{u}$. Like in RBA, the numerical solution of cFBA problems therefore comprises a series of linear programs that have to be solved after a discretization of the dynamics by means of, for example, a collocation scheme.

EXAMPLE 2.4. (Iterative RBA, (Liu, 2020), see also dynamic ME (Yang et al., 2019)). Just as dynamic FBA can be seen as a dynamic extension of classical FBA by iteratively applying the algorithm with constraints following the external conditions, resource balance analysis (RBA, see Goelzer et al. (2011)) can also be applied consecutively. In doing that, the limit case of infinitesimally short sub-intervals leads to a fully dynamic framework. Numerically, this limiting process is skipped and one only solves RBA problems on a series of short—but finite—time intervals. Note that, as cFBA, RBA uses periodicity conditions like (**2.6**) which implies that, in constant external conditions, only one RBA problem needs to be solved. The full solution in this case is given by an exponential curve for $\boldsymbol{y}(t)$. Note that there are fewer degrees-of-freedom for the cell when compared to deFBA or cFBA, as the fixed concentration values for the metabolites in the case of iterative RBA also block internal dynamics of the metabolic network.

In the notation of the constraint-based framework (**2.3**), iterative RBA can be written as

$$\begin{aligned}
\boldsymbol{u}(t) &= \begin{pmatrix} \alpha \\ \boldsymbol{v}_0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_{2:} \end{pmatrix}, \\
\mathcal{A}(t) &= \left\{ \begin{array}{rrcl} & \dot{\boldsymbol{y}} & = & \mathbf{S}_{\boldsymbol{y}} \cdot \boldsymbol{v} \\ & \mathbf{0} & = & \mathbf{S}_{\boldsymbol{x}} \cdot \boldsymbol{v} \\ & \mathbf{0} & \leq & \boldsymbol{y} \\ (\dot{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{u}): & \boldsymbol{lb}(t) & \leq & \boldsymbol{v} \leq \boldsymbol{ub}(t) \\ & \boldsymbol{h}(t) & \geq & \mathbf{H}_{\boldsymbol{y}}(t) \cdot \boldsymbol{y} + \mathbf{H}_{\boldsymbol{v}}(t) \cdot \boldsymbol{v} \\ & \boldsymbol{v} & = & \boldsymbol{u}_{2:} \cdot \exp(\lambda \cdot (t - t_0)) \\ & \lambda & = & \ln(\boldsymbol{u}_1) / (t_{\text{end}} - t_0) \end{array} \right\}, \\
\boldsymbol{\phi}_{\text{bndry}} &= \begin{pmatrix} \boldsymbol{y}(t_0) - \boldsymbol{y}_0 \\ \boldsymbol{y}_0 - \boldsymbol{y}(t_0) \end{pmatrix}, \\
f(\boldsymbol{y}(t), \boldsymbol{u}) &= -\boldsymbol{u}_1.
\end{aligned}$$

Note that the control variables $\boldsymbol{u}$ are not time-dependent, i.e., they enter the model as control parameters rather than functions that need to be optimized in the sense of optimal control.

2.3 Time-Optimal Adaptation: Definition and Forms

Previous frameworks for constraint-based optimization did not explicitly include the time interval as part of the optimization objective. In the following, we introduce Time-Optimal Adaptation (TOA) as a framework to analyze transition between different cellular states in the shortest possible time. TOA is motivated by the assumption that under certain environmental conditions, cells may have evolved to reach target amounts $\boldsymbol{y}^{\text{goal}}$ in the shortest possible time, starting from initial amounts $\boldsymbol{y}^{\text{init}}$. This transition might either take place in a variable environment, encoded by a time-dependent set $\mathcal{A}(t)$, or in a constant environment. Likewise, the target and initial amounts may either have to fulfill additional optimality criteria, or may correspond to pre-defined or experimentally measured states. Mathematically, we capture such a strategy in the following way.

Time-Optimal Adaptation

Given an initial/current amount of molecules $\boldsymbol{y}^{\text{init}} \in \mathbb{R}^{n_y}$ and a target amount $\boldsymbol{y}^{\text{goal}} \in \mathbb{R}^{n_y}$, the optimization objective is to transition from the former to the latter as quickly as possible.

$$\min_{\boldsymbol{y}(\cdot),\boldsymbol{u}(\cdot),T>t_0} T \tag{2.7a}$$

$$\text{s. t.} \quad \begin{aligned} \boldsymbol{y}(t_0) &= \boldsymbol{y}^{\text{init}}, \\ \boldsymbol{y}(T) &= \boldsymbol{y}^{\text{goal}} \end{aligned} \tag{2.7b}$$

$$\text{and } \left(\dot{\boldsymbol{y}}(t), \boldsymbol{y}(t), \boldsymbol{u}(t)\right) \in \mathcal{A}(t) \text{ for (almost) all } t \in [t_0, T], \textit{see } (2.3a) \tag{2.7c}$$

The constraints **(2.7c)** and **(2.7b)** can be framed within the abstract constraint-based framework (**2.3**) by including $\boldsymbol{y}^{\text{init}}$ and $\boldsymbol{y}^{\text{goal}}$ using

$$\boldsymbol{\phi}_{\text{bndry}}\left(\boldsymbol{y}(t_0), \boldsymbol{y}(t_{\text{end}}), \boldsymbol{u}(t_0), \boldsymbol{u}(t_{\text{end}})\right) = \begin{pmatrix} \boldsymbol{y}(t_0) - \boldsymbol{y}^{\text{init}} \\ \boldsymbol{y}^{\text{init}} - \boldsymbol{y}(t_0) \\ \boldsymbol{y}(t_{\text{end}}) - \boldsymbol{y}^{\text{goal}} \\ \boldsymbol{y}^{\text{goal}} - \boldsymbol{y}(t_{\text{end}}) \end{pmatrix},$$

whereas the global objective function, cf. f in (**2.3**), does not explicitly contain any of the variables $\boldsymbol{y}$ or $\boldsymbol{u}$. Instead, the general framework of constraint-based modeling (**2.3**) is extended through time-optimal adaptation by using the end point of the time interval of interest itself as the optimization objective function. In contrast to the frameworks with non-time-dependent objective function as defined in (**2.3c**), TOA provides solutions $(\boldsymbol{y}(t), \boldsymbol{u}(t))$ only on the time interval $[t_0, T]$ instead of (arbitrary) $[t_0, t_{\text{end}}]$.

Remark 2.5. Within this work, we assume that the target amounts $\boldsymbol{y}^{\text{goal}}$ are accessible. Specifically, we assume that a time $t_{\text{end}} \geq T$ exists such that all values within the optimization problem defining TOA are well-defined. We note that the accessibility of the target state is a classical problem in time-optimal control, and accessibility is a prerequisite for applying TOA. In practice, the target state will often be defined by means of an RBA solution and we conjecture that these target states will be accessible.

Remark 2.6. Within this work, we use the term "adaptation" in a control-theoretic sense. That is, the term refers to changes in the intracellular amounts or concentrations in response to the environmental conditions, respecting the given constraints. In an evolutionary context, such changes are typically considered as "acclimation".

Remark 2.7. We do not require the constraint set $\mathcal{A}(t)$ in (**2.3a**) to have any specific form. This means that time-optimal adaptation can be defined irrespective of the concrete modeling paradigm underneath the simulation. Practically, even discrete time/state systems fit well within TOA. To be concise, however, we concentrate in the following on frameworks closely related to deFBA and cFBA. In Example 2.8, we therefore introduce TOA also in a simplified setting that directly builds upon d(e)FBA, cf. (Waldherr et al., 2015; Höffner et al., 2016). From the viewpoint of the general framework (**2.3**), this is a special case of deFBA with a modified objective function.

EXAMPLE **2.8.** (TOA as an extension of deFBA). Assume that there is no distinction between "fast" and "slow" components within the metabolic network. In this case, the dynamics of its molecular amounts can be described purely by ordinary differential equations $\dot{\boldsymbol{y}}(t) = \mathbf{S} \cdot \boldsymbol{v}(t)$. As for classical flux balance analysis, the fluxes are constrained by upper and lower bounds $\boldsymbol{lb}$, $\boldsymbol{ub}$ that might depend on the possibly changing environment, i.e., $\boldsymbol{ub}(t) \leq \boldsymbol{v}(t) \leq \boldsymbol{ub}(t)$. If $\boldsymbol{y}$ contains compounds with enzymatic function, the flux rates (or weighted sums thereof) may additionally be constrained by (weighted sums of) components of $\boldsymbol{y}$. Such bounds can be collected into a single set of linear inequalities by introducing suitable matrices/vectors $\mathbf{H}_{\mathbf{y}}(t)$, $\mathbf{H}_{\mathbf{v}}(t)$, $\boldsymbol{h}(t)$, i.e.,

$$\mathbf{H}_y(t) \cdot \boldsymbol{y}(t) + \mathbf{H}_v(t) \cdot \boldsymbol{v}(t) \leq \boldsymbol{h}(t),$$

see (Waldherr et al., 2015, **Section 2.3**) for a detailed description. To account for $\boldsymbol{y}$ being total amounts, $\boldsymbol{y}$ is constrained to positive values, i.e., $\boldsymbol{y}(t) \geq \mathbf{0}$. As outlined above, TOA requires fixed initial and terminal values for the molecular amounts, mathematically captured by

$$\boldsymbol{y}(t_0) = \boldsymbol{y}^{\text{init}}, \qquad \boldsymbol{y}(T) = \boldsymbol{y}^{\text{goal}}.$$

In summary, TOA can be aggregated in this simplified case to the following constrained optimization problem

$$\min_{\boldsymbol{y}(\cdot),\boldsymbol{v}(\cdot),T>t_0} T$$

$$\text{s.t.} \quad \left.\begin{aligned} \dot{\boldsymbol{y}}(t) &= \mathbf{S} \cdot \boldsymbol{v}(t) \\ \boldsymbol{lb}(t) &\leq \boldsymbol{v}(t) \leq \boldsymbol{ub}(t) \\ \boldsymbol{h}(t) &\geq \mathbf{H}_y(t) \cdot \boldsymbol{y}(t) + \mathbf{H}_v(t) \cdot \boldsymbol{v}(t) \\ \boldsymbol{y}(t) &\geq \mathbf{0} \end{aligned}\right\}$$

$$\left.\begin{aligned} \dot{\boldsymbol{y}}(t_0) &= \boldsymbol{y}^{\text{init}} \\ \dot{\boldsymbol{y}}(T) &= \boldsymbol{y}^{\text{goal}} \end{aligned}\right\}$$

The notation "$\min_{\boldsymbol{y}(\cdot),\boldsymbol{v}(\cdot),T>t_0}$" can be understood in the sense of optimal control, i.e., one is searching for the optimal objective value among all (differentiable) functions $\boldsymbol{y}(t)$, $t \in [t_0, T]$, and

(measurable) functions $\boldsymbol{v}(t)$, $t \in [t_0, T]$. The framework identifies possible time courses for the fluxes $\boldsymbol{v}(t)$ and amounts $\boldsymbol{y}(t)$ such that (i) stoichiometry, (ii) flux bounds, and (iii) enzyme activities are included in the model and such that the transition from one given amount to another is as fast as possible.

2.4 Applications and Case Studies

Next we introduce two particularly relevant applications of TOA.

APPLICATION 2.1. (Cell Doubling). A first natural application of TOA is cell doubling, where the objective is to double all cellular components in minimal time, such that

$$\boldsymbol{y}^{\text{goal}} = 2 \cdot \boldsymbol{y}^{\text{init}}.$$

The resulting trajectory thus can be interpreted as one cell cycle. Neither the initial, nor the target amount have to be optimal with regard to other objectives. Within the TOA framework, cell doubling can be considered either in a constant environment, or with time-dependent external conditions. We note that applications of constraint-based optimization of metabolism typically do not distinguish between solutions for a single cell vs. solutions for a homogeneous population of cells. Similarly, the time courses for cell doubling predicted by TOA can either be interpreted for a single cell or a homogeneous, synchronized population of cells. If cells are not synchronized, that is each cell within the population is at a different time point with respect to its cell cycle, we have to average over the population or, equivalently, over a full cell cycle, to obtain *in silico* measurements of a population.

APPLICATION 2.2. (Transitions after a nutrient shift). A second important application of TOA is to consider a sudden change in the external conditions, i.e., from a given constant nutrient availability for $t < 0$ to a different one for $t \geq 0$. In this scenario, TOA can be utilized to predict the transition of the intracellular amounts $\boldsymbol{y}^{\text{init}}$ to new target amounts $\boldsymbol{y}^{\text{goal}}$. The new target amounts might either be optimal with respect to the new environmental conditions (as defined by RBA), or be provided otherwise (for example by experimental observations). In both scenarios, the target amounts are typically defined in terms of concentrations instead of absolute amounts. Hence, we must also formulate the boundary conditions in terms of $\boldsymbol{c}$(t),

$$\begin{aligned} \boldsymbol{y}(t_0) &= \boldsymbol{y}^{\text{init}}, \\ \boldsymbol{c}(T) &= \frac{1}{bio(T)} \cdot \boldsymbol{y}(T) = \frac{1}{\boldsymbol{w}^\top \cdot \boldsymbol{y}^{\text{goal}}} \cdot \boldsymbol{y}^{\text{goal}} = \boldsymbol{c}^{\text{goal}}. \end{aligned} \tag{2.8}$$

As shown in **Supplementary Appendix S3**, it is possible to rearrange conditions (2.8) such that a linear equality system in the unknowns ($\mathbf{y}(t_0)$, $\boldsymbol{y}(T)$) is obtained. Therefore, the concentration-based definition has no immediate drawbacks regarding the numerical solution.

We must further consider that an as-quick-as-possible transition from one intracellular concentration to another does not incorporate the overall (i.e., biomass-) growth of the cell and thus might not represent an evolutionarily plausible strategy. Rather, the transition to new external conditions involves a balance between fast transition to a (better adapted) novel state and the requirement to increase (or not decrease) the total biomass of the cell. To obtain a general framework, we therefore propose a two-objective optimization problem:

$$\begin{aligned} \min_{\boldsymbol{y}(\cdot),\boldsymbol{u}(\cdot),T>0} \begin{pmatrix} T \\ -\alpha \end{pmatrix} \\ s.t.\ \ \boldsymbol{y}(0) &= \boldsymbol{y}^{\text{init}} \\ \boldsymbol{y}(T) &= \alpha \cdot \boldsymbol{y}^{\text{goal}} \\ \text{and } \left(\dot{\boldsymbol{y}}(t), \boldsymbol{y}(t), \boldsymbol{u}(t)\right) &\in \mathcal{A}(t) \text{ for (almost) all } t \in [0, T], \end{aligned} \tag{2.9}$$

where $\boldsymbol{y}^{\text{init}}$ denotes a normalized vector of intracellular amounts which describe the cells for the environment $t < 0$. The "normalization" here can, for example, be understood as $\boldsymbol{w}^\top \cdot \boldsymbol{y}^{\text{init}} = 1$. Accordingly, $\boldsymbol{y}^{\text{goal}}$ denotes a normalized vector for the environmental conditions after the nutrient shift. "Minimality" in (**2.9**) is to be understood in the sense of Pareto: a triple ($\boldsymbol{y}(t)$, $\boldsymbol{u}(t)$, T) is optimal if T cannot be decreased without decreasing α such that $\boldsymbol{y}(T) = \alpha \cdot \boldsymbol{y}^{\text{goal}}$, and vice-versa if α cannot be increased without also increasing the end time T. The set of all optimal solutions of (**2.9**) describes the different compromises between fast adaptation and continued growth.

Remark 2.9. Note that the boundary conditions **(2.8)** do not entail any direct condition concerning $bio(t_{\text{end}}) = \boldsymbol{w}^\top \cdot \boldsymbol{y}(t_{\text{end}})$. If the metabolic network allows for a quick degradation of compounds, it might be optimal (in the sense of TOA) to shrink (in terms of absolute biomass) before actually adapting to the new concentrations, or even to completely disintegrate all metabolic compounds to zero. Such a behavior would be in line with the description of time-optimal time courses as induced by **(2.8)**. To remedy this, a linear inequality

$$\begin{aligned} bio(t_{\text{end}}) &\geq bio(t_0) \\ \text{or} \quad \boldsymbol{w}^\top \cdot \boldsymbol{y}(t_{\text{end}}) &\geq \boldsymbol{w}^\top \cdot \boldsymbol{y}(t_0) \end{aligned}$$

can be added, illustrating again the power of constraint-based modeling. Whenever necessary, this was done for the in silico experiments in **Section 3**.

2.5 Numerical Solution

The optimization problem (**2.7**) of TOA contains a general condition on the dynamics of $\boldsymbol{y}$ in the form of **(2.7c)**. To design an algorithm able to cope with this generality, we assume that a numerical method is available that can simulate this dynamic behavior subject to boundary conditions on a given fixed time interval $[t_1, t_2] \subseteq [t_0, T]$ and/or to determine whether such a solution exists. Provided this condition (and tacitly assuming that the relevant feasible end time points T lie in a connected set), the minimal value for T can be found using any one-dimensional root finding algorithm. For its simplicity and guaranteed convergence, we propose to use the

following bisection method for the determination of T in TOA:

Input: • an initial interval $[t_{\min}, t_{\max}] \subseteq [t_0, t_{\text{end}}]$ for the optimal end time point T
• a tolerance level tol
While $|t_{\max} - t_{\min}| > tol$:
Set $t_{\text{new}} = (t_{\min} + t_{\max})/2$
If (2.7b), (2.7c) is soluble for $T = t_{\text{new}}$:
Set $t_{\max} = t_{\text{new}}$
Else:
Set $t_{\min} = t_{\text{new}}$
Output:
$T = t_{\text{new}}$ as approximation to the optimal end time point

For the initial time interval $[t_{\min}, t_{\max}]$, one needs to assume that **(2.7b)** and **(2.7c)** define an infeasible problem on $[t_0, t_{\min}]$, while the corresponding problem on $[t_0, t_{\max}]$ is feasible. The quick convergence of the bisection method entails that an already very good initial guess is not crucial for an efficient implementation, as long as the simulation task is not too computationally expensive.

If there is legitimate doubt about the result, the algorithm can be re-started with another initial interval or one can change to a more fine-grained sampling for the evaluation of feasible and infeasible points. The numerical results in **Section 3** were preceded by an exhaustive scan of end time points, which indicated that the set of feasible end time points do indeed form a single interval (i.e., a connected set) in all shown examples.

Remark 2.10. The bisection method was chosen here for several reasons over more "classical" methods in time-optimal control: firstly, the "simplicity" aspect of the bisection method does not only refer to it being easily applicable for various extensions of the framework (like time- and/or state-discrete systems, or a framework that incorporates heterogeneity within a community or in space) but also to the implementation. Many existing toolboxes include interfaces to (MI-)LP solvers. Algorithms for dynamic simulations are moreover often highly optimized, such that checking for feasibility over a given time range can be more efficient than implementing a new interface to an optimal control library.

Secondly, the inherently linear structure of problems like deFBA should be preserved. For time-dependent constraint sets $\mathcal{A}(t)$ this is only possible if the time variable t is treated as the independent variable in the optimal control algorithm. In existing optimal control libraries like BOCOP (Bonnans et al., 2017), time-optimal control problems are often transferred to optimal control problems on a unit interval by introducing an artificial independent variable. If the time-dependency of some of the constraints is non-linear, this translates to the optimization problems that need to be solved within the optimal control routine.

We note, however, that in the non-linear case the application of "classical algorithms" for time-optimal control problems like shooting-methods, or those based on the Pontryagin principle might generally outperform the bisection approach taken here.

Remark 2.11. For the solution of the Pareto problem (**2.9**) it is not necessary to implement algorithms for maximizing α, i.e., optimizing the second objective. Instead, one can continue using the algorithm for time-optimal adaptation while simultaneously fixing feasible values of α. With respect to the definition of Pareto-optimality, this means that for any feasible value of one objective, the other one is optimized, corresponding to the so-called ϵ-constraint method in multi-objective optimization, cf. (Ehrgott, 2000).

2.6 Time-Optimal Adaptation Variability Analysis

Minimizing T need not suffice to uniquely determine the time courses in $\boldsymbol{y}$. If this is the case, the variability over time can be captured by enumerating possible time series once the optimal end time point was found. We will refer to this procedure as TOA-Variability Analysis (TOA-VA). In contrast to static flux variability analysis (FVA), there are several ways to define what such an enumeration means. One way would be to determine the maximum and minimum possible value for all components of $\boldsymbol{y}$ and separately at each time point. This, however, would not only lead to time-consuming computations, but would also be difficult to interpret: a numerical solution that is constructed via putting together maxima or minima $\boldsymbol{y}(\tilde{t})$ for all time instances $\tilde{t}$ does not have to fulfill the dynamics defined by the original model. Here, we understand TOA-VA as the minimization and maximization of the integral over all components of $\boldsymbol{y}$, i.e., for all $i = 1, 2, \ldots, n_{\mathbf{y}}$:

$$\min_{\boldsymbol{u}(\cdot),\boldsymbol{y}(\cdot)} \pm \int_{t_0}^{T} y_i(t)\mathrm{d}t \tag{2.10}$$

subject to the dynamic and/or boundary constraints in the original problem. Note that the overall time courses might still not be uniquely defined from (**2.10**).

To explore the variability of the time courses for the concentrations $\boldsymbol{c}$(t), we use the following variant of TOA-VA, which is called relative TOA-VA:

(i) Compute the optimal end time point T of time-optimal adaptation.
(ii) For all $i = 1, 2, \ldots, n_{\mathbf{y}}$: Use TOA-VA as in (**2.10**) to obtain a minimal value $I_{\min,\mathrm{i}}$ and a maximal value $I_{\max,\mathrm{i}}$ for the integral of y_{i} over $[t_0, T]$.
(iii) Calculate for all $i = 1, 2, \ldots, n_{\mathbf{y}}$ the (maximal and minimal) concentrations $c_{\mathrm{i}}(t)$ as given by **(2.1)** where $\boldsymbol{y}(t)$ is calculated from

$$\max_{\boldsymbol{y}(\cdot),\boldsymbol{u}(\cdot)} \int_{t_0}^{T} bio(t)\mathrm{d}t \quad \text{s.t.} \int_{t_0}^{T} y_i(t)\mathrm{d}t \le I_{\min,i},$$
$$\min_{\boldsymbol{y}(\cdot),\boldsymbol{u}(\cdot)} \int_{t_0}^{T} bio(t)\mathrm{d}t \quad \text{s.t.} \int_{t_0}^{T} y_i(t)\mathrm{d}t \ge I_{\max,i}$$

(again subject to the original constraints of the problem).

There is still no guarantee that the solutions to these problems are unique. However, since the concentrations are defined as the ratio of total amounts to the biomass, the above definition is reasonable as one is maximized whilst minimizing the other.

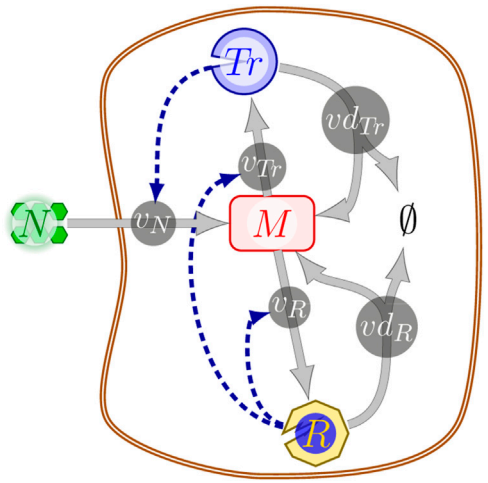


FIGURE 1 | A schematic illustration of the coarse-grained self-replicator model; solid lines represent biochemical reactions between the nodes (biochemical compounds), dashed dark-blue lines indicate that a reaction is catalyzed by the respective compound. Abbreviations: *N*, external nutrient; *M*, metabolic precursor/storage; *Tr*, transporter; *R*: ribosome; v_N, nutrient uptake reaction; v_R, ribosome production reaction; vd_R, ribosome degradation reaction; v_{Tr}, transporter production reactions; vd_{Tr}, transporter degradation reaction.

Note, that this definition implies that the weighted sum of all (maximal or minimal) concentrations no longer needs to add to the total biomass.

2.7 Implementation

The calculations for all experiments in **Section 3** were done in Python 3.8.1 on a laptop computer. Scripts that reproduce the numerical experiments below are available on GitHub, https://github.com/MarkusKoebis/StaticTOA_py The numerical solutions were determined from a complete parameterization (using the trapezoidal rule) of the compounds and fluxes over the entire time range of interest using $n = 100$ steps on an equidistant grid. This leads to a sparse LP problem which was solved using `gurobipy` on Gurobi 9.0.1 solver (Gurobi, 2021) with standard settings (concerning problem formulation and tolerances). Most experiments were repeated (for verification) with tight error tolerances without notable differences. For time-optimal adaptation, no objective vector for the LPs is necessary, so we used the null vector $\mathbf{0}$. For (relative) TOA-VA, the integrals in the objective or constraints were approximated using the same time grid and also the trapezoidal rule.

3 RESULTS

3.1 A Coarse-Grained Self-Replicator Model

We illustrate TOA by means of a coarse-grained self-replicator model (Molenaar et al., 2009; Giordano et al., 2016; Yabo et al., 2022). The model, cf. **Figure 1**, consists of three compounds: *M* (intracellular metabolic precursor), *Tr* (transporter), and *R* (ribosome), as well as five biochemical reactions, and the external nutrient *N*. The uptake of the external nutrient *N* is catalyzed by the transporter *Tr* and depends on the availability of *N* via a Michaelis-Menten rate equation. Depending on the application, the concentration of the external nutrient *N* may either be constant or vary over time. The synthesis of the catalytic macromolecules *Tr* and *R* is limited by the ribosome amount. Within the model, macromolecules can be disassembled into the precursor *M*. For energetic consistency, however, disassembly results in fewer precursor molecules than required for synthesis, reflecting the energy expenditure of protein synthesis and thereby avoiding futile cycles. We note that within the model, no compound is subject to the quasi steady-state assumption, and the metabolic precursor *M* can accumulate over time. Hence *M* also serves as a storage compound. All constraints of the model can be formulated in terms of linear inequalities. A detailed definition is provided in **Supplementary Appendix S2.1**.

3.2 Constant Environments and RBA

Before the dynamic behavior of the model is studied by means of TOA, we summarize the steady-state properties of the model in a constant environment using Resource Balance Analysis (RBA). RBA provides a method to calculate the steady-state amounts of the cell that maximize the growth rate under constant external conditions, i.e., for a constant external nutrient concentration. In the following, extracellular nutrient is measured relative to the Michaelis constant K_M of the uptake reaction, with N/K_M as a dimensionless parameter.

Figure 2A shows the maximal growth rate λ as a function of the relative nutrient availability. The growth rate follows a Monod equation with a maximum $\lambda^{\max} \approx 0.435\,\mathrm{h}^{-1}$ and an effective (dimensionless) affinity constant $K_A \approx 0.347$, corresponding to the value of the relative nutrient availability N/K_M at which the cell grows at half the maximal growth rate $\lambda^{\max}$.

Figure 2B shows the total amounts of the three intracellular components *M*, *Tr*, and *R* as a function of the (relative) nutrient availability. The amounts were scaled such that the total biomass always equals one unit (e.g., 1 g cellular dry mass). As expected, when maximizing the growth rate, the level of the precursor/storage component *M* is always zero. This reflects the fact that the precursor *M* has no catalytic activity, and any non-zero amount of *M* would consume resources that otherwise could be allocated to transport or protein translation.

The amounts of the other intracellular components *Tr* and *R* follow the well-known growth laws of microbiology (Scott and Hwa, 2011). The concentrations are a function of the growth rate, and hence the external nutrient availability, the well-known linear relationship is shown in **Supplementary Appendix S4**. With increasing nutrient availability, the relative amount of transporter decreases, whereas the relative amount of ribosome increases.

3.3 TOA in Constant Environments

Our first case study using TOA is to consider the doubling of a microbial cell in minimal time. We assume that the self-replicator model in **Figure 1** has pre-described initial amounts $\boldsymbol{y}(t_0) = \boldsymbol{y}_0$ which simultaneously identify the pre-defined initial state $\boldsymbol{y}^{\mathrm{init}}$. The objective is to double all cellular components as fast as possible, cf. Application 2.1. The environment is assumed to be

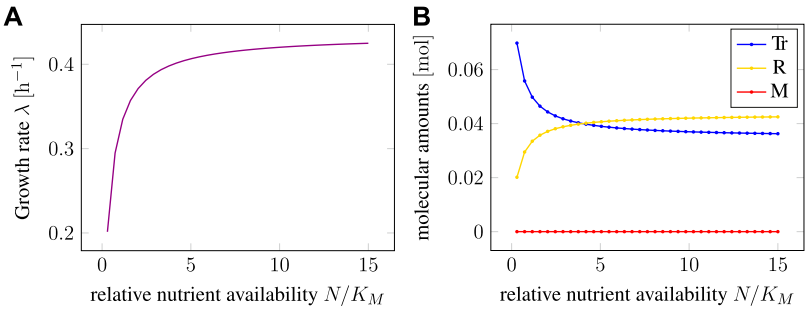


FIGURE 2 | (A) Maximal growth rate λ as a function of the (relative) extracellular nutrient availability as predicted by RBA. **(B)** Cellular amounts of intracellular compounds as functions of relative nutrient availability. Extracellular nutrient is measured relative to the Michaelis constant K_M of the uptake reaction.

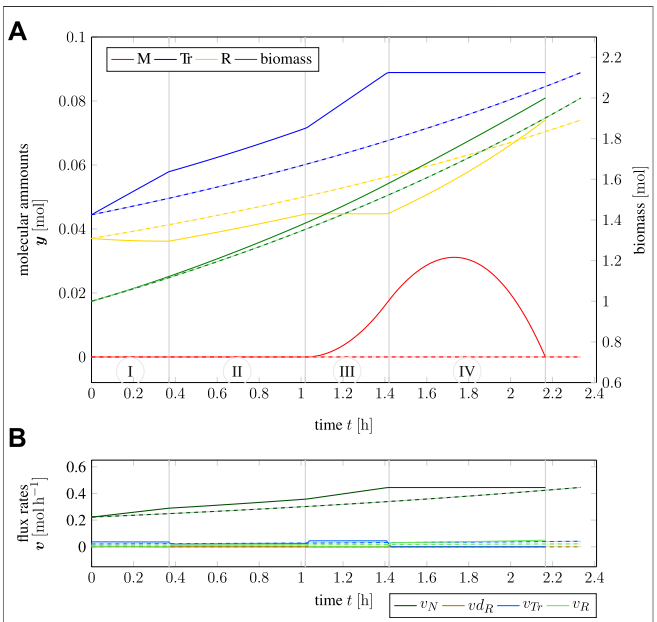


FIGURE 3 | Cell cycle of an "optimistic" cell; **(A)** amounts and biomass as a function of time, **(B)** flux rates as a function of time; solid lines indicate the solution of TOA, dashed lines indicate iterative RBA (exponential growth) with the same "optimistic" initial values.

constant with a relative (external) nutrient availability $N/K_M = 1$. The initial (and final) amounts are not assumed to be optimal for the given environment. Instead, $\boldsymbol{y}(t_0)$ is obtained by solving an RBA problem corresponding to $N/K_M = 2.0$. In other words, the cell is assumed to be adapted to a higher nutrient level than is present in the current environment. In the following, we will refer to such cells as "optimists".

Figure 3 shows the time course of intracellular components for one cell doubling. The predicted time-optimal amounts of metabolic compounds are shown as solid lines (red, blue, and yellow), the total biomass is shown in green. The dashed lines correspond to a solution obtained by iterative RBA (cf. **Example 2.4**), which corresponds to exponential growth of all cellular components with no further internal degrees of freedom. **Figure 3B** shows the respective flux rates over the simulated time range. Solid lines again indicate the solution of TOA, while dashed lines (exponential curves) correspond to the solution found with iterative RBA.

Using TOA, the time for one cellular doubling is $T = 2.17$ h. In contrast, the solution based on iterative RBA results in a slightly longer doubling time of $T = 2.34$ h, showing that internal degrees of freedom shorten the calculated division time. The time course of $\boldsymbol{y}(t)$ over one cell doubling can be subdivided into four time intervals (marked as I-IV in **Figure 3A**). At the beginning (marked as interval "I"), cell growth is limited by the lack of transporter *Tr* due to the "optimistic" initial configuration of the cell. Hence, ribosome *R* is actively disassembled into precursor *M* to increase the synthesis of *Tr*. In interval "II", the cell is perfectly adapted to the given nutrient environment and grows exponentially, before the re-adaptation to the target composition $\boldsymbol{y}^{\text{goal}} = 2\,\boldsymbol{y}^{\text{init}}$ begins in interval "III". Within interval "III", the cell still has an overabundance of *Tr*, which allows it to accumulate the precursor *M*. In the final interval "IV", synthesis of transporter *Tr* ceases and all resources are devoted to the synthesis of the ribosome *R*, until the target amounts $\boldsymbol{y}^{\text{goal}}$ are reached.

The biological plausibility of these time courses is discussed in **Section 4**. Here we only summarize the following results: Given the initial amounts $\boldsymbol{y}^{\text{init}}$, cell doubling using TOA in time-invariant environments gives rise to complex intracellular dynamics different from solutions obtained by iterative RBA. Importantly, these solutions involve a transient accumulation of the precursor *M* as a storage compound–a phenomenon not observed with iterative RBA. The minimal division time predicted by time-optimal adaptation is shorter than division times obtained by iterative RBA.

So far, we considered a particular initial amount $\boldsymbol{y}^{\text{init}}$ such that the cell was adapted to a higher nutrient availability than actually present in the environment ("optimist"). To obtain a broader view, we evaluated cell doubling using TOA in different time-invariant environments with initial (and final) amounts adapted to different external nutrient availability. The results are shown in **Figure 4**. Solid lines correspond to intracellular amounts using TOA, dashed lines correspond to a solution obtained with iterative RBA (exponential growth without internal degrees of

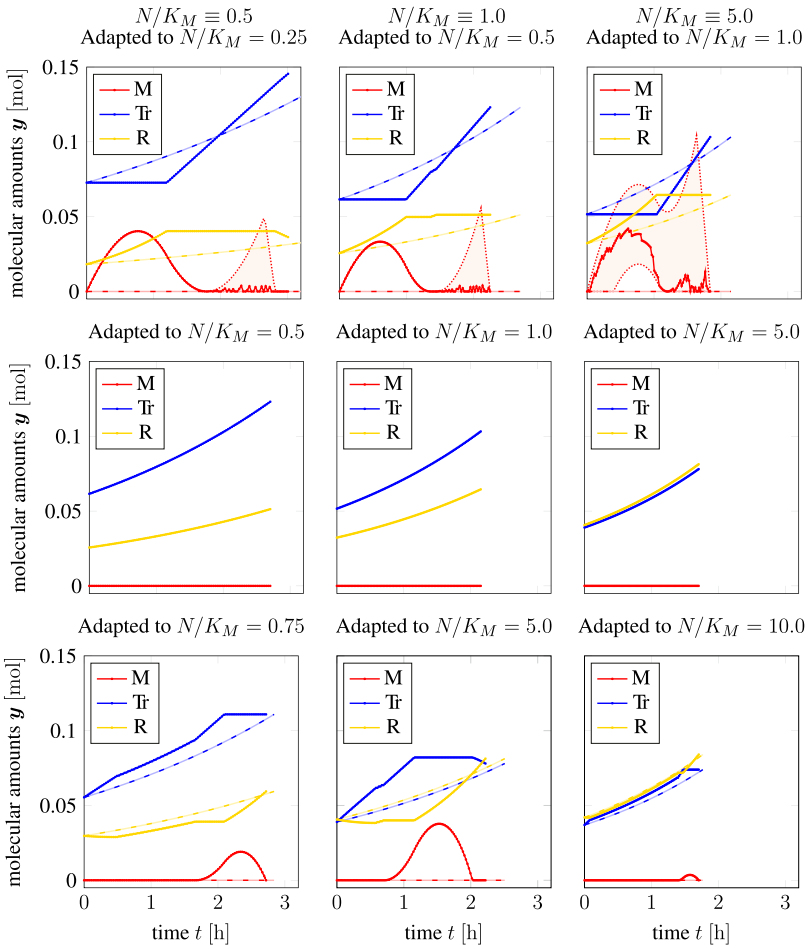


FIGURE 4 | Time course solutions of time-optimal adaptation and a cell doubling experiment under different constant external nutrient conditions; solid lines: TOA, shaded areas: TOA-VA, dashed lines: iterative RBA (simulated until cell doubling was achieved); upper row: pessimistically adapted, middle row: perfectly adapted (recovery of iterative RBA), bottom row: optimistically adapted for constant relative nutrient availability of $N/K_M = 0.5$ (left column), $N/K_M = 1.0$ (middle column), and $N/K_M = 5.0$ (right column).

freedom). Shaded areas correspond to variability in the sense of TOA-VA (cf. **Section 2.6**), i.e., possible solutions that equally satisfy all constraints and the optimality criterion. In this case, the solid lines display a "nominal" solution, i.e., one that was provided by the algorithm before an additional variability analysis (we note that since the numerical solution is based on a feasibility problem, the LP solver has no incentive to favor a smooth solution to any other).

Columns in **Figure 4** correspond to different relative nutrient availability levels: the first column to a nutrient availability $N(t)/K_M \equiv 0.5$; the second column to $N(t)/K_M \equiv 1.0$, and the third to $N(t)/K_M \equiv 5.0$. The rows in **Figure 4** correspond to different "expectations" of the cells, that is, which external nutrient availability the initial (and final) amounts are adapted to. Specifically, the first row corresponds to "pessimists". That is, cells adapted to a nutrient availability below the one present in the environment, while retaining the objective to double all cellular components in minimal time. The second row corresponds to cells perfectly adapted to the environmental nutrient availability. The final row corresponds to "optimists", i.e., cells adapted to a higher nutrient availability than present in the environment.

The latter scenario corresponds to the example already shown in **Figure 3**. We again observe an initial increase in the transporter synthesis, followed by a delayed onset of ribosome synthesis. Importantly, in each case, we can see a transient accumulation of storage $M(t)$ that is absent in solutions obtained by iterative RBA. In the case of perfectly adapted cells (middle row), solutions obtained by TOA are equivalent to solutions obtained by iterative RBA. For "pessimistic" cells (top row), we again observe complex time courses. In particular, cells adapted to lower nutrient levels than present in the environment exhibit an overabundance of transporter. Hence, we observe an initial rapid uptake of nutrient and transient accumulation of the precursor M. In the initial interval, resources are primarily allocated to the synthesis of ribosomes. Only in the later

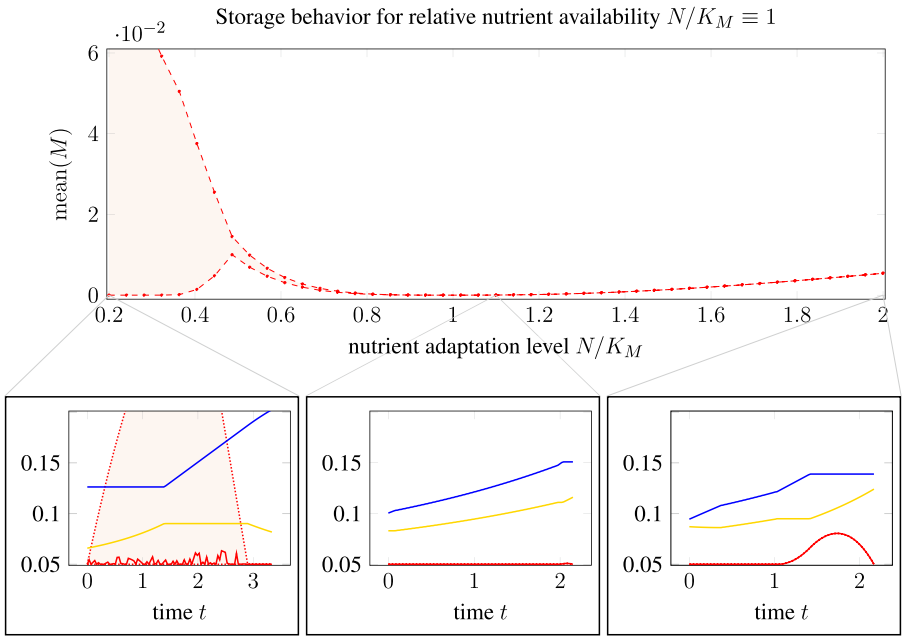


FIGURE 5 | Influence of optimistic and pessimistic goal states in cell doubling: Main plot: mean relative storage accumulation, see **(3.1)**, as a function of nutrient adaptation level. Bottom row: Three selected time courses, cf. **Figure 4**, for nutrient adaptation levels N/K_M of 0.2, 1.1, and 2.0. For $N/K_M < 1$, the quantity *mean*(M) is no longer unique such that a shaded area indicates the possible range, as TOA-VA also predicts a range of possible solutions (shaded area in the bottom left plot).

interval, the transporter is synthesized to the required amounts (even at the expense of ribosomes that may be disassembled into precursors). The transient accumulation of precursor M exhibits considerable variability and the solutions of TOA are no longer unique.

A detailed discussion about the biological plausibility of these time courses is again relegated to **Section 4**. Here we only note that, despite the simplicity of the model, the solutions exhibit a wide variety of qualitatively different complex temporal behaviors, including the transient accumulation of the precursor M.

3.4 The Role of Expectation: Optimists vs. Pessimists

We further investigate two key observations obtained in the previous experiments: the transient accumulation of precursor M as a storage compound, as well as the impact of the initial cellular state on the predicted doubling time.

Firstly, **Figure 5** shows the average storage concentration predicted for a population of cells adapted to different nutrient availabilities ($N/K_M \in (0.2, 2.0)$, x-axis) in an environment with an actual relative nutrient availability $N/K_M \equiv 1.0$. To calculate the average storage concentration predicted by TOA for a population of cells, we assume that the (*in silico*) measurements are taken from a heterogeneous population of unsynchronized cells that are (equidistributed) at various stages of their cell cycle. To take this non-uniform age distribution into account, the population average was computed, cf. (Powell, 1956), as

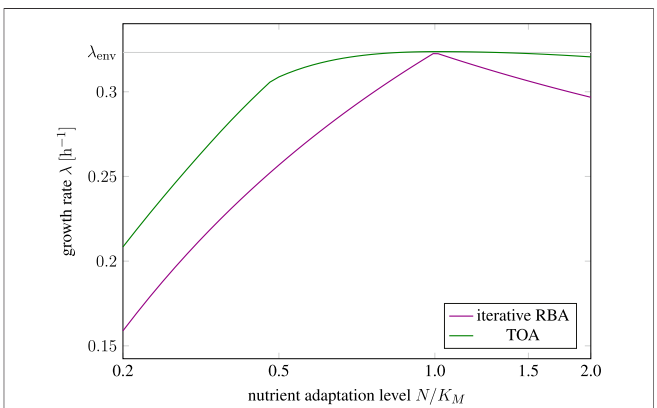


FIGURE 6 | Growth rate of differently adapted cells as predicted by cell-doubling experiments using TOA and iterative RBA in an environment with relative nutrient availability $N/K_M = 1.0$; $\lambda_{env} \approx 0.32$ h^{-1} denotes the maximal growth rate as predicted by RBA.

$$\text{mean}(M) = \frac{1}{T} \dot{c} \int_0^T \frac{M(t)}{\boldsymbol{w}^\top \cdot \boldsymbol{y}(t)} \mathrm{d}t, \tag{3.1}$$

where $\boldsymbol{y}(t)$ is a solution obtained by relative TOA-VA, cf. **Section 2.6**.

As shown in **Figure 5**, we observe (the possibility of) a nonzero average storage concentration for all cellular states that are not perfectly adapted to the respective environment. For optimistic cells adapted to a higher nutrient availability than present in the environment, the average storage concentration increases slightly

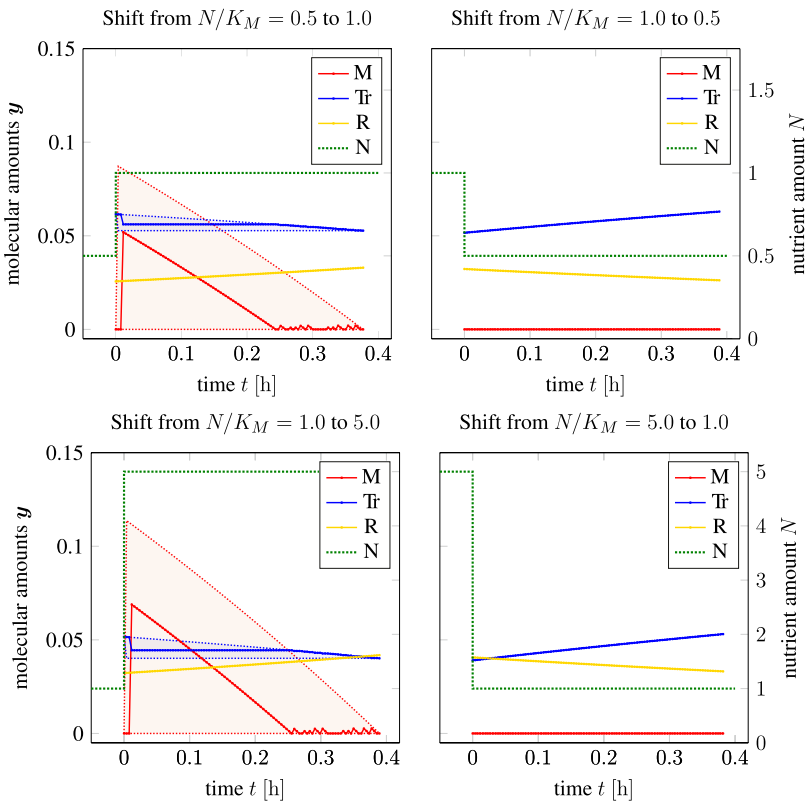


FIGURE 7 | Adaptation to a single nutrient jump (shown as a dashed green line), left column: adaptations from poorer to richer medium, right column: adaptation to scarcer environment; shaded areas: solutions in the sense of TOA-VA. We note that for $t < 0$, TOA makes no assumptions about $\boldsymbol{y}(t)$.

with the distance to the perfectly adapted state. The effect is more pronounced for pessimistic cells adapted to a lower nutrient availability than present in the environment. In this case, the solutions of TOA are not unique and the range of average storage is indicated as a shaded area. For "pessimist" cells, the large average storage is due to a high abundance of transporter molecules, which implies that uptake and accumulation of precursor is not restricted.

Secondly, **Figure 6** shows the predicted growth rate for cells adapted to a different relative nutrient availability ($N/K_M \in (0.2, 2.0)$) than present in the environment ($N/K_M \equiv 1.0$). The straight line indicates the growth rate of cells that are perfectly adapted, resulting in a maximal growth rate $\lambda = \lambda_{\text{env}} \approx 0.32\ \text{h}^{-1}$. The maximal growth rates for cells adapted to a different environment (misadaptation) are shown as a solid green line for solutions obtained with TOA and as a purple line for solutions obtained with iterative RBA.

We observe that misadaptation always results in a reduced growth rate, as compared to a perfectly adapted cell. However, solutions obtained by TOA always outperform solutions obtained by iterative RBA, demonstrating that internal degrees of freedom and transient accumulation of storage shorten the predicted doubling time. Furthermore, the decrease in growth rate is more pronounced for "pessimistic" adaptation, that is, for cells that are adapted to a lower nutrient level than present in the environment. In contrast, "optimistic" adaptation, that is, cells are adapted to a higher levels than present in the environment, together with TOA results in growth rates close to perfectly adapted cells–indicating that "optimistic" adaptation carries a lower evolutionary cost than "pessimistic" adaptation.

3.5 Time-Optimal Adaptation at a Nutrient Shift

As our second application, we consider a nutrient shift, i.e., a sudden change in the external conditions from a given constant nutrient availability for $t < 0$ to a different one for $t \geq 0$. TOA is utilized to predict the time-optimal transition of a cell perfectly adapted to the initial state at $t < 0$ to a state perfectly adapted to maximize growth in the new environment for $t \geq 0$. As noted in **Section 2.4**, the target state for the new environment is typically defined in terms of concentrations rather than absolute amounts, because it is unknown whether or how much the cells are able to grow during adaptation.

Figure 7 shows the resulting time courses for the coarse-grained self-replicator model used in the previous sections. Shown are time-optimal shifts from a low nutrient availability to a higher nutrient availability (left column in **Figure 7**), as well as time-optimal shifts from a high nutrient availability to a lower nutrient availability (right column in **Figure 7**). Non-unique solutions are again displayed as shaded areas indicating the maximum and minimum range in which solutions can be

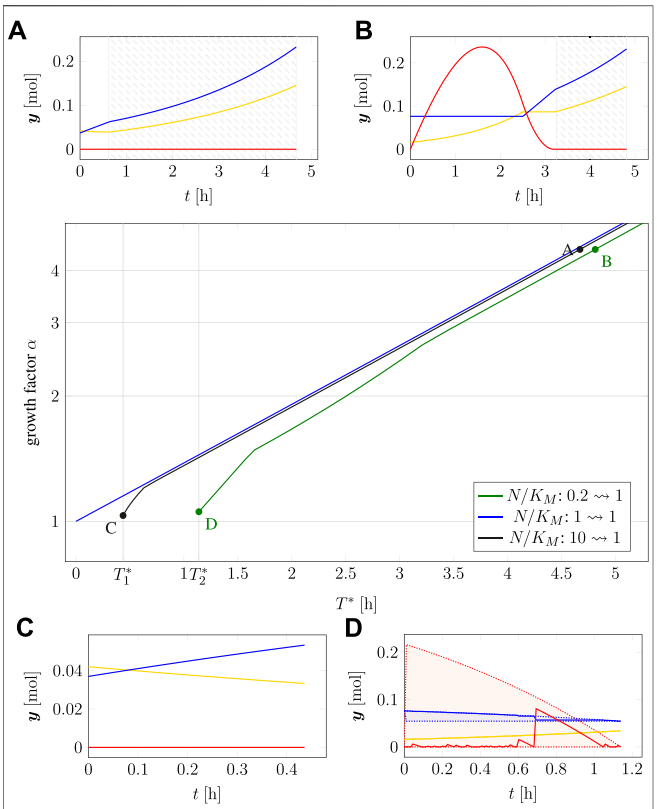


FIGURE 8 | Two-objective optimization of adaption time T^* and total biomass growth factor α for time-optimal adaptation at a nutrient shift. Main plot: Pareto fronts for three different initial adaptations (measured in N/K_M) of 0.2, 1.0 and 10. Subplots **(A–D)**: Time courses at different points on the Pareto fronts, cf. **Figure 7**. The shaded areas in subplots A and B indicate time intervals where the cell is perfectly adapted (exponential growth).

found (TOA-FVA, see **Section 2.6**). We observe that the time-optimal transition from lower to higher nutrient availability again entails a transient accumulation of storage.

As detailed in **Section 2.4**, time-optimal adaptation alone is not sufficient as an evolutionary principle to explain cellular adaptation after a nutrient shift. Rather, we consider a two-objective optimization (in the sense of Pareto) with the conflicting objectives of a fastest possible adaptation to the new state vs. a maximal increase in total cellular biomass.

Figure 8 (main panel) shows the resulting Pareto fronts for different transitions in terms of the minimal time T^* for adaptation vs. the maximal increase in cellular biomass given by the factor α, cf. (**2.9**). Panels A–D in **Figure 8** show selected time courses of intracellular amounts at different positions of the Pareto front. In the subplots A and B, the shaded areas indicate that the cell is perfectly adapted to the environment in the sense of RBA, i.e., from the start of the shaded areas, the cell exhibits balanced exponential growth at the maximal growth rate and no further internal dynamics take place. The absence of internal dynamics explains that, for larger values of α or T^*, the lines in the main plot become asymptotically parallel.

In the absence of a nutrient shift (i.e., the transition N/K_M: 1 → 1, blue line in **Figure 8**), the minimal time for adaptation is $T^* = 0$ with a growth factor $\alpha = 1$, in this case the relationship between transition times $T^* > 0$ and increase in biomass is consistent with exponential growth (note the logarithmic scale on the y-axis).

For a nutrient shift from high to low nutrient availability (N/K_M: 10 → 1, black line) the minimal transition time is $T^* = T_1^* \approx 0.44$ h. **Figures 8A–C** show two representative transitions on the Pareto front with panel A corresponding to a scenario that prioritizes an increase in biomass (factor α) over the transition time T^*, and panel C corresponding to a scenario that prioritizes a minimal transition time over the accumulation of biomass.

For a nutrient shift from low to higher nutrient availability (N/K_M: 0.2 → 1, green line) the minimal transition time is $T^* = T_2^* \approx 1.14$ h. **Figures 8B,D** show two representative transitions on the Pareto front with panel B corresponding to a scenario that prioritizes an increase in biomass (factor α) over the transition time T^*, and panel D corresponding to a scenario that prioritizes a minimal transition time over the accumulation of biomass. In either case, the optimal transition involves a transient accumulation of the storage compound M.

Consistent with results in the previous section, **Figure 8** also shows that "optimistic" adaptation carries a lower evolutionary cost than "pessimistic" adaptation. A cell adapted to high nutrient availability exhibits only a slightly reduced biomass increase when transitioning into a low nutrient environment, as compared to a cell already adapted to this environment. In contrast, a cell adapted to a lower nutrient environment exhibits a more pronounced reduction in accumulated biomass when transitioning into higher nutrient availability, as compared to either a cell that is already adapted to the higher nutrient availability, or likewise as compared to a cell that was previously adapted to even higher nutrient availability.

4 DISCUSSION

In this work we introduced TOA, a novel approach to simulate and predict time-optimal adaptation of microbial metabolism and growth. While time-optimal modeling has been considered before, see, among others, (Klipp et al., 2002) (temporal gene expression), (Pavlov and Ehrenberg, 2013) (fast proteome adaptation to environmental change), (Waldherr et al., 2015) (maximize survival time under nutrient depletion), (Basan et al., 2020) (minimization of lag/response-time), and (Djema et al., 2020) (bio-reactor applications), our work builds upon the recent advances in dynamic constraint-based modeling, such as dFBA, deFBA and cFBA, cf. **Section 2.2**. TOA is versatile and extends most approaches currently employed in constraint-based modeling of microbial metabolism and growth.

In particular, while the analysis of balanced steady-state growth dominates current experimental and computational studies, in most natural environments microbes have to continuously adapt to perturbations and changes in nutrient availability. TOA allows us to study such transitions in the context of established constraint-based models of microbial metabolism. Similar to other constraint-based methods, the solutions obtained from TOA are not based on mechanistic

understanding of the regulatory system that governs the respective transition, but are derived from the assumption that, under certain conditions, a time-optimal transition may be evolutionary beneficial. We emphasize that an application of TOA does not necessarily imply that a time-optimal transition is the only or most important evolutionary objective. Rather, and again similar to other optimality-based methods, the solutions of TOA provide a computational "gold standard", (Giordano et al., 2016), to which experimentally observed behavior can be compared.

Within this work, we exemplified the use of TOA by considering two prototypical applications: the doubling of a cell in a constant environment (cf. **Application 2.1**), as well as the time-optimal adaptation to a nutrient shift (cf. **Application 2.2**). Following previous works (Molenaar et al., 2009; Giordano et al., 2016; Yabo et al., 2022), the application of TOA was illustrated using a coarse-grained self-replicator model. The results illustrate the utility of TOA to generate and explore biological hypotheses.

The premise underlying the *in silico* experiments of our first application, cell doubling in a constant environment, was that microbial cells are not necessarily precisely adapted to the given environment, but may nonetheless have evolved a regulatory scheme that allows them to double their intracellular composition in minimal time. Based on this premise, the application of TOA gives rise to several predictions, we observe 1) complex intracellular dynamics different from solutions obtained by iterative RBA, 2) that transient accumulation of storage compounds reduces the predicted doubling time, and 3) that (mis-)adaptation to a higher nutrient availability than actually present in the environment carries a lower evolutionary cost than (mis-)adaptation to a lower nutrient availability.

Due to the simplicity of the coarse-grained model, we do not necessarily expect the specific time courses obtained for the model to be exact predictions of biological reality. In particular, we acknowledge that the coarse-grained model lacks further intracellular constraints that affect progress through the cell cycle (for example, checkpoints and a detailed representation of DNA replication and segregation) that also impact metabolic processes. Nonetheless, we are confident that the results reveal several insights that reflect biological reality. Specifically, the role of storage compounds in cellular metabolism is difficult to explore using existing constraint-based models. Here, the application of TOA demonstrates that, beyond the role of storage in diurnal oscillations, cf. (Rügen et al., 2015; Reimers et al., 2017) and as a safeguard for periods of nutrient scarcity, storage may play an important role even under constant environmental conditions. As shown with TOA, intracellular dynamics and transient accumulation of nutrients may contribute to a reduction of doubling time. Indeed, and different from typical steady-state solutions of current constraint-based methods, cells do exhibit coordinated metabolic dynamics over a cell cycle (Papagiannakis et al., 2017).

The application of TOA was further exemplified by simulations of time-optimal cellular adaptation to a nutrient shift. Similar to the results obtained for constant environments, TOA demonstrates that transient accumulation of storage can reduce the time required for adaptation–a finding supported by experimental evidence that storage compounds, such as glycogen, indeed provide short-term benefits in changing environments (Sekar et al., 2020).

In particular, the rapid uptake and storage of nutrients following an upshift in nutrient supply (as shown in **Figure 7**, left column) is reminiscent of "luxury uptake" or "over-compensation". The latter phenomenon is well known (Powell et al., 2009) and occurs when cells are starved and re-exposed to a limiting nutrient, such as phosphate. "Luxury uptake" and "over-compensation" after starvation can be exploited, for example, for nutrient removal from wastewater (Powell et al., 2009). Our analysis shows that such "over-compensation" or "overshoot" phenomena are readily explained using principles of (optimal) cellular resource allocation, and do not necessarily require explanations that invoke competition between individuals to rationalize rapid nutrient uptake after starvation.

We conjecture that, while the specific trajectories of the cellular response to environmental shifts might be different under specific conditions, for example, due to additional constraints not present in the model, many of the principles revealed by TOA remain valid in more elaborate models of cellular growth transitions–and thereby provide an important reference to identify optimal vs. suboptimal behavior. Indeed, it was previously shown that growth transition kinetics of *E. coli* are indeed suboptimal under the studied conditions (Erickson et al., 2017)–a finding that could only be obtained by comparison to an optimal reference solution. As shown in this work, TOA can also be readily incorporated into a multi-objective framework (in the sense of Pareto) that allows us to incorporate additional objectives.

Finally, the results of TOA demonstrate that the costs of mis-adaptation to an environment are not symmetric, neither for cell doubling in a constant environment (**Figure 6**), nor for adaptation after a nutrient shift (**Figure 8**). In either case, a cell that is adapted to a higher level of (extracellular) nutrient than available in the environment ("optimist") has only a minor disadvantage compared to an already perfectly adapted cell. Vice versa, however, cells that are adapted to a lower level of (extracellular) nutrient than available in the environment ("pessimist") have a pronounced disadvantage compared to a perfectly adapted cell. This asymmetry indicates that adaptation to a low nutrient environment is only advantageous if the low nutrient state persists for an extended period of time. This asymmetry is supported by experimental evidence. For example, it has been suggested that some microorganisms, such as *Lactococcus lactis*, preserve a large overcapacity of ribosomes and glycolytic enzymes to be ready to rapidly respond and grow when conditions improve (Goel et al., 2015), and thereby implement an "optimistic strategy".

5 CONCLUSIONS AND OUTLOOK

Constraint-based optimization plays an important role to elucidate and eventually predict cellular behavior. As an extension of previous modeling frameworks, we introduced time-optimal adaptation. TOA is motivated by the assumption that under certain conditions it is evolutionary favorable to adapt to a new cellular state in minimal time. In its general form, TOA can be applied in a very broad sense and thereby extends most of the existing constraint-based modeling frameworks.

As shown in this work, TOA allowed us to obtain insight into several biological phenomena, such as the accumulation of storage in constant environments and "overshoot" accumulation of nutrients after starvation, which cannot be readily explained using existing methods–thereby demonstrating the utility of TOA for future analysis.

While the examples discussed within this work focused on constant environments and simple nutrient shifts, TOA can also be applied in time-dependent environments and can be readily extended to include further constraints. Likewise, as shown in this work, TOA can be included within multi-objective optimization in the sense of Pareto.

Possible further extensions include "t-max adaptation", i.e., to maximize, for example, survival time under nutrient starvation, as well as more general constraints on the target state (for example, to attain a minimal amount of a specific intermediate in minimal time, while the amounts other cellular components are not specified).

We are therefore confident that TOA and its possible extensions are a valuable contribution in the context of constraint-based modeling with manifold applications beyond the examples discussed in this work.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

All authors contributed equally to the conceptualization of the method, the writing and editing of the manuscript. MK conducted the numerical experiments. RS provided interpretations and discussion.

FUNDING

The work of MK was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship of the author at the Norwegian University of Science and Technology. The work of RS is funded by the grant STE 2062/2-1 of the German Research Foundation (DFG). We acknowledge support by the German Research Foundation (DFG) and the Open Access Publication Fund of Humboldt-Universität zu Berlin.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.866676/full#supplementary-material

REFERENCES

Basan, M., Honda, T., Christodoulou, D., Hörl, M., Chang, Y.-F., Leoncini, E., et al. (2020). A Universal Trade-Off between Growth and Lag in Fluctuating Environments. *Nature* 584, 470–474. doi:10.1038/s41586-020-2505-4

Bonnans, F., Martinon, P., and Grélard, V. (2017). *BOCOP: An Open Source Toolbox for Optimal Control*.

Bordbar, A., Monk, J. M., King, Z. A., and Palsson, B. O. (2014). Constraint-based Models Predict Metabolic and Associated Cellular Functions. *Nat. Rev. Genet.* 15, 107–120. doi:10.1038/nrg3643

Djema, W., Bernard, O., and Giraldi, L. (2020). Separating Two Species of Microalgae in Photobioreactors in Minimal Time. *J. Process Control* 87, 120–129. doi:10.1016/j.jprocont.2020.01.003

Ehrgott, M. (2000). *Multicriteria Optimization*. Springer Berlin Heidelberg. doi:10.1007/978-3-662-22199-0

Erickson, D. W., Schink, S. J., Patsalo, V., Williamson, J. R., Gerland, U., and Hwa, T. (2017). A Global Resource Allocation Strategy Governs Growth Transition Kinetics of *Escherichia coli*. *Nature* 551, 119–123. doi:10.1038/nature24299

Gerdts, M. (2011). *Optimal Control of ODEs and DAEs*. Berlin, Boston: De Gruyter.

Giordano, N., Mairet, F., Gouzé, J.-L., Geiselmann, J., and de Jong, H. (2016). Dynamical Allocation of Cellular Resources as an Optimal Control Problem: Novel Insights into Microbial Growth Strategies. *PLoS Comput. Biol.* 12 (28), e1004802. doi:10.1371/journal.pcbi.1004802

Goel, A., Eckhardt, T. H., Puri, P., de Jong, A., Branco dos Santos, F., Giera, M., et al. (2015). Protein Costs Do Not Explain Evolution of Metabolic Strategies and Regulation of Ribosomal Content: Does Protein Investment Explain an Anaerobic Bacterial Crabtree Effect? *Mol. Microbiol.* 97, 77–92. doi:10.1111/mmi.13012

Goelzer, A., Fromion, V., and Scorletti, G. (2011). Cell Design in Bacteria as a Convex Optimization Problem. *Automatica* 47, 1210–1218. doi:10.1016/j.automatica.2011.02.038

Gurobi (2021). *Gurobi Optimizer Reference Manual. Tech. Rep.* Beaverton, OR: Gurobi Optimization LLC.

Hermes, H., and Lasalle, J. P. (1969). *Functional Analysis and Time Optimal Control*. New York: Academic Press.

Höffner, K., Harwood, S. M., and Barton, P. I. (2012). A Reliable Simulator for Dynamic Flux Balance Analysis. *Biotechnol. Bioeng.* 110, 792–802. doi:10.1002/bit.24748

Höffner, K., Khan, K. A., and Barton, P. I. (2016). Generalized Derivatives of Dynamic Systems with a Linear Program Embedded. *Automatica* 63, 198–208. doi:10.1016/j.automatica.2015.10.026

Jeanne, G., Goelzer, A., Tebbani, S., Dumur, D., and Fromion, V. (2018). Dynamical Resource Allocation Models for Bioreactor Optimization. *IFAC-PapersOnLine* 51, 20–23. doi:10.1016/j.ifacol.2018.09.020

Klipp, E., Heinrich, R., and Holzhütter, H.-G. (2002). Prediction of Temporal Gene Expression. *Eur. J. Biochem.* 269, 5406–5413. doi:10.1046/j.1432-1033.2002.03223.x

Lerman, J. A., Hyduke, D. R., Latif, H., Portnoy, V. A., Lewis, N. E., Orth, J. D., et al. (2012). In Silico method for Modelling Metabolism and Gene Product Expression at Genome Scale. *Nat. Commun.* 3. doi:10.1038/ncomms1928

Liu, L., and Bockmayr, A. (2020). Regulatory Dynamic Enzyme-Cost Flux Balance Analysis: A Unifying Framework for Constraint-Based Modeling. *J. Theor. Biol.* 501, 110317. doi:10.1016/j.jtbi.2020.110317

Liu, L. (2020). *Unifying Metabolic Networks, Regulatory Constraints, and Resource Allocation*. Berlin, Germany: Freie Universität Berlin. Ph.D. thesis. doi:10.17169/REFUBIUM-27061

Mahadevan, R., Edwards, J. S., and Doyle, F. J. (2002). Dynamic Flux Balance Analysis of Diauxic Growth in *Escherichia coli*. *Biophysical J.* 83, 1331–1340. doi:10.1016/s0006-3495(02)73903-9

Mahadevan, R., and Schilling, C. H. (2003). The Effects of Alternate Optimal Solutions in Constraint-Based Genome-Scale Metabolic Models. *Metab. Eng.* 5, 264–276. doi:10.1016/j.ymben.2003.09.002

Molenaar, D., van Berlo, R., de Ridder, D., and Teusink, B. (2009). Shifts in Growth Strategies Reflect Tradeoffs in Cellular Economics. *Mol. Syst. Biol.* 5, 323. doi:10.1038/msb.2009.82

Orth, J. D., Thiele, I., and Palsson, B. Ø. (2010). What Is Flux Balance Analysis? *Nat. Biotechnol.* 28, 245–248. doi:10.1038/nbt.1614

Papagiannakis, A., Niebel, B., Wit, E. C., and Heinemann, M. (2017). Autonomous Metabolic Oscillations Robustly Gate the Early and Late Cell Cycle. *Mol. Cell* 65, 285–295. doi:10.1016/j.molcel.2016.11.018

Pavlov, M. Y., and Ehrenberg, M. (2013). Optimal Control of Gene Expression for Fast Proteome Adaptation to Environmental Change. *Proc. Natl. Acad. Sci. U.S.A.* 110, 20527–20532. doi:10.1073/pnas.1309356110

Powell, E. O. (1956). Growth Rate and Generation Time of Bacteria, with Special Reference to Continuous Culture. *J. General Microbiol.* 15, 492–511. doi:10.1099/00221287-15-3-492

Powell, N., Shilton, A., Chisti, Y., and Pratt, S. (2009). Towards a Luxury Uptake Process via Microalgae - Defining the Polyphosphate Dynamics. *Water Res.* 43, 4207–4213. doi:10.1016/j.watres.2009.06.011

Reimers, A. M., Knoop, H., Bockmayr, A., and Steuer, R. (2017). Cellular Trade-Offs and Optimal Resource Allocation during Cyanobacterial Diurnal Growth. *Proc. Natl. Acad. Sci. U. S. A.* 114, E6457–E6465. doi:10.1073/pnas.1617508114

Rügen, M., Bockmayr, A., and Steuer, R. (2015). Elucidating Temporal Resource Allocation and Diurnal Dynamics in Phototrophic Metabolism Using Conditional FBA. *Sci. Rep.* 5 (16), 15247. doi:10.1038/srep15247

Scott, M., and Hwa, T. (2011). Bacterial Growth Laws and Their Applications. *Curr. Opin. Biotechnol.* 22, 559–565. doi:10.1016/j.copbio.2011.04.014

Segel, L. A., and Slemrod, M. (1989). The Quasi-Steady-State Assumption: A Case Study in Perturbation. *SIAM Rev.* 31, 446–477. doi:10.1137/1031091

Sekar, K., Linker, S. M., Nguyen, J., Grünhagen, A., Stocker, R., and Sauer, U. (2020). Bacterial Glycogen Provides Short-Term Benefits in Changing Environments. *Appl. Environ. Microbiol.* 86. doi:10.1128/aem.00049-20

Varma, A., and Palsson, B. O. (1994). Stoichiometric Flux Balance Models Quantitatively Predict Growth and Metabolic By-Product Secretion in Wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.* 60, 3724–3731. doi:10.1128/aem.60.10.3724-3731.1994

Waldherr, S., Oyarzún, D. A., and Bockmayr, A. (2015). Dynamic Optimization of Metabolic Networks Coupled with Gene Expression. *J. Theor. Biol.* 365, 469–485. doi:10.1016/j.jtbi.2014.10.035

Yabo, A. G., Caillau, J.-B., Gouzé, J.-L., de Jong, H., and Mairet, F. (2022). Dynamical Analysis and Optimization of a Generalized Resource Allocation Model of Microbial Growth. *SIAM J. Appl. Dyn. Syst.* 21, 137–165. doi:10.1137/21m141097x

Yang, L., Ebrahim, A., Lloyd, C. J., Saunders, M. A., and Palsson, B. O. (2019). DynamicME: Dynamic Simulation and Refinement of Integrated Models of Metabolism and Protein Expression. *BMC Syst. Biol.* 13, 2. doi:10.1186/s12918-018-0675-6

Yegorov, I., Mairet, F., and Gouzé, J.-L. (2018). Optimal Feedback Strategies for Bacterial Growth with Degradation, Recycling, and Effect of Temperature. *Optim. Control Appl. Meth* 39, 1084–1109. doi:10.1002/oca.2398

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Köbis, Bockmayr and Steuer. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 13 September 2022
doi: 10.3389/fmolb.2022.878148

Dynamic Modelling of DNA Repair Pathway at the Molecular Level: A New Perspective

Paola Lecca[1]* *and Adaoha E. C. Ihekwaba-Ndibe*[2]*

[1]Faculty of Computer Science, Free University of Bozen-Bolzano, Bolzano, Italy, [2]Faculty of Health and Life Sciences, Coventry University, Coventry, United Kingdom

OPEN ACCESS

Edited by:
Ernesto Perez-Rueda,
Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Arvind Panday,
Harvard Medical School, United States
Valentina Tozzini,
Consiglio Nazionale delle Ricerche, Italy

***Correspondence:**
Paola Lecca
Paola.Lecca@unibz.it
Adaoha E. C. Ihekwaba-Ndibe
Adaoha.Ihekwaba@coventry.ac.uk

Specialty section:
This article was submitted to Biological Modeling and Simulation, a section of the journal Frontiers in Molecular Biosciences

Received: 17 February 2022
Accepted: 22 June 2022
Published: 13 September 2022

Citation:
Lecca P and Ihekwaba-Ndibe AEC (2022) Dynamic Modelling of DNA Repair Pathway at the Molecular Level: A New Perspective. Front. Mol. Biosci. 9:878148. doi: 10.3389/fmolb.2022.878148

DNA is the genetic repository for all living organisms, and it is subject to constant changes caused by chemical and physical factors. Any change, if not repaired, erodes the genetic information and causes mutations and diseases. To ensure overall survival, robust DNA repair mechanisms and damage-bypass mechanisms have evolved to ensure that the DNA is constantly protected against potentially deleterious damage while maintaining its integrity. Not surprisingly, defects in DNA repair genes affect metabolic processes, and this can be seen in some types of cancer, where DNA repair pathways are disrupted and deregulated, resulting in genome instability. Mathematically modelling the complex network of genes and processes that make up the DNA repair network will not only provide insight into how cells recognise and react to mutations, but it may also reveal whether or not genes involved in the repair process can be controlled. Due to the complexity of this network and the need for a mathematical model and software platform to simulate different investigation scenarios, there must be an automatic way to convert this network into a mathematical model. In this paper, we present a topological analysis of one of the networks in DNA repair, specifically homologous recombination repair (HR). We propose a method for the automatic construction of a system of rate equations to describe network dynamics and present results of a numerical simulation of the model and model sensitivity analysis to the parameters. In the past, dynamic modelling and sensitivity analysis have been used to study the evolution of tumours in response to drugs in cancer medicine. However, automatic generation of a mathematical model and the study of its sensitivity to parameter have not been applied to research on the DNA repair network so far. Therefore, we present this application as an approach for medical research against cancer, since it could give insight into a possible approach with which central nodes of the networks and repair genes could be identified and controlled with the ultimate goal of aiding cancer therapy to fight the onset of cancer and its progression.

Keywords: DNA damage, DNA repair genes, dynamical networks, ODE models, parametric sensitivity analysis, centrality measure analysis

1 INTRODUCTION

DNA molecules packaged in our chromosomes carry our genetic blueprint, and their preservation is essential for the coordination of cellular function and organization of life (Branzei and Foiani, 2008). As DNA is the repository of our genetic information, we would expect its structure to be highly stable. This is not the case for the DNA (Reed and Waters, 2005). Damage to DNA is a constant threat (Lindahl, 1993; Alberts, 2015). The DNA molecule is intrinsically reactive, as it is very susceptible to chemical and physical factors, which can lead to DNA lesions, such as base loss, base modification, and double-strand DNA breaks (Hoeijmakers, 2009; Çağlayan and Wilson, 2015; Ross and Truant, 2016; Yadav et al., 2020). Physiological conditions such as oxygen-rich, aqueous, or pH 7.4 (Lindahl, 1993) as well as chemical events such as hydrolysis and exposure to reactive oxygen species (ROS) or other reactive metabolites can damage DNA. Exogenous chemicals or endogenous metabolic processes trigger chemical reactions. Although exogenous stressors can be extremely powerful, endogenous threats are constant and unabating. It is estimated that a single cell experiences up to 10^5 spontaneous or induced DNA lesions per day (Lindahl, 1993; Bont, 2004; Kovalchuk, 2016; Chatterjee and Walker, 2017).

DNA damage has far-reaching consequences, such as preventing RNA polymerase from transcribing the correct messenger RNA sequence to produce the correct protein. In the longer term, cellular malfunctions such as cancer initiation, inborn defects, and ageing that result after damaged DNA replicates are examples of unpredictable long-term consequences of DNA damage; as base misincorporation causes mutations which alter the genetic code (Reed and Waters, 2005). Therefore, a coordinated response to DNA damage is necessary in order to ensure cellular viability and prevent diseases. Cells, fortunately, possess a robust system of mechanisms that function together to reduce the adverse consequences of DNA damage and ensure that their genetic information is faithfully replicated, thus maintaining the integrity of their genome (Ganai and Johansson, 2016). This coordinated effort, known as DNA damage response (DDR) operates by sensing and signalling the genotoxic events, and the damage is then resolved either by DNA repair machineries, or cell death if DNA cannot be repaired. DNA repair functions as part of the DNA damage response (DDR) (Liu, 2001; Chatterjee and Walker, 2017; Reed and Waters, 2005; Hoeijmakers, 2001, 2009).

DNA repair has so far been shown to exist in both prokaryotic and eukaryotic organisms, with over 150 proteins directly involved in safeguarding the genome (Sancar et al., 2004; Friedberg et al., 2005; Wood et al., 2005; Yousefzadeh et al., 2021). DNA repair processes restore DNA back to its normal sequence and structure after damage (Friedberg et al., 2006), and are characterised traditionally by the type of damage they repair. There are five major DNA repair pathways available to cells to deal with DNA damage burdens. Each of these processes recognises a particular type of DNA lesions, and together work in preventing mutagenesis. They include 1) direct reversal repairs that repairs lesion induced mainly by alkylating agents, 2) Base excision repair (BER), for small base modifications like single-strand breaks (SSBs) and non-bulky damaged DNA bases, 3) Nucleotide excision repair (NER), that corrects bulky, helix-distorting DNA lesions, 4) mismatch repair (MMR), that repairs base-base mismatch and insertion or deletion loops (IDLs), 5) Recombinational repair, which is divided into non-homologous end joining (NHEJ) and homologous recombination repair (HR), both of which repairs DNA double-strand breaks (DSBs). Other types of DSB repairs include alternative non-homologous end-joining (alt-NHEJ, MMEJ) and translesion synthesis (TLS), which operates as a tolerance mechanism for DNA damage (Jackson and Bartek, 2009; Hosoya and Miyagawa, 2014; Li et al., 2021).

For frequently occurring DNA damage, direct reversal of DNA damage by specialised proteins is the most efficient and most straightforward method of DNA repair. However, this approach is only used by a small proportion of DNA repair types. Most damage to DNA is repaired by the removal of damaged bases and is followed by resynthesis of the removed/excised region (replacement) (Cooper, 2000). The pathways involved in the removal of base damage are base excision repair (BER), nucleotide excision repair (NER) and mismatch repair (MMR) (Cooper, 2000). The rest of the pathways repair damage to DNA structure/backbone. DNA damage can cause breaks in the DNA backbone, single-strand breaks in one strand, or double-strand breaks on both strands. Single-strand breaks are repaired by mechanisms sharing common steps in the BER pathway; however, DSBs are especially harmful as, by definition, no unbroken complementary strand exit which can serve as a template for repair when both strands break (Bennett et al., 1993; Friedberg et al., 2006). For cells with DNA already replicated prior to cell division, the duplicate copy can easily supply the missing information. So in these cells, DSBs can be repaired by HR, involving the exchange of DNA strands.

Even so, very efficient repair mechanisms can sometimes fail to provide a clean template for DNA synthesis. Replication errors can make it past these mechanisms, as DNA repair can also undergo mutations and become dysregulated. DNA repair gene mutations have been known to cause a variety of rare inherited human syndromes. Some of which include premature ageing phenotypes, increased sensitivity to ionising radiation exposure, and increased cancer risk (Friedberg et al., 2006; Lok and Powell, 2012; Carusillo and Mussolino, 2020). It has also been found that inherited defects in each of the DNA repair pathways are associated with distinct genome instability syndromes (Yousefzadeh et al., 2021), syndromes characterised by developmental defects (Bouwman and Jonkers, 2012; Ghosal and Chen, 2013; Wolters and Schumacher, 2013; Wood, 2018).

The dysregulation of DNA repair gene networks underlies many human genetic diseases that affect a wide range of body systems but all share a common trait, predisposition to cancer (Chatterjee and Walker, 2017). Almost all human cancers are spontaneous, not inherited, and are caused by environmental or genetic factors. It is of great public health interest to determine which genetic variations increase cancer risk in normal populations, and DNA repair genes are likely contenders.

Therefore, elucidating the molecular mechanism behind DNA repair defects may provide a framework for understanding the complex pattern of genetic variations that contribute to spontaneous human cancers.

In this study, we demonstrate how to translate a network (mathematically definable as a hyper-graph) into a set of first-order differential equations of the mass action law type. Once the model has been established, we present its numerical solution and carry out a sensitivity analysis of its kinetic rates, whose numerical values are mostly unknown. The results of the analysis of network dynamics complement those produced by the calculation of centrality measures, and together they produce a set of genes of similar biological interest, and in perspective also of medical interest, due to their characteristics of topological centrality and vulnerability to stimuli. The paper is organized as follows: in **Section 2** we describe the mechanisms of double-strand break repair pathway homologous recombination repair necessary to understand and interpret the results of the computational analysis, in **Section 3** we describe the rules on which the automatic translation of the network into a rate equation system is based and the methods of sensitivity analysis of the model. In **Section 4** we present the results of the analysis, and, finally, in **Section 5**, we draw some conclusions.

2 DOUBLE-STRAND BREAK REPAIR PATHWAY HOMOLOGOUS RECOMBINATION REPAIR

DSBs are the most serious DNA damage, as both DNA strands are impaired simultaneously. Therefore, due to the magnitude of differing factors leading to DSBs, the effectiveness of their repair is crucial for cell survival and the functioning and prevention of DNA fragmentation, chromosomal translocation and deletion. DSBs can be repaired in mammalian cells by NHEJ, HR, and single-strand annealing (SSA). Unrepaired SSBs result in much more cytotoxic DSBs formation during the S-phase progression of the cell cycle (Kennedy and D'Andrea, 2006). Homologous recombination is a process by which DSBs are repaired through the alignment of homologous sequences of DNA (Dietlein and Reinhardt, 2014) and occurs primarily during the late S to G2 phase of the cell cycle (Cerbinskaite et al., 2012; Chatterjee and Walker, 2017).

Homologous recombination is the second major DSB repair pathway and requires a second, homologous DNA sequence to function as donor template. There are two phases to this process, the first phase triggered by sensor proteins that belong to the MRN complex, and the second phase by the stimulation of resection steps, initiated in the first phase and subsequently extended. HR generally involves the following stages:

1. DSBs are recognised and sensed by the MRN complex (Kim et al., 1994), which activates ATM kinase, initiating the DSB end resection steps, where CtBP-interacting protein (CtIP) and the MRN complex work together to generate single-strand DNA (ssDNA) at the DSB ends ((Zhao et al., 2020).
2. The exposed ssDNA is recognised by and coated with DNA replication protein A (RPA) complex, which recruits the major homologous recombination regulator RAD52 to the site to facilitate HR repair (Maréchal and Zou, 2014; Rossi et al., 2021).
3. The nucleoprotein filament RAD51, is then assembled, mediated by BReast CAncer type 2 susceptibility protein (BRCA2), to replace RPA on ssDNA to perform homology sequence searching and strand invasion (Kowalczykowski, 2015).
4. DSBs are then restored by branch migration, DNA synthesis, ligation, and resolution of Holliday junctions (Zhao et al., 2020).

Following the recognition and sensing of DSBs, a process known as DNA end resection is activated, a critical function in HR (Liu and Huang, 2016; Zhao et al., 2020). DNA end resection catalyses the nucleolytic degradation of the broken ends of DSBs (by the CtIPMRN complex) in the 5′ to 3′ direction generating 3′ single-stranded DNA (ssDNA). The 3′ ssDNA then provides a platform for the recruitment of HR repair-related proteins (Huertas, 2010; Liu and Huang, 2016; Zhao et al., 2020). Following the generation of ssDNA, downstream nucleases and helicases, such as exonuclease 1 (EXO1) or DNA replication ATP-dependent helicase/nuclease DNA replication helicasenuclease 2 (DNA2) and Bloom syndrome protein (BLM), are conscripted to extend the 3' ssDNA for HR repair (Huertas and Jackson, 2009; Yun and Hiom, 2009; Zhao et al., 2020). The identities of these DNA helicases and nucleases are yet to be clearly defined in humans (as in yeast), partly because there are many candidate proteins. Although five RecQ helicase homologs have so far been identified in yeast (Bloom helicase [BLM], Werner helicase/nuclease [WRN], RECQ1, RECQ4, and RECQ5) (Chu and Hickson, 2009; Lu and Davis, 2021), convincing evidence point up BLM in resection (Gravel et al., 2008; Nimonkar et al., 2008). Following resection, the exposed single-strand DNA (ssDNA) is recognised and bound by RPA complex for protection.

RPA plays a significant role in coordinating DNA resection processes and simultaneously preserving the integrity of the resultant ssDNA (Sun et al., 2019). RPA is a heterotrimeric ssDNA binding protein essential to nearly all DNA processing events and associates with ssDNA with very high affinity (Kd sim 109–1010 M) (Maréchal and Zou, 2014). It is comprised of three protein subunits, RPA70, RPA32 and RPA14 and contains multiple oligonucleotideoligosaccharide (OB)-folds that interact with both ssDNA and proteins (Kim et al., 1994; Fanning, 2006; Feldkamp et al., 2014; Maréchal and Zou, 2014). RPA is flexible (Brosey et al., 2013). Its versatile nature allows it to coordinate the recruitment, activation and exchange of many proteins whose combined activities allow for the protection and propagation of eukaryotic genomes (Maréchal and Zou, 2014). How multiple RPAs associate on ssDNA and coordinate its vast array of processes remains to be determined (Sun et al., 2019). However, a critical feature of RPA is that, though it can bind nucleic acids with very high affinity, it can

TABLE 1 | Interactive graphical user interface of NADS software showing the options and the task concerning the generation of ODE equations and their solution.

```
##############################################################
Welcome to Network Analyzer and Dynamics Simulator
##############################################################

Which network do you want to analyze?

1: Homologous recombination interactions
2: Test example
3: Exit

Selection: 1

################################
Principal network characteristics
################################

Number of nodes: 25
Number of edges: 250

Types of interactions and their frequency:

       controls expression of 1
       controls state change of 134
       in catalysis with 41
       interacts with 74

##############################
Centrality measures calculation
##############################

Degree

Betweennes

Clustering coefficient

Eingenvector centrality

Vibrational centrality
... pseudo-inverse of graph Laplacian ...

Subgraph centrality

Information centrality
... pseudo-inverse for information centrality estimation ...
Export results to files ...

Select last decile nodes from the centralities measure distributions ...
Nodes highly scoring by Clustering coefficient : empty set.

Select first decile nodes from the centralities measure distributions ...

... save the results of selection in barplots.

Saving plots of centralities distributions.
```

TABLE 2 | Interactive graphical user interface of NADS software showing the options and the task concerning the analysis of network topology.

```
#######################################################
            Build and solve network dynamics
#######################################################

Network size reduction: the union of the nodes of 99-th percentile
of centrality distributions will be selected.

Do you want to reduce the size of the network to analyze?
(It is highly recommended to select Yes if you want to reduce
the computational times.)

1: Yes
2: No
3: Exit

Selection: 2

The number of nodes of the network is:  25
The number of edges of the network is:  250

Start to convert this network into a set of ordinary differential
equations (ODE) ...

The initial values of the variables are selected randomly in the
range [1, 100].
Do you want to change this range?

1: Yes
2: No
3: Exit

Selection: 2

The input file does not contain information on the kinetic rate constants:
random values from a given range will be selected (default is [0,0.01]).
Do you want to change this range?

1: Yes
2: No
3: Exit

Selection:  2

... script dynamics.R has been created.......: 11.801 sec elapsed
Solve the ODE of the network dynamics ...: 11.961 sec elapsed
```

easily be displaced by other enzymes for further downstream processing (Sun et al., 2019).

When ssDNA length is sufficient for HR repair, the end resection process is terminated (Zhao et al., 2015). Although the regulation of DNA end resection termination are not yet clearly understood, several studies suggest that under physiological conditions, end resection is terminated by RAD51-RPA switching (Zhao et al., 2015). This switching is regulated by BRCA2-DSS1. DSS1 - SEM1 in yeast - is a small, highly acidic protein that competes with ssDNA, by mimicking ssDNA in order to remove RPA from the genuine ssDNA (Zhao et al., 2015; Stefanovie et al., 2019; Le et al., 2020; Rossi et al., 2021). The DSS1 then binds to BRCA2 in order to facilitate RAD51 filament formation (Liu et al., 2010; Stefanovie et al., 2019; Rossi et al., 2021). DSS1 does not seem to bind DNA on its own but appears to enhance ssDNA binding activities of BRCA2 and RAD52 to promote DSB repair (Zhao et al., 2015). BRCA2 then recruits RAD51 to complete the switch

(Zhao et al., 2015; Rossi et al., 2021). For cells with DNA already replicated prior to cell division, RAD51 will oligomerise and form a nucleoprotein filament on the resected, single-stranded DNA (ssDNA) end of the DSB, and search for the homologous DNA sequence on the undamaged sister chromatid, performs strand exchange (invasion), and produce a joint molecule called a D-loop (Rossi et al., 2021). DNA polymerase will then use the homologous DNA strand as a template from the D-loop, and the 3′-end of the broken DNA strand as a primer to commence DNA repair synthesis (Rossi et al., 2021). The other end of the double-strand break is then apprehended by RAD52, joining it to the D-loop, through the annealing process, causing the displaced strand to act as a template for the second strand synthesis (Rossi et al., 2021). When DNA synthesis is complete, the D-loops are then dissociated by RAD54, a protein that interacts with RAD51 to promote branch migration, or interacts with helicases like BLM (van Brabant et al., 2000; Bugreev et al., 2006; Kawale and Sung, 2020; Rossi et al., 2021). DNA is further extended by DNA polymerase, annealed to the ssDNA part of the second broken DNA, gap filled and finally restored (Rossi et al., 2021). In **Figure 1** we summarise what is described in this section about the DSB signalling mechanisms.

Homologous recombination is able to repair DSBs error-free using the undamaged sister chromatid (Dietlein and Reinhardt, 2014). As the accuracy of homologous recombination repair is important for DSBs (Sugiyama and Kantake, 2009), if it is impaired, chemotherapeutic opportunities may arise (Huang and Zhou, 2021).

3 GRAPH REPRESENTATION AND MATHEMATICAL MODEL

We considered HDR through Homologous Recombination (HRR) network as available in Pathways Commons (Cerami et al., 2010) in the SIF (Simple Interaction Format) format at the link in the reference (Orlic-Milacic, 2015). See these data also reported in **Supplementary Tables S1–S3**.

We implemented an R script, that takes as input the HR network and is able to.

- analyse the topology of the network through the calculation of standard and new centrality measures. The standard node centrality measures considered in this study are the degree (in-, out- and total), the betweenness, the clustering coefficient, the eingenvector centrality, the vibrational centrality, the subgraph centrality, and the information centrality (see (Marsden, 2005; Koschützki and Schreiber, 2008; Ghasemi et al., 2014; Wang et al., 2014; Fornito et al., 2016; Jalili et al., 2016; Ashtiani et al., 2018) for a concise but comprehensive report on the meaning and the use of these measure in molecular biology). We considered also a new centrality measures, such as vibrational centrality, introduced by Estrada in (Estrada and Hatano, 2010) that we will discuss in more detail in the next section (we also refer the reader to (Lecca and Re, 2019) for a review on vibrational centrality); for the reader's convenience, we list the definition of these centrality indices in **Supplementary Table S4**, that are also extensively covered in many textbooks on graph theory, and in various articles in the applied sciences. We refer the reader to Estrada's numerous works, a comprehensive compendium of which can be found in the book (Estrada and Hatano, 2010; Estrada, 2011);
- automatically generate a system of rate equations, specifically first order mass action differential equations, describing the dynamics of the network.

and a R script implementing parametric sensitivity analysis of the dynamics model.

In the dynamics model, by default the kinetic rate constants k a well as the initial values of the proteins and molecules concentrations are set equal to random values in fixed ranges. Nevertheless these ranges can be modified by the user as shown by the interactive console output reported in **Table 1**. However, we note that the interval of definition of the uniform distribution cannot exceed the maximal range of parameter variability within which the system of rate equations has a solution. We refer the reader to a previous work of us (Lecca et al., 2016) for more details on this.

In a first experiment, the initial values of the proteins concentration (not experimentally known) has been drawn randomly in a range [1, 100] a. u., and the simulation time interval was [0, 10] a. u. In a second experiment, the numerical simulation of the model was performed by assigning an initial quantity between 18 and 20 (expressed in arbitrary units a. u.) to each node and for $t \in [0, 1400]$ (in arbitrary units). The solution of the system of 25 differential equations converges for values of rate constants in the range $]0, 0.01$ a. u. in the first experiment, and in the range $]10^{-6}, 10^{-5}$ a.u. in the second experiment. To the best of our knowledge, the *in vivo* concentrations of the proteins that are part of the HR network are not known. Combined with the lack of knowledge of the values of rate constants, this means that we cannot assign units to these values that reflect *in vivo* kinetics. However, the initial values of the protein concentrations have been chosen from the order of magnitude at which various *in vitro* experiments operate and express these concentrations in nM and time in seconds. The literature we have referred to includes the works of (Yang et al., 2010; Nguyen et al., 2013; Foertsch et al., 2019).

We changed the parameters one at a time while keeping the values of the others fixed. Since for each parameter p_h (h ranges from 1 to the number of parameters in the equations), we sampled N_P values, and consequently we performed N_P model simulations. Let us denote with $x_s(t)$, ($s = 1, 2, \ldots, d$) the time series expressing the numerical solutions of the rate equations, where d is the number of the proteins in the network. The index of sensitivity of $x_s(t)$ with respect to the change of h-th parameter from the value p_h to the value p'_h is calculated as in (Lecca et al., 2016) by

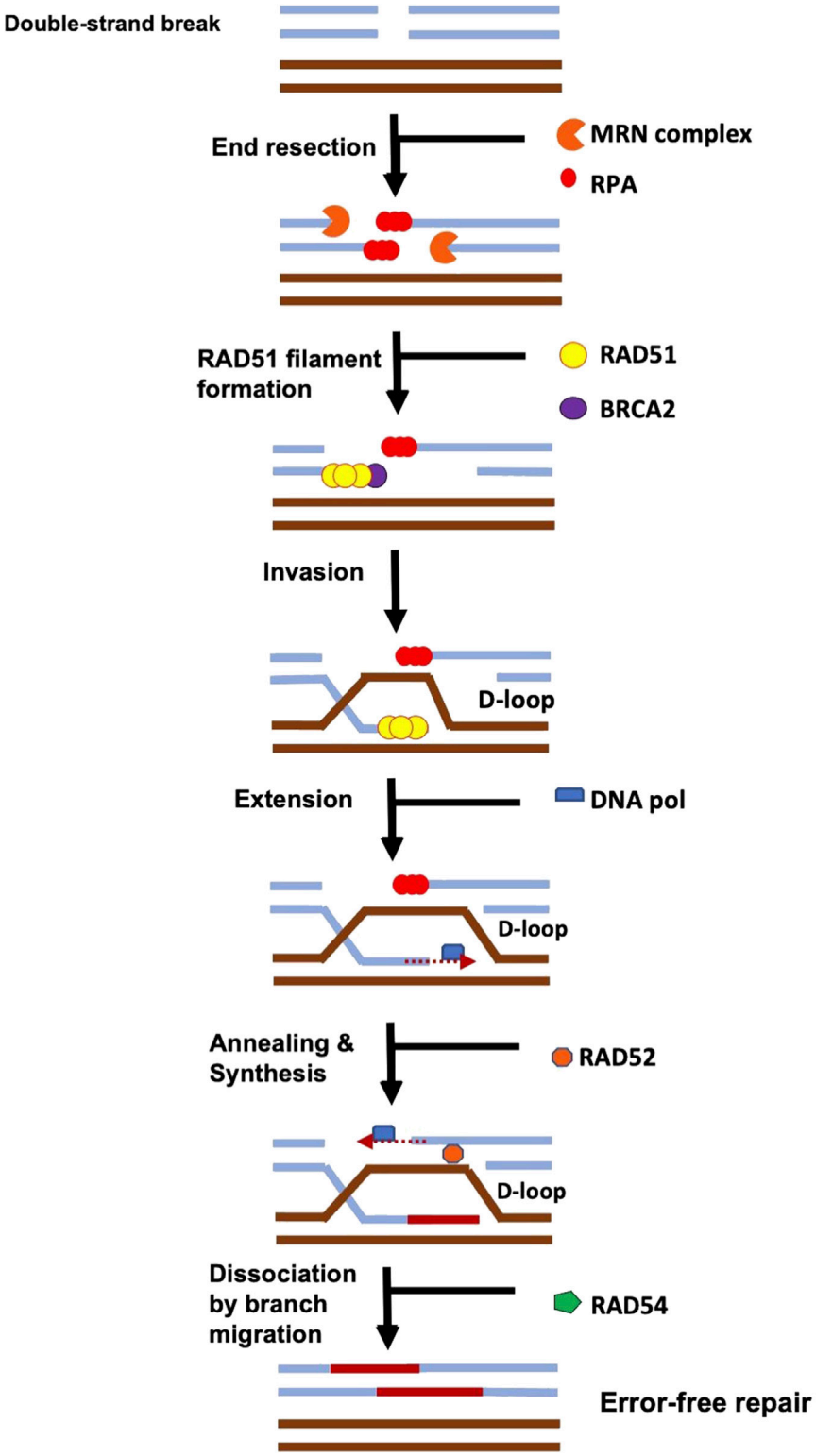


FIGURE 1 | The signalling of a DSB is initiated via the binding of the MRN complex which initiates resection. During HR, the ends of the double-strand break (DSB) are resected by nucleases, exposing single-strand DNA (ssDNA) that becomes bound by RPA. The mediator protein, BRCA2 initiates the loading of RAD51 onto ssDNA, helping to displace RPA. RAD51 oligomerizes, forms a nucleoprotein filament, and then searches for the homologous DNA sequence on the intact chromosome. RAD51 filament invades the intact dsDNA and forms a D-loop structure. It is further processed by DNA polymerases, chromatin remodelers (RAD54), nucleases, and ligases to restore it back to its original sequences. (Adapted from (Rossi et al., 2021).

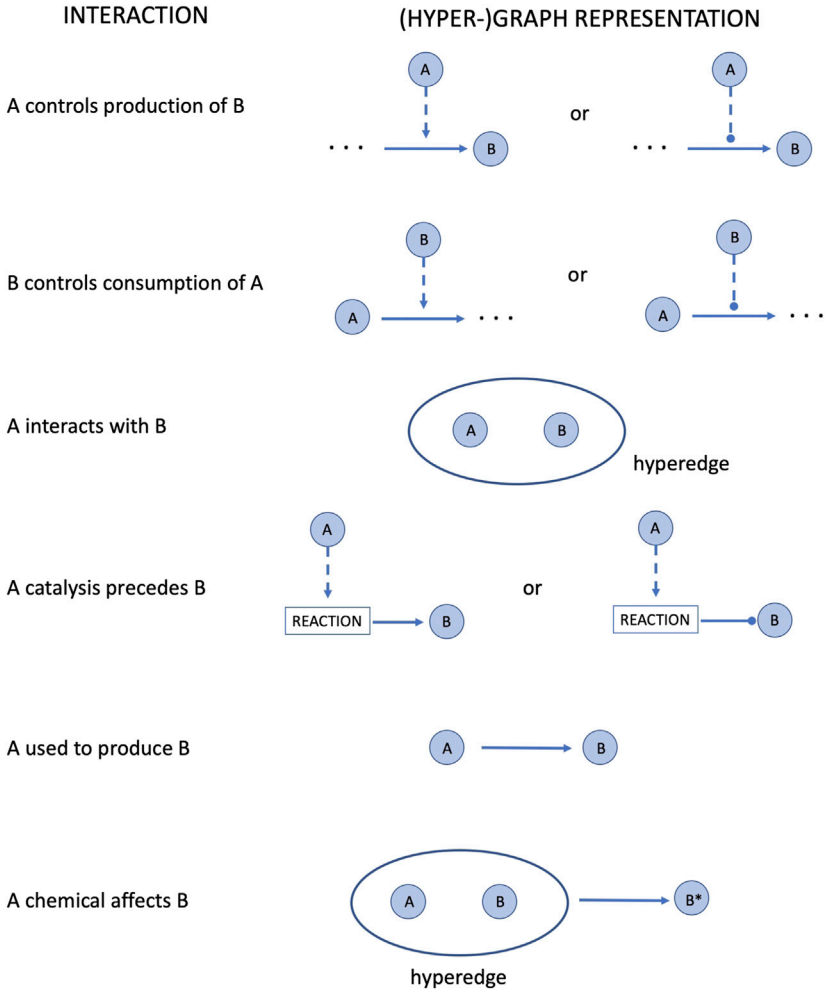


FIGURE 2 | Conversion of the SIF format interactions into a (hyper-)graph structure.

$$SI_{ish} = \frac{1}{N}\sum_{k=1}^{N}\left(\frac{1}{N_P - 1}\sum_{r=1}^{N_P}\left(x_s^{(r)}(t_k|p_h \leftarrow p_h') - \overline{x_s}^{(r)}(p_h \leftarrow p_h')\right)^2\right)^{\frac{1}{2}} \quad (1)$$

where N is the length of the time series $x_s(\mathrm{t})$, and

$$\overline{x_s}(p_h \leftarrow p_h') = \frac{1}{N_P}\sum_{k=1}^{N_P} x_s(t_k|p_h \leftarrow p_h'). \quad (2)$$

where "$p_h \leftarrow p_h'$" means "p_h replaced by p_h'". The **Eq. 1** defines the mean of the standard deviations of the distributions of the simulated values of a protein/gene abundance at time points t_k.

In case the user knows the values of the rate constants, he/she can add them as an extra column to the SIF format of the input files. At the moment of writing, for most interactions the values of the kinetic constants are not known and there are no time-resolved data from which it is possible to infer them. It is precisely this context that justifies our choice to study the dynamics of the system in a range of values of the model parameters and more generally to provide a software that can be used as a platform for in silico experiments.

We implemented Network Analyser and Dynamics Simulator (NADS) consists of three modules written in R language:

TABLE 3 | Translation of BioPAX interactions into simple ordinary differential equations. See in **Figure 2** the (hyper-)graph representation of these interactions.

Interaction	Differential equation
A controls production of B	$\frac{dB}{dt} = kA$, $\frac{dA}{dt} = 0$
B controls consumption of A	$\frac{dA}{dt} = -kB$, $\frac{dB}{dt} = 0$
A interacts with B	$\frac{dA}{dt} = \frac{dB}{dt} = -kAB$
A catalysis precedes B	$\frac{dB}{dt} = \pm kA$
A used to produce B	$\frac{dB}{dt} = kA$
A chemical affects B	$\frac{dA}{dt} = 0$, $\frac{dB}{dt} = -kB = -\frac{dB^*}{dt}$

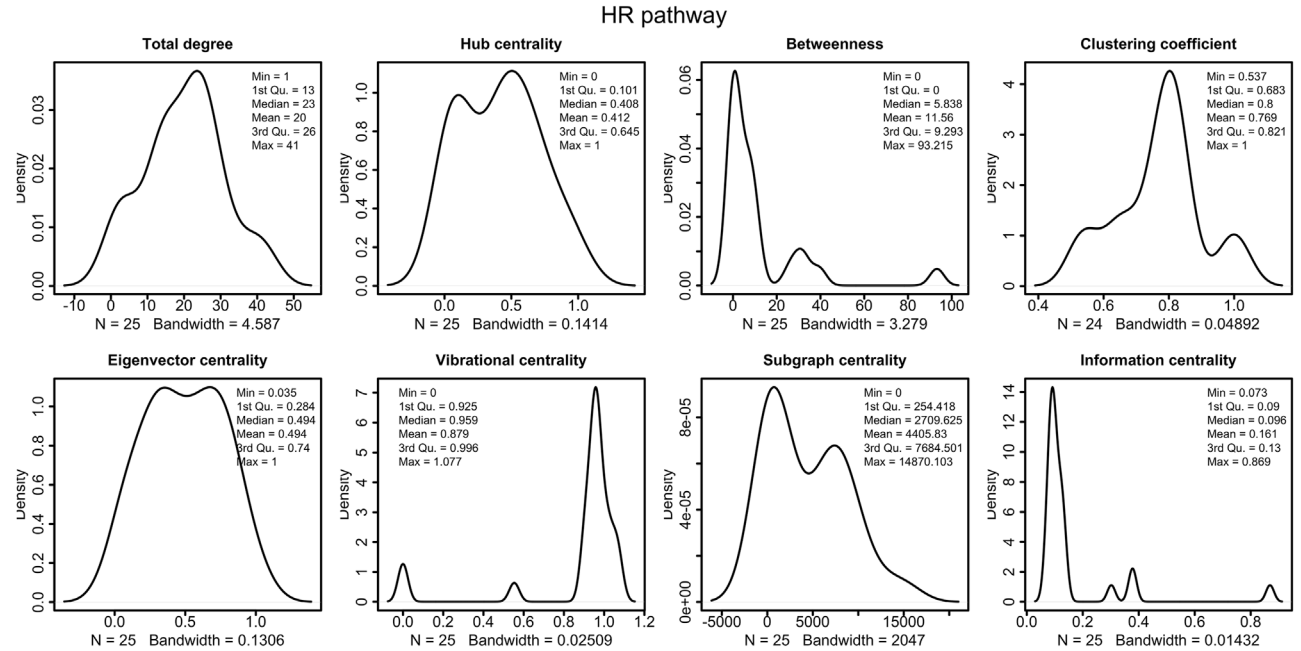


FIGURE 3 | Distributions of the centrality measures of HR pathway (Orlic-Milacic, 2015). We observe that the majority of node have low betweenness, low information centrality and high vibrational centrality.

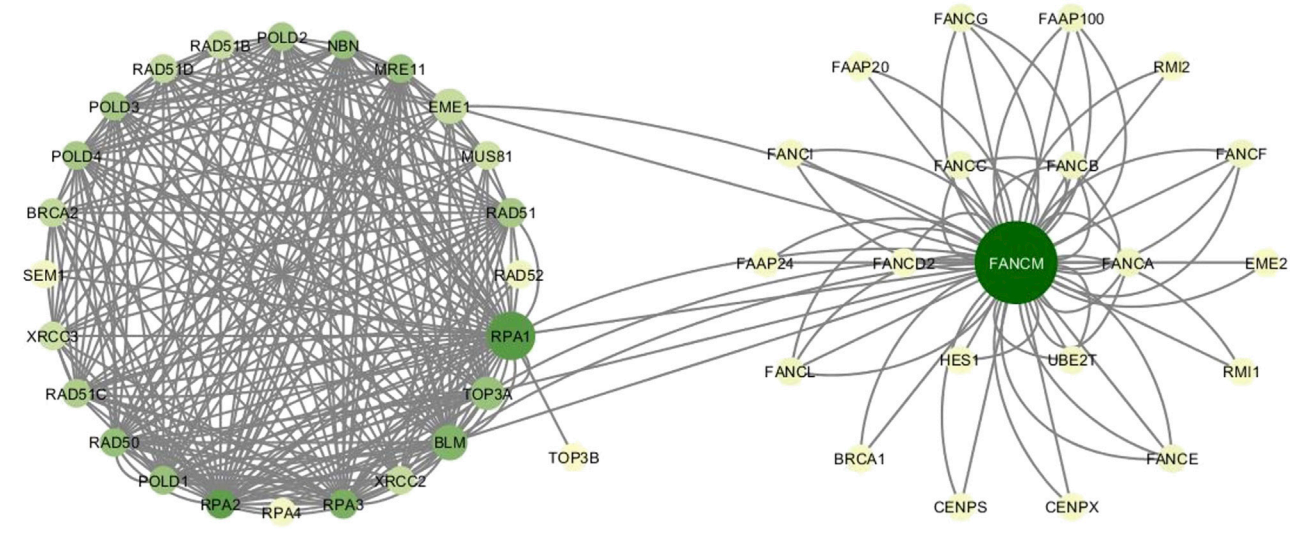


FIGURE 4 | HR network and FANCM network. Colors vary from yellow to green according to increasing degree values. Node sizes grow as the betweenness centrality of nodes.

- network_analysis_functions.R: this module implements the functions that processes the SIF data-frames to make them suitable to their conversion into a graph. This module also implement the functions to rank the nodes according to their centrality measures;
- graph_parser.R: this module converts the SIF format network into an R script that solves the corresponding ordinary differential equations;
- network_analysis.R: this module first calls network_analysis_functions.R and performs the networks analysis, and secondly it calls the module graph_parser.R that generates the script dynamics.R containing the differential equations of the network dynamics.

The user can launch the software simply by running in RStudio the script network_analysis.R, and then by answering

TABLE 4 | Values of the centrality measures for the HR pathway in (Orlic-Milacic, 2015). In bold, we marked the genes/proteins with the highest scores.

Protein	Total degree	In-degree	Out-degree	Hub centrality	Betweenness	Clustering coefficient	Eigenvector centrality	Vibrational centrality	Subgraph centrality	Information centrality
BLM	30	19	11	4.08E-01	39.5138622	0.2183908	0.81043641	**1.0771148**	7408.13084	0.08351073
EME1	13	0	13	5.21E-01	0	0.5384615	0.2589557	0.9579369	2709.62496	0.13185164
MRE11	26	16	10	3.69E-01	9.2926175	0.1969231	0.73968044	1.0587524	6020.65076	0.08957011
MUS81	13	1	12	5.18E-01	0	0.5384615	0.2589557	0.9547013	2291.00787	0.13177729
NBN	27	17	10	3.67E-01	9.2926175	0.1823362	0.76703983	1.0603542	6020.65076	0.08786049
POLD1	26	9	17	6.83E-01	7.8121197	0.36	0.67253587	0.9655278	7684.50086	0.08965382
POLD2	23	5	18	7.10E-01	5.8377525	0.4624506	0.57100865	0.9672104	9065.81813	0.09552092
POLD3	23	8	15	6.45E-01	5.8377525	0.4624506	0.57100865	0.961071	5521.89505	0.09524672
POLD4	23	7	16	6.66E-01	5.8377525	0.4624506	0.57100865	0.963361	6513.86378	0.09533952
RAD50	27	13	14	5.79E-01	10.2729789	0.2108262	0.75819149	1.0206621	**11427.9652**	0.08811688
RAD51	23	12	11	4.03E-01	30.3997662	0.4426877	0.49447273	0.9038147	1403.03278	0.09426237
RAD51B	14	12	2	8.58E-02	0.3636364	0.7032967	0.31841389	0.9960094	74.47586	0.12453711
RAD51C	21	9	12	4.96E-01	1.6614219	0.4095238	0.47619405	0.9027594	1911.88011	0.09955242
RAD51D	16	11	5	2.15E-01	0.5127787	0.7166667	0.3593928	0.9248884	282.05774	0.11524626
RAD52	4	2	2	1.01E-01	0	**1**	0.1247068	0.9567526	254.41809	0.30199153
RPA1	**41**	20	21	9.02E-01	**93.2152627**	0.1512195	0.89284007	0.9069973	8617.68902	0.0726257
RPA2	**41**	**22**	19	8.46E-01	32.6964097	0.1487805	**1**	0.9404821	8508.98841	0.07306925
RPA3	34	11	**23**	**1.00E+00**	25.979693	0.2174688	0.819185	0.9803911	**14870.10272**	0.07928594
RPA4	3	0	3	1.27E-01	0	**1**	0.10547581	0.9592952	700.52299	0.37822514
SEM1	3	1	2	9.14E-02	0	**1**	0.06803348	0.9398974	63.99639	0.37757664
TOP3A	24	12	12	4.64E-01	9.2926175	0.2318841	0.68062612	1.0251952	8739.44955	0.09349123
XRCC2	15	14	1	5.60E-02	0.8476272	0.7142857	0.34875334	1.0105203	55.0254	0.11916455
XRCC3	13	12	1	4.68E-02	0.3333333	0.8076923	0.2842042	0.5530372	0	0.13016119
BRCA2	16	16	0	1.40E-16	0	0.45	0.36198393	0	0	0.11465992
TOP3B	1	1	0	8.77E-18	0	NA	0.03472425	0	0	**0.86898249**

to the questions in the interactive interface as shown in **Table 2**.

We provide also the fourth module implementing the parametric sensitivity analysis, named sensitivity_analysis.R, which takes as an input the system of equations automatically generated by network_analysis.R.

The main module is the script network_analysis.R. As soon as the user runs it from the R Studio GUI (R Studio, 2022) or from a terminal, an interactive output is displayed as in **Tables 2**, **3**. The program asks the user to select the network to be analysed and then.

- it calculates the centralities measures
- it translates the SIF network into a hyper-graph structure according to the rules reported in **Figure 2**
- then, it translates the hyper-graph into a set of ordinary differential equations, according to the rules reported in **Table 3**,
- and, finally, it solves them.

The program returns also the execution times for the tasks expected to be the most computationally demanding, such as integrating the equation.

4 RESULTS

The HR network considered in this study has 25 nodes and 250 edges, as reported in **Supplementary Tables S1–S3**. The right part of the **Figure 4** shows the HR network in circular layout. The network analysis phase of our study calculated the centrality measures distributions show in **Figure 3**, and identified six genes, as shown in **Table 4**:

1. BLM, scoring first for vibrational centrality
2. RAD50 scores first for sub-graph centrality
3. RAD52, scoring first for clustering coefficient
4. RPA1, scoring first for total degree and betweenness
5. RPA2, scoring first for in-degree and eigenvector centrality
6. RPA3, scoring first for out-degree, hub centrality, and sub-graph centrality
7. RPA4, scoring first for clustering coefficient
8. SEM1, scoring first for clustering coefficient.

Of particular interest is the fact that BLM has the highest vibration centre. The interpretation of this result is that BLM is the node most sensitive to stresses and/or stimuli, i.e., according to the vibrational centrality measure, it is the most vulnerable node in the network (Estrada and Hatano, 2010). This result is of particular interest in light of the crucial role this gene plays in the HR network. Indeed, the key role of BLM is well know, and alterations in this protein is linked to different diseases including cancer (Kaur et al., 2021). BLM is a 3′-5′ ATP-dependent RecQ DNA helicase. It is a genome stabilizer playing an essential role in the DNA replication regulation, DNA recombination, and both homologous and non-homologous pathways of DSB repair. The high vulnerability of the BLM node to external stimuli and conditions suggests the need to identify which conditions and/or stimuli may be altering it, in order to preserve its proper functioning and/or to understand how it can be restored if it is

TABLE 5 | This is the PART I of the table of ordinary differential equations of the dynamics of HR network, in R code formalism. The *k* followed by a number denote the kinetic rate constant, and the letter "d" in front of the name of the proteins denote the temporal derivative of it concentration.

```
dBLM  =  - k12*EME1  - k34*MRE11*BLM  - k35*MUS81  - k55*NBN*BLM
 - k57*POLD1  - k74*POLD2  - k92*POLD3  - k107*POLD4  - k132*RAD50*BLM
 - k138*RAD51*BLM  - k162*RAD51D*BLM  - k169*RPA1  - k182*RPA1*BLM
 - k190*RPA2  - k202*RPA2*BLM  - k209*RPA3  - k222*RPA3*BLM
 - k248*TOP3A*BLM  - k249*XRCC2*BLM

dBRCA2  =  - k13*EME1 - k36*MUS81 - k58*POLD1 - k75*POLD2 - k93*POLD3
 - k108*POLD4  - k139*RAD51*BRCA2  - k150*RAD51C
   - k155*RAD51C*BRCA2  - k170*RPA1  - k183*RPA1*BRCA2
 - k191*RPA2  - k203*RPA2*BRCA2  - k210*RPA3  - k235*SEM1*BRCA2
 - k250*XRCC3*BRCA2

dMRE11  =  - k1*BLM  - k14*EME1  - k34*MRE11*BLM  - k37*MUS81
 - k56*NBN*MRE11  - k59*POLD1  - k76*POLD2  - k94*POLD3
 - k109*POLD4  - k133*RAD50*MRE11  - k171*RPA1  - k184*RPA1*MRE11
 - k192*RPA2  - k204*RPA2*MRE11  - k211*RPA3  - k223*RPA3*MRE11
 - k237*TOP3A

dNBN  =  - k2*BLM  - k15*EME1  - k32*MRE11  - k38*MUS81  - k55*NBN*BLM
 - k56*NBN*MRE11  - k60*POLD1  - k77*POLD2  - k95*POLD3  - k110*POLD4
 - k130*RAD50  - k134*RAD50*NBN  - k172*RPA1  - k185*RPA1*NBN
 - k193*RPA2  - k205*RPA2*NBN  - k212*RPA3  - k224*RPA3*NBN
 - k238*TOP3A

dPOLD1  =  - k3*BLM  - k25*MRE11  - k47*NBN  - k72*POLD1*POLD3
 - k73*POLD1*POLD4  - k89*POLD2*POLD1  - k123*RAD50  - k186*RPA1*POLD1
 - k206*RPA2*POLD1  - k225*RPA3*POLD1  - k239*TOP3A

dPOLD2  =  - k4*BLM  - k26*MRE11  - k48*NBN  - k89*POLD2*POLD1
 - k90*POLD2*POLD3  - k91*POLD2*POLD4  - k124*RAD50  - k240*TOP3A

dPOLD3  =  - k5*BLM  - k27*MRE11  - k49*NBN  - k72*POLD1*POLD3
 - k90*POLD2*POLD3  - k122*POLD4*POLD3  - k125*RAD50  - k241*TOP3A

dPOLD4  =  - k6*BLM  - k28*MRE11  - k50*NBN  - k73*POLD1*POLD4
 - k91*POLD2*POLD4  - k122*POLD4*POLD3  - k126*RAD50  - k242*TOP3A

dRAD50  =  - k7*BLM  - k16*EME1  - k39*MUS81  - k61*POLD1  - k78*POLD2
 - k96*POLD3  - k111*POLD4  - k132*RAD50*BLM  - k133*RAD50*MRE11
 - k134*RAD50*NBN  - k135*RAD50*RPA1  - k136*RAD50*RPA2  - k137*RAD51
 - k173*RPA1  - k194*RPA2  - k213*RPA3  - k226*RPA3*RAD50  - k243*TOP3A

dRAD51  =  - k17*EME1  - k40*MUS81  - k62*POLD1  - k79*POLD2  - k97*POLD3
 - k112*POLD4  - k138*RAD51*BLM  - k139*RAD51*BRCA2  - k140*RAD51*RAD51B
 - k141*RAD51*RAD51D  - k142*RAD51*RAD52  - k143*RAD51*RPA1  - k144*RAD51*RPA2
 - k145*RAD51*SEM1  - k146*RAD51*XRCC2  - k147*RAD51*XRCC3  - k151*RAD51C
 - k156*RAD51C*RAD51  - k174*RPA1  - k195*RPA2  - k214*RPA3  - k227*RPA3*RAD51

dRAD51B  =  - k18*EME1  - k41*MUS81  - k63*POLD1  - k80*POLD2  - k98*POLD3
 - k113*POLD4  - k140*RAD51*RAD51B  - k148*RAD51B*RPA1  - k149*RAD51B*XRCC2
 - k157*RAD51C*RAD51B  - k163*RAD51D*RAD51B  - k175*RPA1  - k196*RPA2
 - k215*RPA3

dRAD51C  =  - k19*EME1  - k42*MUS81  - k64*POLD1  - k81*POLD2  - k99*POLD3
 - k114*POLD4  - k155*RAD51C*BRCA2  - k156*RAD51C*RAD51  - k157*RAD51C*RAD51B
 - k158*RAD51C*RAD51D  - k159*RAD51C*RPA1  - k160*RAD51C*XRCC2
 - k161*RAD51C*XRCC3  - k176*RPA1  - k197*RPA2  - k216*RPA3
```

TABLE 6 | This is the PART II (continuation) of the table of ordinary differential equations of the dynamics of HR network, in R code formalism. The *k* followed by a number denote the kinetic rate constant, and the letter "d" in front of the name of the proteins denote the temporal derivative of it concentration.

```
dRAD51D  =  - k20*EME1 - k43*MUS81  - k65*POLD1  - k82*POLD2
 - k100*POLD3 - k115*POLD4  - k141*RAD51*RAD51D  - k158*RAD51C*RAD51D
 - k162*RAD51D*BLM  - k163*RAD51D*RAD51B  - k164*RAD51D*RPA1
 - k165*RAD51D*XRCC2 - k166*RAD51D*XRCC3  - k177*RPA1
 - k198*RPA2  - k217*RPA3

dRAD52  =  -k142*RAD51*RAD52  - k167*RAD52*RPA1
 - k168*RAD52*RPA2  - k228*RPA3*RAD52

dRPA1  =  - k8*BLM  - k29*MRE11  - k51*NBN  - k66*POLD1  - k83*POLD2
 - k101*POLD3  - k116*POLD4  - k127*RAD50  - k135*RAD50*RPA1
 - k143*RAD51*RPA1 - k148*RAD51B*RPA1  - k152*RAD51C  - k159*RAD51C*RPA1
 - k164*RAD51D*RPA1 - k167*RAD52*RPA1  - k182*RPA1*BLM  - k183*RPA1*BRCA2
 - k184*RPA1*MRE11 - k185*RPA1*NBN  - k186*RPA1*POLD1  - k187*RPA1*TOP3A
 - k188*RPA1*TOP3B - k189*RPA1*XRCC2  - k207*RPA2*RPA1  - k229*RPA3*RPA1
 - k232*RPA4*RPA1 - k236*SEM1*RPA1  - k244*TOP3A

dRPA2  =  - k9*BLM  - k11*BLM  - k30*MRE11  - k33*MRE11  - k52*NBN
 - k54*NBN - k67*POLD1  - k84*POLD2  - k102*POLD3  - k117*POLD4
 - k128*RAD50  - k131*RAD50 - k136*RAD50*RPA2  - k144*RAD51*RPA2
 - k153*RAD51C  - k168*RAD52*RPA2 - k178*RPA1  - k202*RPA2*BLM
 - k203*RPA2*BRCA2  - k204*RPA2*MRE11 - k205*RPA2*NBN  - k206*RPA2*POLD1
 - k207*RPA2*RPA1  - k208*RPA2*TOP3A - k218*RPA3  - k230*RPA3*RPA2
 - k233*RPA4*RPA2  - k245*TOP3A  - k247*TOP3A

dRPA3  =  - k10*BLM  - k31*MRE11  - k53*NBN  - k68*POLD1  - k85*POLD2
 - k103*POLD3  - k118*POLD4  - k129*RAD50  - k154*RAD51C  - k222*RPA3*BLM
 - k223*RPA3*MRE11  - k224*RPA3*NBN  - k225*RPA3*POLD1  - k226*RPA3*RAD50
 - k227*RPA3*RAD51  - k228*RPA3*RAD52  - k229*RPA3*RPA1  - k230*RPA3*RPA2
 - k231*RPA3*TOP3A  - k234*RPA4*RPA3  - k246*TOP3A

dRPA4  =  -k232*RPA4*RPA1  - k233*RPA4*RPA2  - k234*RPA4*RPA3

dSEM1  =  -k145*RAD51*SEM1  - k235*SEM1*BRCA2  - k236*SEM1*RPA1

dTOP3A  =  - k21*EME1  - k44*MUS81  - k69*POLD1  - k86*POLD2
 - k104*POLD3 - k119*POLD4  - k179*RPA1  - k187*RPA1*TOP3A  - k199*RPA2
 - k208*RPA2*TOP3A  - k219*RPA3  - k231*RPA3*TOP3A  - k248*TOP3A*BLM


dEME1  =  -k24*EME1*MUS81

dMUS81  =  -k24*EME1*MUS81

dTOP3B  =  -k188*RPA1*TOP3B
```

altered. The high vulnerability of this node could also be explained by a recent study by Kaur et al. (2021). These authors report that BLM has a dual function both as a tumour suppressor and possibly as a proto-oncogene, being probably involved in the mechanisms of its deregulation in tumours.

The analysis also correctly identifies the SEM1 gene as a node with a high clustering coefficient. Indeed, as reported in (Safran et al., 2021; GeneCard, 2022), SEM1 gene encodes for a protein that is part of a 26S proteasome, which is a multiprotein complex with a function in the ATP-dependent degradation of ubiquitinated proteins. This complex contributes to the maintenance of protein homeostasis by removing misfolded or damaged proteins, which could jeopardize the healthy cellular functions, and by removing proteins no longer need. Therefore, 26S proteasome is involved in numerous cellular processes, including cell cycle progression, apoptosis, or DNA damage repair (Sone et al., 2004).

TABLE 7 | For each gene/protein in the HR pathway in (Orlic-Milacic, 2015) we selected the kinetic rates whose sensitivity index belongs to the 98th percentile of the sensitivity index distribution. The sensitity index is calculated using the **formula (1)**.

Gene	Kinetic rate to which it is highly sensitive
BLM	k12, k34, k107, k138, k222
BRCA2	k13, k58, k108, k139, k250
EME1	k24, k138, k140, k144, k227
MRE11	k14, k34, k109, k204, k223
MUS81	k24, k138, k140, k144, k227
NBN	k15, k32, k34, k60, k110
POLD1	k72, k73, k89, k90, k91
POLD2	k72, k73, k89, k90, k91
POLD3	k72, k89, k90, k91, k122
POLD4	k73, k89, k90, k91, k122
RAD50	k16, k61, k78, k111, k137
RAD51	–
RAD51B	k18, k63, k113, k140, k163
RAD51C	k19, k64, k81, k114, k157
RAD51D	k20, k65, k115, k141, k162
RAD52	k142, k168, k222, k227, k228
RPA1	k8, k29, k66, k83, k116
RPA2	k30, k33, k117, k144, k206
RPA3	k31, k68, k118, k222, k227
RPA4	k29, k227, k232, k233, k234
SEM1	k139, k145, k235, k236, k250
TOP3A	k21, k69, k86, k119, k248
TOP3B	k29, k66, k83, k116, k188
XRCC2	k22, k70, k120, k146, k149
XRCC3	k23, k121, k147, k166, k250

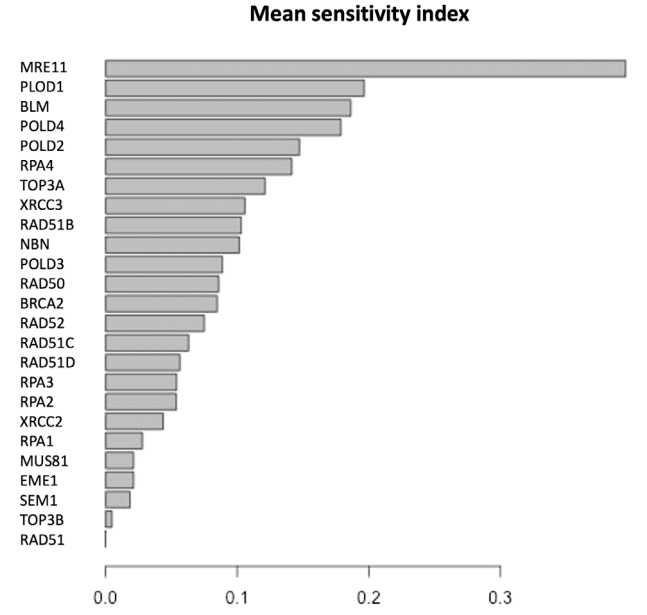


FIGURE 5 | Mean of the sensitivity index distributions for the proteins in HR network (Orlic-Milacic, 2015). These results refer to simulation in the time interval [0, 10] a.u., and initial values of the proteins randomly sampled in the range [1, 100] a. u. and kinetics rates values sampled in the interval [0, 0.01].

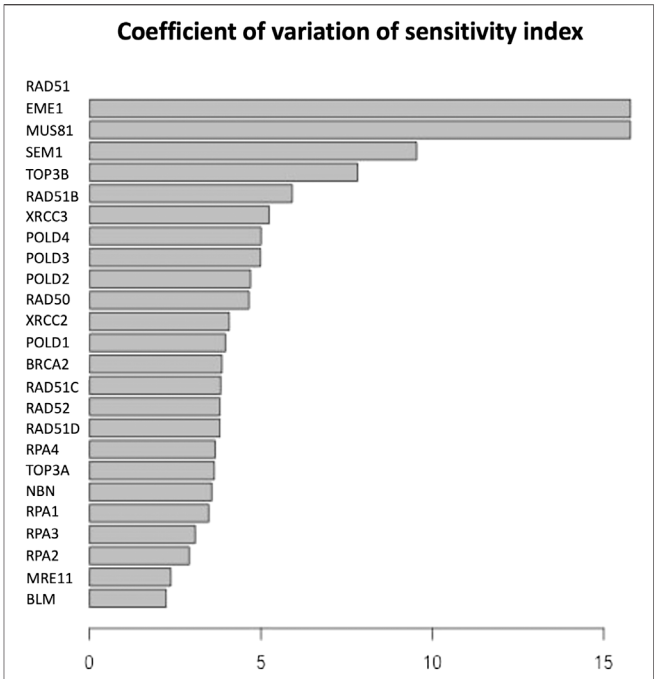


FIGURE 6 | Coefficient of variation of the sensitivity index distributions for the proteins in HR network Orlic-Milacic, (2015). These results refer to simulation in the time interval [0, 10] a.u., and initial values of the proteins randomly sampled in the range [1, 100] a. u. and kinetics rates values sampled in the interval [0, 0.01].

SEM1 was found also as a subunit in experiments of affinity purification of the yeast 19S proteasome, and its human homolog, DSS1, was found to copurify with the human 19S proteasome (Krogan et al., 2004). DSS1 is associated with the tumour suppressor protein BRCA2 involved in DNA DSBs repair. The authors in (Krogan et al., 2004) proved that SEM1 is essential for efficient repair of an HO-generated yeast DSB using both HR and nonhomologous end joining (NHEJ) pathways. Moreover, they showed that deletion of SEM1 contributes to cause defects in (synthetic) growth and hypersensitivity to genotoxins when combined with mutations in certain well-established genes involved in the DNA DSB repair.

Similarly to SEM1, the result of a high clustering coefficient is also expected for RPA4, as RPA4 is also part of a complex (Keshav et al., 1995). RPA4 gene encodes a single-stranded DNA-binding protein that is a subunit of the replication protein A complex (GeneCard, 2022). Replication protein A is essential for DNA DSB repair and plays a crucial role in the activation of cell cycle checkpoints. As regards the RPA complex, we have already seen in the previous sections that the RPA complex controls DNA repair and DNA damage checkpoint activation as well. In particular, the network analysis shows that RPA1 highly scores by total degree and betweenness. These results reflect the fact that RPA1 is an active route of communication exchanges between various nodes in the network. RPA1 is part of the heterotrimeric replication protein A complex (RPA/RP-A). It stabilizes single-

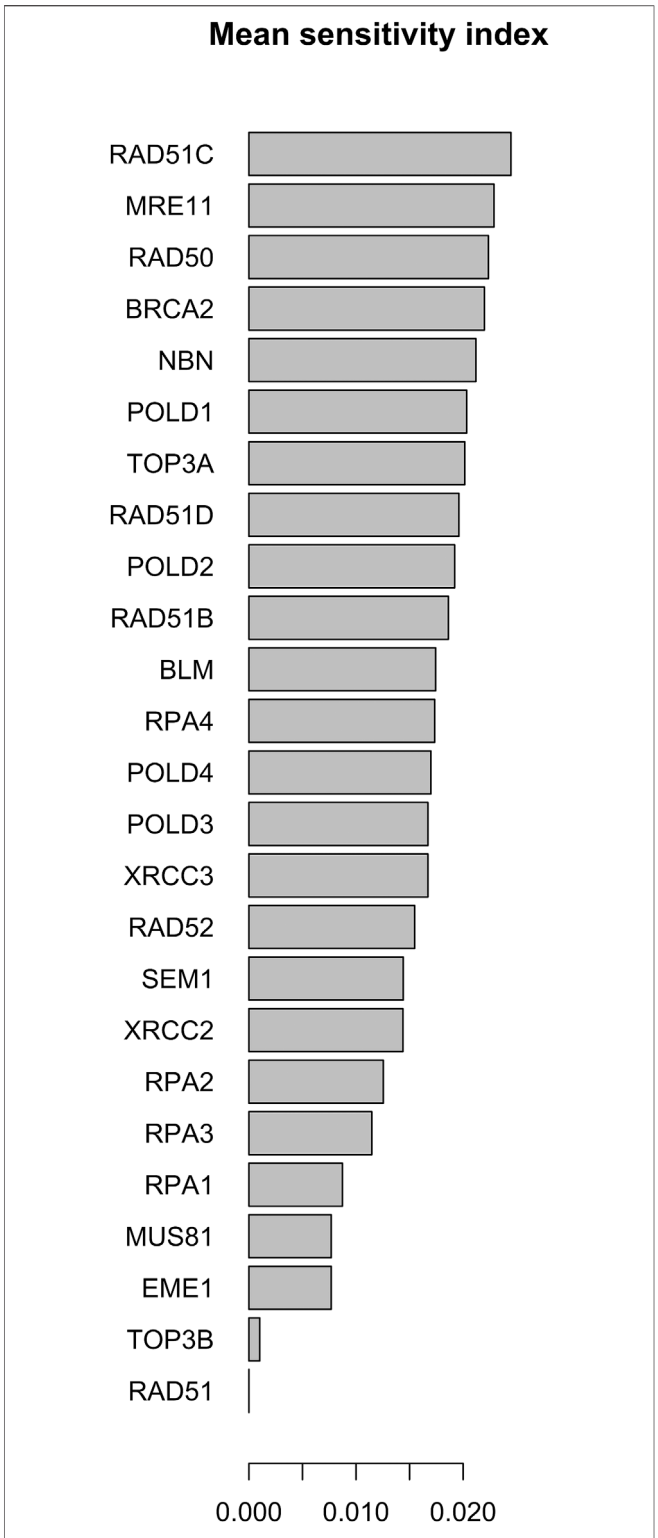


FIGURE 7 | Mean of the sensitivity index distributions for the proteins in HR network Orlic-Milacic, (2015). These results refer to simulation in the time interval [0, 1400] a.u., and initial values of the proteins randomly sampled in the range [18, 20] a. u. and kinetics rate values sample in [10^{-5}, 10^{-6}] a. u.

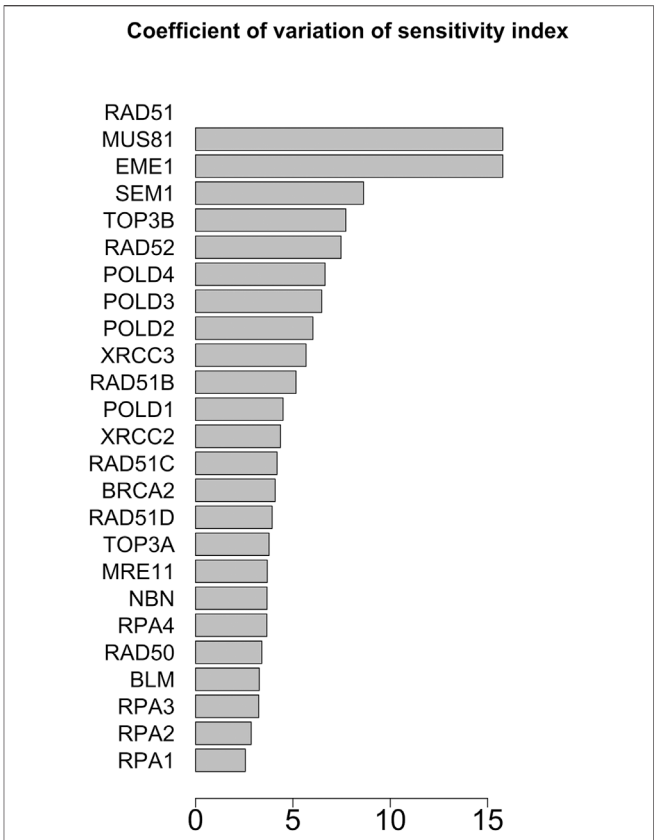


FIGURE 8 | Coefficient of variation of the sensitivity index distributions for the proteins in HR network (Orlic-Milacic, 2015). These results refer to simulation in the time interval [18, 20] a.u., and initial values of the proteins randomly sampled in the range [10^{-5}, 10^{-6}] a. u.

stranded DNA intermediates, that form during DNA replication or upon DNA stress In (Bass et al., 2016; Haahr et al., 2016; Human Protein Atlas, 2022). It prevents the reannealing of single-stranded DNA intermediates and recruits and activates different proteins and complexes forming part of DNA metabolism. Thereby, it is a key protein both in DNA replication and in the cellular response to DNA damage (Lin et al., 1998). RPA2 shows high score for in-degree and eigenvevtor centrality, meaning that it is interacting with protein also highly scoring by eigenvector centrality and degree (Hansen et al., 2020), and thence with proteins which have a great influence in the HR network. Indeed RPA 2 gene has been found highly expressed in low grade carcinomas and its expression has a gradual significant decrease from stage I to stage IV carcinomas. All the three subunits RPA1, RPA2, and RPA3, were more abundant (with statistical significance evidence) in lymph node negative and earlier stage (stage I and II) gastric carcinomas (Fourtziala et al., 2020). Finally, of particular interest and the fact that RPA3 ranks first in terms of centrality out-degree, hub-centrality and sub-graph centrality. Since subgraph centrality of a node is the number of subgraphs a node

TABLE 8 | Values of the centrality measures for the HR pathway in (Orlic-Milacic, 2015) merged with the FANCM pathway in (Pathways Commons, 2022). In bold, we marked the genes with the highest scores.

Protein	Total degree	In-degree	Out-degree	Hub centrality	Betweenness	Clustering coefficient	Eigenvector centrality	Vibrational centrality	Subgraph centrality	Information centrality
BLM	32	19	13	4.83E-01	117.0805289	0.72058824	0.83056114	**1.107752**	7408.39241	0.1204394
EME1	15	0	15	5.92E-01	0	0.48351648	0.28101096	9.68E-01	2727.280253	0.1730775
MRE11	26	16	10	3.71E-01	9.2926175	0.82051282	0.73943396	1.08E+00	6025.93013	0.1319644
MUS81	13	1	12	5.18E-01	0	0.53846154	0.25936455	9.56E-01	2292.062046	0.1868928
NBN	27	17	10	3.69E-01	9.2926175	0.82051282	0.76658977	1.08E+00	6025.93013	0.1296602
POLD1	26	9	17	6.82E-01	7.8121197	0.76470588	0.67199489	9.86E-01	7688.630996	0.1322165
POLD2	23	5	18	7.10E-01	5.8377525	0.76470588	0.57047643	9.81E-01	9071.417868	0.1400556
POLD3	23	8	15	6.44E-01	5.8377525	0.76470588	0.57047643	9.78E-01	5524.141817	0.1396592
POLD4	23	7	16	6.65E-01	5.8377525	0.76470588	0.57047643	9.80E-01	6516.909994	0.1397968
RAD50	27	13	14	5.81E-01	10.2729789	0.81318681	0.75778635	1.04E+00	**11437.382893**	0.1300547
RAD51	23	12	11	4.02E-01	30.3997662	0.65497076	0.4937734	9.17E-01	1403.106256	0.1386277
RAD51B	14	12	2	8.53E-02	0.3636364	0.82051282	0.31833841	1.01E+00	74.55065	0.1774293
RAD51C	21	9	12	4.93E-01	1.6614219	0.81904762	0.47562207	9.12E-01	1911.911354	0.1453357
RAD51D	16	11	5	2.13E-01	0.5127787	0.81904762	0.3597479	9.38E-01	282.219746	0.1656386
RAD52	4	2	2	1.00E-01	0	**1**	0.12449815	9.63E-01	254.418337	0.3917785
RPA1	43	20	**23**	9.74E-01	215.4152627	0.49802372	0.91307865	9.47E-01	8620.183617	0.1074375
RPA2	41	22	19	8.45E-01	32.6964097	0.64210526	**1**	9.82E-01	8511.793141	0.1095086
RPA3	34	11	23	**1**	25.979693	0.64210526	0.81828849	9.98E-01	14877.81448	0.1182323
RPA4	3	0	3	1.27E-01	0	**1**	0.10543731	9.63E-01	700.745966	0.4815787
SEM1	3	1	2	9.07E-02	0	**1**	0.0682632	9.28E-01	64.063815	0.4810717
TOP3A	26	12	14	5.39E-01	43.5259509	0.73626374	0.70171805	1.04E+00	8739.804106	0.1322725
XRCC2	15	14	1	5.59E-02	0.8476272	0.82417582	0.3492084	1.03E+00	55.032467	0.1704502
XRCC3	13	12	1	4.65E-02	0.3333333	0.80769231	0.28356645	5.52E-01	0	0.1844967
BRCA1	2	0	2	7.17E-02	0	NA	0.02248201	9.53E-01	15.252907	0.6489791
CENPS	2	0	2	7.17E-02	0	NA	0.02248201	9.53E-01	15.252907	0.6489791
CENPX	2	0	2	7.17E-02	0	NA	0.02248201	9.53E-01	15.252907	0.6489791
EME2	2	0	2	7.17E-02	0	NA	0.02248201	9.53E-01	15.252907	0.6489791
FAAP100	4	2	2	7.17E-02	0	NA	0.04496402	1.00E+00	14.252907	0.3943335
FAAP20	2	0	2	7.17E-02	0	NA	0.02248201	9.53E-01	15.252907	0.6489791
FAAP24	2	0	2	7.17E-02	0	NA	0.02248201	9.53E-01	15.252907	0.6489791
FANCA	4	2	2	7.17E-02	0	NA	0.04496402	1.00E+00	14.252907	0.3943335
FANCB	4	2	2	7.17E-02	0	NA	0.04496402	1.00E+00	14.252907	0.3943335
FANCC	4	2	2	7.17E-02	0	NA	0.04496402	1.00E+00	14.252907	0.3943335
FANCD2	2	1	1	3.59E-02	0	NA	0.02248201	9.77E-01	3.563227	0.6504898
FANCE	4	2	2	7.17E-02	0	NA	0.04496402	1.00E+00	14.252907	0.3943335
FANCF	4	2	2	7.17E-02	0	NA	0.04496402	1.00E+00	14.252907	0.3943335
FANCG	4	2	2	7.17E-02	0	NA	0.04496402	1.00E+00	14.252907	0.3943335
FANCI	3	1	2	7.17E-02	0	NA	0.03372301	9.77E-01	14.252907	0.4827827
FANCL	4	2	2	7.17E-02	0	NA	0.04496402	1.00E+00	14.252907	0.3943335
FANCM	**66**	**44**	22	4.35E-16	**481**	0.01811594	0.29119966	5.13E-01	131.839388	0.0955692
RMI1	2	0	2	7.17E-02	0	NA	0.02248201	9.53E-01	15.252907	0.6489791
RMI2	2	0	2	7.17E-02	0	NA	0.02248201	9.53E-01	15.252907	0.6489791
UBE2T	3	2	1	3.59E-02	0	NA	0.03372301	1.00E+00	3.563227	0.4815848
BRCA2	16	16	0	1.09E-16	0	0.69230769	0.36151501	-4.08E-18	0	0.1645300
TOP3B	1	1	0	6.80E-18	0	NA	0.03524702	0.00E+00	0	**1.1169817**
HES1	2	2	0	1.36E-17	0	NA	0.02248201	0.00E+00	0	0.6438225
MAX	43	22	23	0.9743705	215.4152627	0.82417582	0.91307865	1.08361	14877.81448	0
SCORE										

participates in (weighted according to their size) (Estrada and Rodríguez-Velázquez, 2005), it means that RPA3 take part into a number of subgraphs of significant size relatively to the whole network size. From our analysis it results that RPA3 versus RPA1 and RPA2, although its roles are similar to those of RPA2, has thus a great influence on pathways of crucial importance more than on subset of unconnected nodes or single nodes.

The analysis also highlights RAD50 that is a component of the MRN complex The protein complex is involved in numerous enzymatic activities required for nonhomologous joining of DNA ends. It is protein is essential for DNA double-strand break repair, cell cycle checkpoint activation, telomere maintenance, and meiotic recombination. (Carney et al., 1998; de Jager et al., 2001; Estrada and Ross, 2018; Bian et al., 2019; Beikzadeh and Latham, 2021; National library of Medicine, 2022). This role is reflected by th ehigh sungraph centrality that measures the centrality of a node by taking into account the number of subgraphs the node participates in. Specifically, the subgraph centrality of a node is the number of closed loops originating at

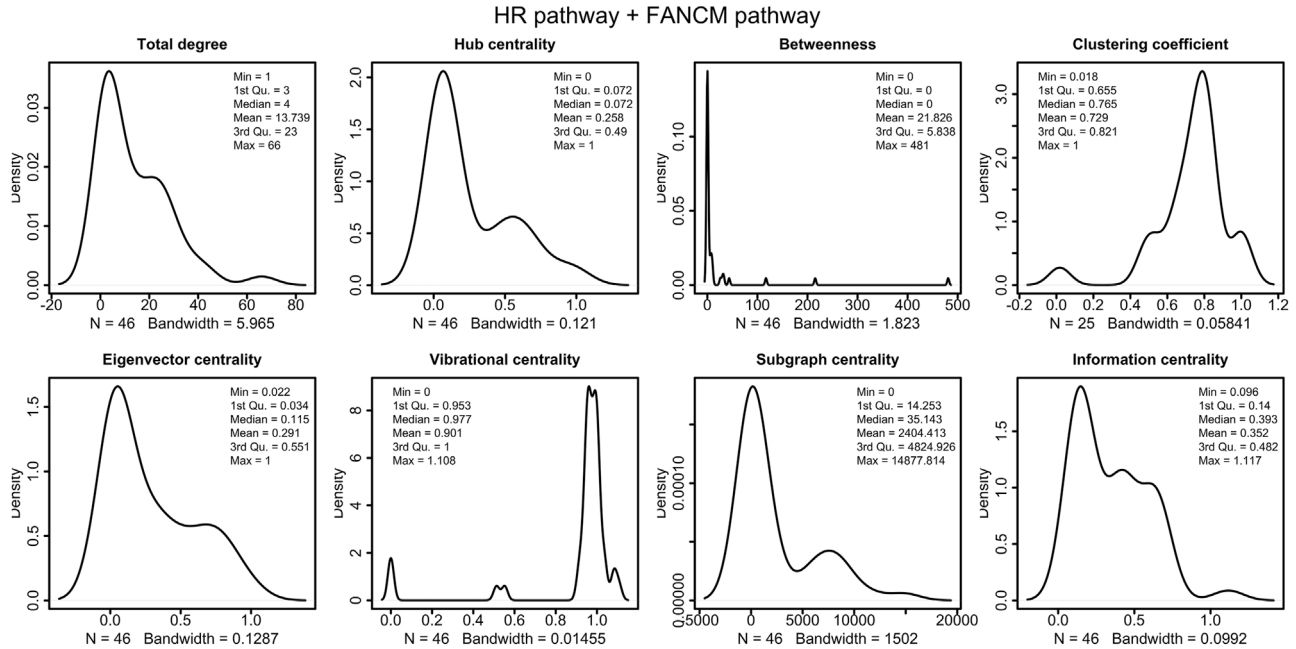


FIGURE 9 | Distributions of the centrality measures of HR pathway (Orlic-Milacic, 2015) merged with FANCM pathway (Pathways Commons, 2022). We observe that the majority of node have low betweenness, and high vibrational centrality.

the node, where longer loops are exponentially downweighted. Consequently, subgraph centrality measures how close a node is to the other nodes in the network.

The results of this analysis reveal a correspondence between the measure of centrality and the role of the protein. On the basis of this, when the role of the protein is known, this information can be used to work out the correctness of the computational analysis. When, on the other hand, the role of the protein is not known, knowledge of its centrality measurements can suggest the type or set of types of possible roles.

In **Tables 5**, **6** we show the rate equations of the HR network dynamics generated as the automatic translation of the network (see the script dynamics.R in the GitLab repository of NADS software). In file Simulations_of_Dynamics_HR_pathway.pdf provided in the Supplementary Material, we show the time evolution curves of each node of the HR network obtained as a solution of the equations.

The parameter sensitivity analysis was conducted by perturbing each parameter in the convergence range of the solution and yielded the results shown in **Table 7**; **Figures 5**, **6**. In **Table 7** we report the kinetic rate constants for which the sensitivity index belongs ot the 98th percentile of the sensitivity index distribution. They correspond to the most sensitive parameters, i.e. to the interactions whose alterations can significantly alter the dynamics of the network. To find out which interactions they refer to, the reader can refer to the **Supplementary Tables S1–S3**. **Figure 5** we shows that MRE11, followed by POLD1, BLM has the highest average sensitivity index. In **Figure 6** we show the coefficient of variation of each protein in the HR network. The coefficient of variation, being the ratio of the standard deviation to the mean, measures the extent of variability in relation to the mean of the population. The higher the coefficient of variation, the greater the dispersion. BLM, followed by MRE11 and RPA2 exhibits the lowest coefficient of variation of the sensitivity index. This mean that BLM, MRE11 and RPA2 have high sensitivity indices, and that the distribution of the sensitivity indices is well shaped around its mean, i.e. these protein exhibit almost the same sensitivity for all the parameters of the model. In this study we have therefore found that the BLM and RPA2 are sensitive nodes and that their sensitivity has two components: a topological sensitivity expressed by vibrational centrality, eigenvector centrality and clustering coefficient, and a dynamic sensitivity expressed by the parameter sensitivity index.

As with the BLM and RPA2 proteins, sensitivity analysis also highlights the MRE11 protein, which is highly sensitive to kinetic parameters, and its vibrational centrality is 1.0587524, very close to the maximum value exhibited by BLM (see **Table 4**). Its eigenvector centrality is 0.7397 which, although not the maximum, is very close to it (see **Table 4**). Indeed, MRE11 is an integral part of the protein complex of RAD50-MRE11A-NBS1 known as the MRN complex (Porras, 2014; Shibata et al., 2014; Mukherjee et al., 2019). It plays a key role in homologous recombination, and it is generally believed that MRE11 initiates double-strand breaks resection. In particular, the authors show that the loss of MRE11 reduces the efficiency of homologous recombination in human TK6 cells without affecting double-strand breaks resection, indicating a role for MRE11 in homologous recombination also at a post-resection step.

The high value of the eigenvector centrality fork BLM, RPA2 and MRE11 confirms the crucial role of these proteins in the network and expresses the fact that they are pointed by

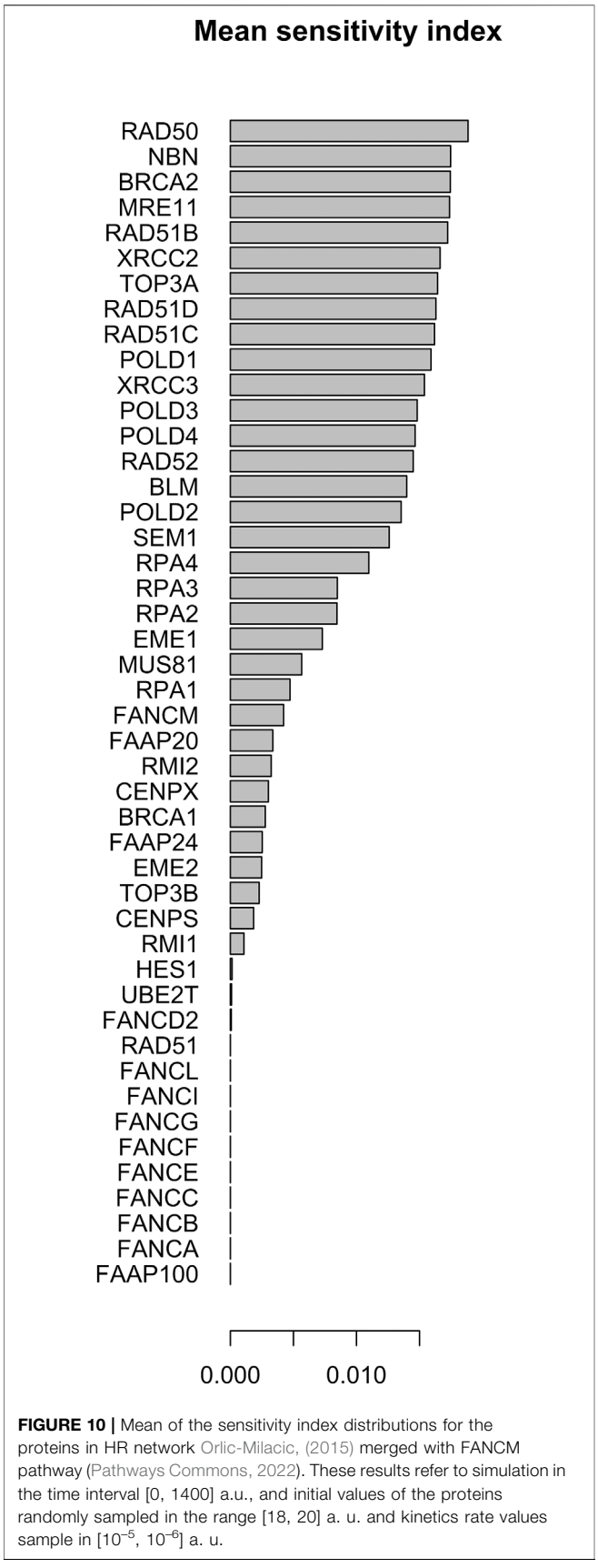


FIGURE 10 | Mean of the sensitivity index distributions for the proteins in HR network Orlic-Milacic, (2015) merged with FANCM pathway (Pathways Commons, 2022). These results refer to simulation in the time interval [0, 1400] a.u., and initial values of the proteins randomly sampled in the range [18, 20] a. u. and kinetics rate values sample in $[10^{-5}, 10^{-6}]$ a. u.

nodes that have a high value of the eigenvector centrality too. Indeed, if a node is pointed to by many nodes (which also have high eigenvector centrality) then that node will have high eigenvector centrality (Fletcher and Wennekers, 2018). The high sensitivity to the parameters characterising the dynamics of the interactions between these proteins and the partners pointing to them indicates the great influence that these partner nodes have on these proteins. Interestingly, RAD51 does not result sensitive to any parameter. The RAD51 encodes a protein that is essential for repairing damaged DNA. Recent findings have indicated RAD51 protein is overexpressed in a variety of tumours Chen et al. (2017). The overexpression of RAD51 causes improper and hyper-recombination, and thus contributes to genomic instability and

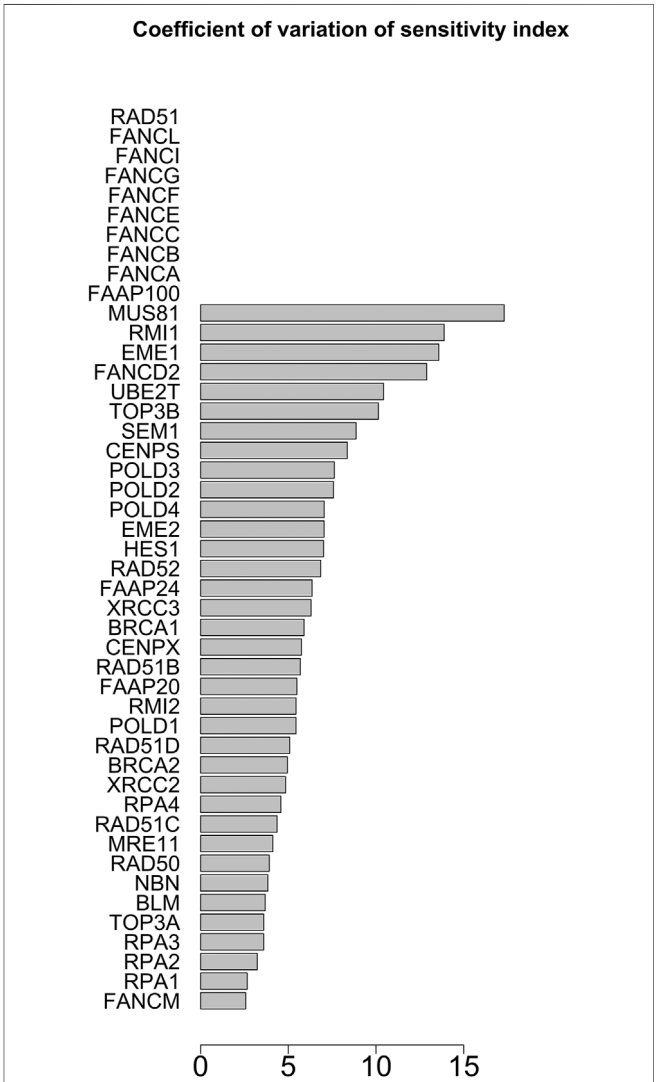


FIGURE 11 | Coefficient of variation of the sensitivity index distributions for the proteins in HR network (Orlic-Milacic, 2015) merged with FANCM pathway (Pathways Commons, 2022). These results refer to simulation in the time interval [18, 20] a.u., and initial values of the proteins randomly sampled in the range $[10^{-5}, 10^{-6}]$ a. u.

genetic diversity. Genomic instability might, in turn, drive regular cells towards neoplastic transformation or further contributes to cancer metastatic progression (Chen et al., 2007). The RAD51 protein binds to the DNA at the site of a break and encapsulates it in a protein sheath, initiating the repair process MedlinePlus (2022); Uniprot (2022). RAD51 protein interacts with BRCA1 and BRCA2, to fix damaged DNA. The BRCA2 protein regulates the activity of the RAD51 protein by transporting it in the nucleus to sites of DNA damage. Although the interaction between the BRCA1 protein and the RAD51 protein has still to be elucidate, research suggests that BRCA1 may also activate RAD51 in response to DNA damage (Cousineau et al., 2005; Chappell et al., 2016). The result of the sensitivity analysis found seems to contradict the important role of this protein in these interactions. Indeed, for example, one might expect a high sensitivity of RAD51 to the k139 due to its interaction with BRCA2 (see **Supplementary Table S2**). One explanation for this contradiction could be that since the mechanisms of interaction of RAD51 with these proteins are not fully known, the model used in this study could be an oversimplification of the interaction of RAD51 with its partners. If more accurate models in the future confirm the low sensitivity of RAD51 to the parameters of the rate equations describing the dynamics of the network, it will be necessary to investigate the physical and biological characteristics that make it so stable to perturbations. The fact that RAD51 has a low value of vibrational centrality in this study is a factor in favour of the possible confirmation of this case.

In **Figures 7**, **8** we report the results of the sensitivity analysis obtained selecting different ranges of initial conditions and parameters values. The plots highlights RAD51C, MRE11, RAD50 and BRCA2 as the most sensitive nodes to the parameters. We comment in the Section Remarks the expected differences and similarities in the results of sensitivity analysis when we change the intervals of initial conditions and parameters.

4.1 Analysis of the HR Pathway Merged With FANCM Pathway

We repeated the analysis on the HR's network extended by adding the pathways of FANCM gene (Fanconi Anaemia Group M Protein), obtained from Pathways Commons (Pathways Commons, 2022) given its important role in genome duplication, repair mechanisms and its involvement in the development of Fanconi anaemia, which several studies report to be a syndrome related to cancer predisposition (Deans and West, 2009; Xue et al., 2015; Bhattacharjee and Nandi, 2017; Pan et al., 2017; Wang et al., 2018). Finally, a recent study of Panday et al. reports that FANCM regulates repair pathway choice at stalled replication forks (Ling et al., 2016; Panday et al., 2021). FANCM and BLM have a similar role and cooperatively act in the DNA repair mechanisms (Panday et al., 2021), and through this analysis we want to investigate on this similarity.

The new network including HR and FANCM pathways is made up of 46 nodes and 316 edges. The left part of **Figure 4** shows the FANCM pathway and its connection with HR network. As reported in **Table 8**, the calculation of centrality measures led to the following results:

- BLM has the highest score in vibrational centrality
- FANCM has the highest score in total degree, in degree and betweenness
- RAD50 scores first for subgraph centrality
- RAD52 scores first for clustering coefficient
- RPA1 scores first for out-degree
- RPA2 scores first for eigenvector centrality
- RPA3 scores first for hub centrality
- RPA4 scores first for clustering coefficient
- SEM1 scores first for clustering coefficient.

The distribution of the centrality measures on the entire network is shown in **Figure 9**. These results not only re-emphasise as central the genes/proteins already identified in the analysis of the HR network alone in (Orlic-Milacic, 2015), but also highlight the central role of FANCM and as a node of particular relevance due to their high in-degree and high betweenness. By assigning an initial quantity between 18 and 20 (expressed in arbitrary units) to each node, the solution of the system of 46 differential equations converges for values of rate constants in the range 10^{-4} and 10^{-5} a.u. In file Simulations_of_Dynamics_HR_FANCM_pathway.pdf provided in the Supplementary Material, we show the time evolution curves (obtained as a solution of the equations) of each node of the HR pathway merged with FANCM pathway. The parameter sensitivity analysis was conducted by perturbing each parameter in the convergence range of the solution and yielded the results shown in **Figures 10**, **11**. We found that the nodes most sensitive to the parameters are RAD50, NBN, BRCA2, MRE11 and RAD51B. Compared to what was obtained in the analysis of the HR network alone, we find here that BLM is no longer at the top in terms of parameter sensitivity while still maintaining a central role in terms of vibration centrality.

4.2 Remarks

The method provides a range of values for the rate constants within which the solution to the problem exists at the given set of initial values for the node concentration/abundance. We can interpret this range as that of the 'most probable' range of values if.

- the initial conditions are known
- the analysed network does not exclude important interactions occurring *in vivo* and if the system is subjected to the conditions of the real system *in vitro* and *in vivo*. The network considered is only an extract of a much more complex network (still not completely known) that operates *in vivo* and in interaction with environmental factors.

We also note that we do not dispose of experimental time curves that can be used to calibrate the model. Calibration from

experimental data rather than sensitivity analysis would be the most appropriate method to use to obtain an estimated (even interval) parameter estimate. In the absence of both experimental data. We agree with the Reviewer that sensitivity analysis provides information on the minimum set of parameters to be inferred from experimental data, since the parameters to which the model is less sensitive are less influential.

Finally, we also observe that having fixed a set of initial values for the concentrations/abundances of the network components, more than one set of intervals for the rate constants could guarantee the convergence of the numerical method of solving the system of differential equations. It is also true that by changing the initial values of the concentrations, the range of values of the rate constants for which the system converges could change. The results that we report in this new version of the manuscript show, for example, that if the range of the initial concentration values is between 0 and 100, the numerical solution is found for rate constant values between 0 and 0.01, whereas if the range of the initial concentration values is a few tens, the numerical solution is found for rate constant values between 10^{-6} and 106–5. A reduction of 10 in the order of magnitude of the initial concentration values thus corresponds to a reduction of 10^{-3} in the order of magnitude of the rate constants. This is an indication that the system is underdetermined, and in fact consists of more parameters than the number of variables and in the complete absence of experimental data. All this also shows that calibrating the model in the light of experimental data is the best way to hope for a set of 'probable' values of the rate constants.

The work shown in this study therefore does not so much emphasise the numerical solutions, but, through a mathematical model, wants to test the susceptibility of the network components to the parameters and wants to integrate it with the role that the network components have (estimated by the centrality measurements).

5 CONCLUSION

This report presented an application of network analysis and mathematical modelling to the double-strand break repair pathway homologous recombination repair (HR). The complexity of the network of repair mechanisms itself, as well as the complexity of its interactions with the surrounding environment (Li et al., 2009; Chatterjee and Walker, 2017; Kusakabe et al., 2019; Poetsch, 2020; Roux et al., 2021), and the mutations of its components make its mathematical modelling particularly difficult, especially when based on rate equations. It is therefore of great necessity to have a tool that can implement these two important steps:

1. network analysis including standard centrality measures and new measures to quantify the robustness and responsiveness of the network to stimuli and stresses not dependent on the network topology
2. automatic construction of a mathematical model, for its analysis, and which allows to carry out refinements and modifications, when new data and new experimental knowledge make it necessary.

The implementation of these step is an innovative perspective for the analysis of DNA repair mechanisms. So far in the literature, there are many studies and analyses focused on the genetic and genomic aspects of the pathways, but studies on the mathematical modelling of its dynamics are absent. Our study therefore aims to fill this gap, since the mechanisms of DNA repair are governed by genes, proteins and pathways in continuous communication with the environment. For this reason, the analysis of the dynamics of the network is particularly useful, since it can quantify the vulnerability of the network and the modes of response to stimuli and exogenous stress. To the best of our knowledge there are no schemes for translating a graph associated with a biological network into a set of dynamic equations. The main reason for this is that there is no unambiguously defined semantics of a graphic representation of a biological network, i.e. there is no unambiguous definition of the graphic symbolism in terms of the mathematical equation describing the interaction indicated by that symbolism. The translation model we propose in this study is a basic model that describes the interactions indicated by the graph with linear first-order differential equations in the parameters. The code that implements this translation, however, gives the user the possibility to modify the model where he/she deems it appropriate in the light of available biological knowledge, or in the case he/she like to generate new hypothetical scenarios. We believe that the availability of a tool such as NADS can support the investigation of such a complex network that is subject to continuous interaction with external agents, not only to understand its dynamics, but also to predict its evolution and identify points of vulnerability for the benefit of the medical applications that this research may provide.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors. the software used int eh analysis is available at https://gitlab.inf.unibz.it/Paola.Lecca/network-analyzer-and-dynamics-simulator.

AUTHOR CONTRIBUTIONS

PL designed the methods and implemented the R codes for the analysis of the biological network, the automatic translation of the hypergraph into a set of rate equations, the numerical solution of the model and the sensitivity analysis. AI-N provided the biological background and provided the interpretation of the results obtained, in the light of an extensive literature search. Both authors contributed to the conceptualisation of the paper, its writing and revision.

FUNDING

This study has been supported by the COMPANET RTD 2019 Project funds, issued to PL (Principal Investigator) by the Faculty of Computer Science, Free University of Bozen-Bolzano, Italy.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.878148/full#supplementary-material

REFERENCES

Alberts, B. (2015). *Molecular biology of the Cell*. New York, NY: Garland Science, Taylor and Francis Group.

Ashtiani, M., Salehzadeh-Yazdi, A., Razaghi-Moghadam, Z., Hennig, H., Wolkenhauer, O., Mirzaie, M., et al. (2018). A Systematic Survey of Centrality Measures for Protein-Protein Interaction Networks. *BMC Syst. Biol.* 12, 80. doi:10.1186/s12918-018-0598-2

Bass, T. E., Luzwick, J. W., Kavanaugh, G., Carroll, C., Dungrawala, H., Glick, G. G., et al. (2016). ETAA1 Acts at Stalled Replication Forks to Maintain Genome Integrity. *Nat. Cell Biol.* 18, 1185–1195. doi:10.1038/ncb3415

Beikzadeh, M., and Latham, M. P. (2021). The Dynamic Nature of the Mre11-Rad50 DNA Break Repair Complex. *Prog. Biophysics Mol. Biol.* 163, 14–22. doi:10.1016/j.pbiomolbio.2020.10.007

Bennett, C. B., Lewis, A. L., Baldwin, K. K., and Resnick, M. A. (1993). Lethality Induced by a Single Site-specific Double-Strand Break in a Dispensable Yeast Plasmid. *Proc. Natl. Acad. Sci. U.S.A.* 90, 5613–5617. doi:10.1073/pnas.90.12.5613

Bhattacharjee, S., and Nandi, S. (2017). DNA Damage Response and Cancer Therapeutics through the Lens of the Fanconi Anemia DNA Repair Pathway. *Cell Commun. Signal* 15, 41. doi:10.1186/s12964-017-0195-9

Bian, L., Meng, Y., Zhang, M., and Li, D. (2019). MRE11-RAD50-NBS1 Complex Alterations and DNA Damage Response: Implications for Cancer Treatment. *Mol. Cancer* 18, 169. doi:10.1186/s12943-019-1100-5

Bouwman, P., and Jonkers, J. (2012). The Effects of Deregulated DNA Damage Signalling on Cancer Chemotherapy Response and Resistance. *Nat. Rev. Cancer* 12, 587–598. doi:10.1038/nrc3342

Branzei, D., and Foiani, M. (2008). Regulation of DNA Repair throughout the Cell Cycle. *Nat. Rev. Mol. Cell Biol.* 9, 297–308. doi:10.1038/nrm2351

Brosey, C. A., Yan, C., Tsutakawa, S. E., Heller, W. T., Rambo, R. P., and Tainer, J. A. (2013). A new structural framework for integrating replication protein A into DNA processing machinery. *Nucleic Acids Res.* 41 (4), 2313–2327. doi:10.1093/nar/gks1332

Bugreev, D. V., Mazina, O. M., and Mazin, A. V. (2006). Rad54 Protein Promotes Branch Migration of Holliday Junctions. *Nature* 442, 590–593. doi:10.1038/nature04889

Çağlayan, M., and Wilson, S. H. (2015). Oxidant and Environmental Toxicant-Induced Effects Compromise DNA Ligation during Base Excision DNA Repair. *DNA Repair* 35, 85–89. doi:10.1016/j.dnarep.2015.09.010

Carney, J. P., Maser, R. S., Olivares, H., Davis, E. M., Le Beau, M., Yates, J. R., 3rd, et al. (1998). The Hmre11/hrad50 Protein Complex and Nijmegen Breakage Syndrome: Linkage of Double-Strand Break Repair to the Cellular DNA Damage Response. *Cell* 93, 477–486. doi:10.1016/s0092-8674(00)81175-7

Carusillo, A., and Mussolino, C. (2020). DNA Damage: From Threat to Treatment. *Cells* 9, 1665. doi:10.3390/cells9071665

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., et al. (2010). Pathway Commons, a Web Resource for Biological Pathway Data. *Nucleic Acids Res.* 39, D685–D690. doi:10.1093/nar/gkq1039

Cerbinskaite, A., Mukhopadhyay, A., Plummer, E. R., Curtin, N. J., and Edmondson, R. J. (2012). Defective Homologous Recombination in Human Cancers. *Cancer Treat. Rev.* 38, 89–100. doi:10.1016/j.ctrv.2011.04.015

Chappell, W. H., Gautam, D., Ok, S. T., Johnson, B. A., Anacker, D. C., and Moody, C. A. (2016). Homologous Recombination Repair Factors Rad51 and BRCA1 Are Necessary for Productive Replication of Human Papillomavirus 31. *J. Virol.* 90, 2639–2652. doi:10.1128/jvi.02495-15

Chatterjee, N., and Walker, G. C. (2017). Mechanisms of DNA Damage, Repair, and Mutagenesis. *Environ. Mol. Mutagen.* 58, 235–263. doi:10.1002/em.22087

Chen, C.-Y., Ping, Y.-H., Lee, H.-C., Chen, K.-H., Lee, Y.-M., Chan, Y.-J., et al. (2007). Open Reading Frame 8a of the Human Severe Acute Respiratory Syndrome Coronavirus Not Only Promotes Viral Replication but Also Induces Apoptosis. *J. Infect. Dis.* 196, 405–415. doi:10.1086/519166

Chen, Q., Cai, D., Li, M., and Wu, X. (2017). The Homologous Recombination Protein RAD51 Is a Promising Therapeutic Target for Cervical Carcinoma. *Oncol. Rep.* 38, 767–774. doi:10.3892/or.2017.5724

Chu, W. K., and Hickson, I. D. (2009). RecQ Helicases: Multifunctional Genome Caretakers. *Nat. Rev. Cancer* 9, 644–654. doi:10.1038/nrc2682

Cooper, G. (2000). *The Cell : A Molecular Approach*. Washington, D.C. Sunderland, Mass: ASM Press Sinauer Associates.

Cousineau, I., Abaji, C., and Belmaaza, A. (2005). BRCA1 Regulates RAD51 Function in Response to DNA Damage and Suppresses Spontaneous Sister Chromatid Replication Slippage: Implications for Sister Chromatid Cohesion, Genome Stability, and Carcinogenesis. *Cancer Res.* 65, 11384–11391. doi:10.1158/0008-5472.can-05-2156

De Bont, R. (2004). Endogenous DNA Damage in Humans: a Review of Quantitative Data. *Mutagenesis* 19, 169–185. doi:10.1093/mutage/geh025

de Jager, M., van Noort, J., van Gent, D. C., Dekker, C., Kanaar, R., and Wyman, C. (2001). Human Rad50/Mre11 Is a Flexible Complex that Can Tether DNA Ends. *Mol. Cell* 8, 1129–1135. doi:10.1016/s1097-2765(01)00381-1

Deans, A. J., and West, S. C. (2009). FANCM Connects the Genome Instability Disorders Bloom's Syndrome and Fanconi Anemia. *Mol. Cell* 36, 943–953. doi:10.1016/j.molcel.2009.12.006

Dietlein, F., and Reinhardt, H. C. (2014). Molecular Pathways: Exploiting Tumor-specific Molecular Defects in DNA Repair Pathways for Precision Cancer Therapy. *Clin. Cancer Res.* 20, 5882–5887. doi:10.1158/1078-0432.ccr-14-1165

Estrada, E., and Hatano, N. (2010). A Vibrational Approach to Node Centrality and Vulnerability in Complex Networks. *Phys. A Stat. Mech. its Appl.* 389, 3648–3660. doi:10.1016/j.physa.2010.03.030

Estrada, E., and Rodríguez-Velázquez, J. A. (2005). Subgraph Centrality in Complex Networks. *Phys. Rev. E* 71, 056103. doi:10.1103/PhysRevE.71.056103

Estrada, E., and Ross, G. J. (2018). Centralities in Simplicial Complexes. Applications to Protein Interaction Networks. *J. Theor. Biol.* 438, 46–60. doi:10.1016/j.jtbi.2017.11.003

Estrada, E. (2011). *The Structure of Complex Networks*. New York, United States: Oxford University Press.

Fanning, E. (2006). A Dynamic Model for Replication Protein a (RPA) Function in DNA Processing Pathways. *Nucleic Acids Res.* 34, 4126–4137. doi:10.1093/nar/gkl550

Feldkamp, M. D., Mason, A. C., Eichman, B. F., and Chazin, W. J. (2014). Structural Analysis of Replication Protein a Recruitment of the DNA Damage Response Protein SMARCAL1. *Biochemistry* 53, 3052–3061. doi:10.1021/bi500252w

Fletcher, J. M., and Wennekers, T. (2018). From Structure to Activity: Using Centrality Measures to Predict Neuronal Activity. *Int. J. Neur. Syst.* 28, 1750013. doi:10.1142/s0129065717500137

Foertsch, F., Kache, T., Drube, S., Biskup, C., Nasheuer, H. P., and Melle, C. (2019). Determination of the Number of RAD51 Molecules in Different Human Cell Lines. *Cell Cycle* 18, 3581–3588. doi:10.1080/15384101.2019.1691802

Fornito, A., Zalesky, A., and Bullmore, E. T. (2016). "Centrality and Hubs," in *Fundamentals of Brain Network Analysis* (San Diego: Academic Press), 137–161. doi:10.1016/B978-0-12-407908-3.00005-4

Fourtziala, E., Givalos, N., Alexakis, N., Griniatsos, J., Alevizopoulos, N., Kavantzas, N., et al. (2020). Replication Protein A (RPA1, RPA2 and RPA3) Expression in Gastric Cancer: Correlation with Clinicopathologic Parameters and Patients' Survival. *J. BUON* 25, 1482–1489.

Friedberg, E. C., Aguilera, A., Gellert, M., Hanawalt, P. C., Hays, J. B., Lehmann, A. R., et al. (2006). DNA Repair: From Molecular Mechanism to Human Disease. *DNA Repair* 5, 986–996. doi:10.1016/j.dnarep.2006.05.005

Friedberg, E. C., Walker, G. C., Siede, W., Wood, R. D., Schultz, R. A., and Ellenberger, T. (2005). *DNA Repair and Mutagenesis*. Washington, DC: ASM Press.

Ganai, R. A., and Johansson, E. (2016). DNA Replication-A Matter of Fidelity. *Mol. Cell* 62, 745–755. doi:10.1016/j.molcel.2016.05.003

[Dataset] GeneCard (2022). *GeneCards – the Human Gene Database*. Available at: https://www.genecards.org/.

Ghasemi, M., Seidkhani, H., Tamimi, F., Rahgozar, M., and Masoudi-Nejad, A. (2014). Centrality Measures in Biological Networks. *Curr. Bioinforma.* 9, 426–441. doi:10.2174/1574893611308660013

Ghosal, G., and Chen, J. (2013). DNA Damage Tolerance: a Double-Edged Sword Guarding the Genome. *Transl. Cancer Res.* 2, 107–129. doi:10.3978/j.issn.2218-676X.2013.04.01

Gravel, S., Chapman, J. R., Magill, C., and Jackson, S. P. (2008). DNA Helicases Sgs1 and BLM Promote DNA Double-Strand Break Resection. *Genes Dev.* 22, 2767–2772. doi:10.1101/gad.503108

Haahr, P., Hoffmann, S., Tollenaere, M. A. X., Ho, T., Toledo, L. I., Mann, M., et al. (2016). Activation of the ATR Kinase by the RPA-Binding Protein ETAA1. *Nat. Cell Biol.* 18, 1196–1207. doi:10.1038/ncb3422

Hansen, D. L., Shneiderman, B., Smith, M. A., and Himelboim, I. (2020). "Calculating and Visualizing Network Metrics," in *Analyzing Social Media Networks with NodeXL* (Elsevier), 79–94. doi:10.1016/b978-0-12-817756-3.00006-6

Hoeijmakers, J. H. J. (2001). DNA Repair Mechanisms. *Maturitas* 38, 17–22. doi:10.1016/s0378-5122(00)00188-2

Hoeijmakers, J. H. J. (2009). DNA Damage, Aging, and Cancer. *N. Engl. J. Med.* 361, 1475–1485. doi:10.1056/nejmra0804615

Hosoya, N., and Miyagawa, K. (2014). Targeting DNA Damage Response in Cancer Therapy. *Cancer Sci.* 105, 370–388. doi:10.1111/cas.12366

Huang, R., and Zhou, P.-K. (2021). DNA Damage Repair: Historical Perspectives, Mechanistic Pathways and Clinical Translation for Targeted Cancer Therapy. *Sig Transduct. Target Ther.* 6, 254. doi:10.1038/s41392-021-00648-7

Huertas, P. (2010). DNA Resection in Eukaryotes: Deciding How to Fix the Break. *Nat. Struct. Mol. Biol.* 17, 11–16. doi:10.1038/nsmb.1710

Huertas, P., and Jackson, S. P. (2009). Human CtIP Mediates Cell Cycle Control of DNA End Resection and Double Strand Break Repair. *J. Biol. Chem.* 284, 9558–9565. doi:10.1074/jbc.m808906200

[Dataset] Human Protein Atlas (2022). *Human Protein Atlas*.

Jackson, S. P., and Bartek, J. (2009). The DNA-Damage Response in Human Biology and Disease. *Nature* 461, 1071–1078. doi:10.1038/nature08467

Jalili, M., Salehzadeh-Yazdi, A., Gupta, S., Wolkenhauer, O., Yaghmaie, M., Resendis-Antonio, O., et al. (2016). Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks. *Front. Physiol.* 7, 375. doi:10.3389/fphys.2016.00375

Kaur, E., Agrawal, R., and Sengupta, S. (2021). Functions of BLM Helicase in Cells: Is it Acting like a Double-Edged Sword? *Front. Genet.* 12, 634789. doi:10.3389/fgene.2021.634789

Kawale, A. S., and Sung, P. (2020). Mechanism and Significance of Chromosome Damage Repair by Homologous Recombination. *Essays Biochem.* 64, 779–790. doi:10.1042/ebc20190093

Kennedy, R. D., and D'Andrea, A. D. (2006). DNA Repair Pathways in Clinical Practice: Lessons from Pediatric Cancer Susceptibility Syndromes. *J. Clin. Oncol.* 24, 3799–3808. doi:10.1200/jco.2005.05.4171

Keshav, K. F., Chen, C., and Dutta, A. (1995). Rpa4, a Homolog of the 34-kilodalton Subunit of the Replication Protein a Complex. *Mol. Cell Biol.* 15, 3119–3128. doi:10.1128/mcb.15.6.3119

Kim, C., Paulus, B. F., and Wold, M. S. (1994). Interactions of Human Replication Protein a with Oligonucleotides. *Biochemistry* 33, 14197–14206. doi:10.1021/bi00251a031

Koschützki, D., and Schreiber, F. (2008). Centrality Analysis Methods for Biological Networks and Their Application to Gene Regulatory Networks. *Gene Regul. Syst. Bio* 2, GRSB.S702. GRSB. doi:10.4137/grsb.s702

Kovalchuk, I. (2016). *Genome Stability*. San Diego, CA: Elsevier, 1–18.

Kowalczykowski, S. C. (2015). An Overview of the Molecular Mechanisms of Recombinational DNA Repair. *Cold Spring Harb. Perspect. Biol.* 7, a016410. doi:10.1101/cshperspect.a016410

Krogan, N. J., Lam, M. H. Y., Fillingham, J., Keogh, M.-C., Gebbia, M., Li, J., et al. (2004). Proteasome Involvement in the Repair of DNA Double-Strand Breaks. *Mol. Cell* 16, 1027–1034. doi:10.1016/j.molcel.2004.11.033

Kusakabe, M., Onishi, Y., Tada, H., Kurihara, F., Kusao, K., Furukawa, M., et al. (2019). Mechanism and Regulation of DNA Damage Recognition in Nucleotide Excision Repair. *Genes Environ* 41, 2. doi:10.1186/s41021-019-0119-6

Le, H. P., Ma, X., Vaquero, J., Brinkmeyer, M., Guo, F., Heyer, W.-D., et al. (2020). DSS1 and ssDNA Regulate Oligomerization of BRCA2. *Nucleic Acids Res.* 48, 7818–7833. doi:10.1093/nar/gkaa555

Lecca, P., Mura, I., Re, A., Barker, G. C., and Ihekwaba, A. E. (2016). Time Series Analysis of the bacillus Subtilis Sporulation Network Reveals Low Dimensional Chaotic Dynamics. *Front. Microbiol.* 7, 1760. doi:10.3389/fmicb.2016.01760

Lecca, P., and Re, A. (2019). *Theoretical Physics for Biological Systems*. London, England: CRC Press.

Li, L., di Guan, Y., sha Chen, X., ming Yang, J., and Cheng, Y. (2021). DNA Repair Pathways in Cancer Therapy and Resistance. *Front. Pharmacol.* 11, 629266. doi:10.3389/fphar.2020.629266

Li, Y., Marion, M.-J., Zipprich, J., Santella, R. M., Freyer, G., and Brandt-Rauf, P. W. (2009). Gene-Environment Interactions Between DNA Repair Polymorphisms and Exposure to the Carcinogen Vinyl Chloride. *Biomarkers* 14, 148–155. doi:10.1080/13547500902811266

Lin, Y.-L., Shivji, M. K. K., Chen, C., Kolodner, R., Wood, R. D., and Dutta, A. (1998). The Evolutionarily Conserved Zinc Finger Motif in the Largest Subunit of Human Replication Protein a Is Required for DNA Replication and Mismatch Repair but Not for Nucleotide Excision Repair. *J. Biol. Chem.* 273, 1453–1461. doi:10.1074/jbc.273.3.1453

Lindahl, T. (1993). Instability and Decay of the Primary Structure of DNA. *Nature* 362, 709–715. doi:10.1038/362709a0

Ling, C., Huang, J., Yan, Z., Li, Y., Ohzeki, M., Ishiai, M., et al. (2016). Bloom Syndrome Complex Promotes FANCM Recruitment to Stalled Replication Forks and Facilitates Both Repair and Traverse of DNA Interstrand Crosslinks. *Cell Discov.* 2, 16047. doi:10.1038/celldisc.2016.47

Liu, J., Doty, T., Gibson, B., and Heyer, W.-D. (2010). Human BRCA2 Protein Promotes RAD51 Filament Formation on RPA-Covered Single-Stranded DNA. *Nat. Struct. Mol. Biol.* 17, 1260–1262. doi:10.1038/nsmb.1904

Liu, T., and Huang, J. (2016). DNA End Resection: Facts and Mechanisms. *Genomics, Proteomics Bioinforma.* 14, 126–130. doi:10.1016/j.gpb.2016.05.002

Liu, Y. (2001). p53 Protein at the Hub of Cellular DNA Damage Response Pathways through Sequence-specific and Non-sequence-specific DNA Binding. *Carcinogenesis* 22, 851–860. doi:10.1093/carcin/22.6.851

Lok, B. H., and Powell, S. N. (2012). Molecular Pathways: Understanding the Role of Rad52 in Homologous Recombination for Therapeutic Advancement. *Clin. Cancer Res.* 18, 6400–6406. doi:10.1158/1078-0432.ccr-11-3150

Lu, H., and Davis, A. J. (2021). Human RecQ Helicases in DNA Double-Strand Break Repair. *Front. Cell Dev. Biol.* 9, 640755. doi:10.3389/fcell.2021.640755

Maréchal, A., and Zou, L. (2014). RPA-Coated Single-Stranded DNA as a Platform for Post-translational Modifications in the DNA Damage Response. *Cell Res.* 25, 9–23. doi:10.1038/cr.2014.147

Marsden, P. V. (2005). "Network Analysis," in *Encyclopedia of Social Measurement* (Elsevier), 819–825. doi:10.1016/b0-12-369398-5/00409-6

[Dataset] MedlinePlus (2022). MedlinePlus Data Base. Available at: https://medlineplus.gov/genetics/gene/rad51/(Accessed 02 10.2022)

Mukherjee, S., Abdisalaam, S., Bhattacharya, S., Srinivasan, K., Sinha, D., and Asaithamby, A. (2019). "Mechanistic Link between DNA Damage Sensing, Repairing and Signaling Factors and Immune Signaling," in *DNA Repair* (Elsevier), 297–324. doi:10.1016/bs.apcsb.2018.11.004

[Dataset] National library of Medicine (2022). *National Library of Medicine*. National Center for Biotechnology Information. Available at: https://www.nlm.nih.gov/.

Nguyen, G. H., Dexheimer, T. S., Rosenthal, A. S., Chu, W. K., Singh, D. K., Mosedale, G., et al. (2013). A Small Molecule Inhibitor of the BLM Helicase Modulates Chromosome Stability in Human Cells. *Chem. Biol.* 20, 55–62. doi:10.1016/j.chembiol.2012.10.016

Nimonkar, A. V., Özsoy, A. Z., Genschel, J., Modrich, P., and Kowalczykowski, S. C. (2008). Human Exonuclease 1 and BLM Helicase Interact to Resect DNA and Initiate DNA Repair. *Proc. Natl. Acad. Sci. U.S.A.* 105, 16906–16911. doi:10.1073/pnas.0809380105

[Dataset] Orlic-Milacic, M. (2015). HDR through Homologous Recombination (Hrr). Available at: https://apps.pathwaycommons.org/search?q=homologous%20recombinations&type=Pathway (Accessed 10 01.2021)

Pan, X., Drosopoulos, W. C., Sethi, L., Madireddy, A., Schildkraut, C. L., and Zhang, D. (2017). FANCM, BRCA1, and BLM Cooperatively Resolve the Replication Stress at the ALT Telomeres. *Proc. Natl. Acad. Sci. U.S.A.* 114, E5940. doi:10.1073/pnas.1708065114

Panday, A., Willis, N. A., Elango, R., Menghi, F., Duffey, E. E., Liu, E. T., et al. (2021). FANCM Regulates Repair Pathway Choice at Stalled Replication Forks. *Mol. Cell* 81, 2428–2444. e6. doi:10.1016/j.molcel.2021.03.044

[Dataset] Pathways Commons (2022). Pathways Commons. .Available at: https://apps.pathwaycommons.org/search?type=Pathway&q=FANCM

Poetsch, A. R. (2020). The Genomics of Oxidative DNA Damage, Repair, and Resulting Mutagenesis. *Comput. Struct. Biotechnol. J.* 18, 207–219. doi:10.1016/j.csbj.2019.12.013

Porras, O. (2014). "DNA Repair Defects," in *Stiehm's Immune Deficiencies*. Editors K. E. Sullivan and E. R. Stiehm (Amsterdam: Academic Press), 199–219. doi:10.1016/b978-0-12-405546-9.00007-8

[Dataset] R Studio (2022). R Studio. Available at: https://rstudio.com/.

[Dataset] Reed, S. H., and Waters, R. (2005). DNA Repair. *eLS*. doi:10.1038/npg.els.0005284

Ross, C. A., and Truant, R. (2016). A Unifying Mechanism in Neurodegeneration. *Nature* 541, 34–35. doi:10.1038/nature21107

Rossi, M. J., DiDomenico, S. F., Patel, M., and Mazin, A. V. (2021). RAD52: Paradigm of Synthetic Lethality and New Developments. *Front. Genet.* 12, 780293. doi:10.3389/fgene.2021.780293

Roux, P., Salort, D., and Xu, Z. (2021). Adaptation to DNA Damage as a Bet-Hedging Mechanism in a Fluctuating Environment. *R. Soc. open Sci.* 8, 210460. doi:10.1098/rsos.210460

Safran, M., Rosen, N., Twik, M., BarShir, R., Stein, T. I., Dahary, D., et al. (2021). "The GeneCards Suite," in *Practical Guide to Life Science Databases* (Springer Singapore), 27. doi:10.1007/978-981-16-5812-9_2

Sancar, A., Lindsey-Boltz, L. A., Ünsal-Kaçmaz, K., and Linn, S. (2004). Molecular Mechanisms of Mammalian DNA Repair and the DNA Damage Checkpoints. *Annu. Rev. Biochem.* 73, 39–85. doi:10.1146/annurev.biochem.73.011303.073723

Shibata, A., Moiani, D., Arvai, A. S., Perry, J., Harding, S. M., Genois, M.-M., et al. (2014). DNA Double-Strand Break Repair Pathway Choice Is Directed by Distinct MRE11 Nuclease Activities. *Mol. Cell* 53, 7–18. doi:10.1016/j.molcel.2013.11.003

Sone, T., Saeki, Y., Toh-e, A., and Yokosawa, H. (2004). Sem1p Is a Novel Subunit of the 26 S Proteasome from saccharomyces Cerevisiae. *J. Biol. Chem.* 279, 28807–28816. doi:10.1074/jbc.m403165200

Stefanovie, B., Hengel, S. R., Mlcouskova, J., Prochazkova, J., Spirek, M., Nikulenkov, F., et al. (2019). DSS1 Interacts with and Stimulates RAD52 to Promote the Repair of DSBs. *Nucleic Acids Res.* 48, 694–708. doi:10.1093/nar/gkz1052

Sugiyama, T., and Kantake, N. (2009). Dynamic Regulatory Interactions of Rad51, Rad52, and Replication Protein-A in Recombination Intermediates. *J. Mol. Biol.* 390, 45–55. doi:10.1016/j.jmb.2009.05.009

Sun, Y., McCorvie, T. J., Yates, L. A., and Zhang, X. (2019). Structural Basis of Homologous Recombination. *Cell. Mol. Life Sci.* 77, 3–18. doi:10.1007/s00018-019-03365-1

[Dataset] Uniprot (2022). Uniprot Database. Available at: https://www.uniprot.org/uniprot/Q06609 (Accessed 02 10.2022)

van Brabant, A. J., Ye, T., Sanz, M., German, J. L., Ellis, N. A., and Holloman, W. K. (2000). Binding and Melting of D-Loops by the Bloom Syndrome Helicase. *Biochemistry* 39, 14617–14625. doi:10.1021/bi0018640

Wang, H., Li, S., Oaks, J., Ren, J., Li, L., and Wu, X. (2018). The Concerted Roles of FANCM and Rad52 in the Protection of Common Fragile Sites. *Nat. Commun.* 9, 2791. doi:10.1038/s41467-018-05066-y

Wang, P., Lü, J., and Yu, X. (2014). Identification of Important Nodes in Directed Biological Networks: A Network Motif Approach. *Plos One* 9, e106132. doi:10.1371/journal.pone.0106132

Wolters, S., and Schumacher, B. (2013). Genome Maintenance and Transcription Integrity in Aging and Disease. *Front. Genet.* 4, 19. doi:10.3389/fgene.2013.00019

Wood, R. D. (2018). Fifty Years since DNA Repair Was Linked to Cancer. *Nature* 557, 648–649. doi:10.1038/d41586-018-05255-1

Wood, R. D., Mitchell, M., and Lindahl, T. (2005). Human DNA Repair Genes, 2005. *Mutat. Research/Fundamental Mol. Mech. Mutagen.* 577, 275–283. doi:10.1016/j.mrfmmm.2005.03.007

Xue, X., Sung, P., and Zhao, X. (2015). Functions and Regulation of the Multitasking FANCM Family of DNA Motor Proteins. *Genes Dev.* 29, 1777–1788. doi:10.1101/gad.266593.115

Yadav, S., Anbalagan, M., Baddoo, M., Chellamuthu, V. K., Mukhopadhyay, S., Woods, C., et al. (2020). Somatic Mutations in the DNA Repairome in Prostate Cancers in African Americans and Caucasians. *Oncogene* 39, 4299–4311. doi:10.1038/s41388-020-1280-x

Yang, Y., Dou, S.-X., Xu, Y.-N., Bazeille, N., Wang, P.-Y., Rigolet, P., et al. (2010). Kinetic Mechanism of DNA Unwinding by the BLM Helicase Core and Molecular Basis for its Low Processivity. *Biochemistry* 49, 656–668. doi:10.1021/bi901459c

Yousefzadeh, M., Henpita, C., Vyas, R., Soto-Palma, C., Robbins, P., and Niedernhofer, L. (2021). DNA Damage-How and Why We Age? *eLife* 10, e62852. doi:10.7554/eLife.62852

Yun, M. H., and Hiom, K. (2009). CtIP-BRCA1 Modulates the Choice of DNA Double-Strand-Break Repair Pathway throughout the Cell Cycle. *Nature* 459, 460–463. doi:10.1038/nature07955

Zhao, F., Kim, W., Kloeber, J. A., and Lou, Z. (2020). DNA End Resection and its Role in DNA Replication and DSB Repair Choice in Mammalian Cells. *Exp. Mol. Med.* 52, 1705–1714. doi:10.1038/s12276-020-00519-1

Zhao, W., Vaithiyalingam, S., San Filippo, J., Maranon, D. G., Jimenez-Sainz, J., Fontenay, G. V., et al. (2015). Promotion of BRCA2-dependent Homologous Recombination by DSS1 via RPA Targeting and DNA Mimicry. *Mol. Cell* 59, 176–187. doi:10.1016/j.molcel.2015.05.032

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Lecca and Ihekwaba-Ndibe. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership