



OPEN ACCESS

EDITED BY

Arijita Sarkar,
University of Southern California,
United States

REVIEWED BY

Danny Misiak,
Martin Luther University of Halle-
Wittenberg, Germany
Harsh Goel,
All India Institute of Medical Sciences,
India

*CORRESPONDENCE

Aritra Bose,
✉ a.bose@ibm.com

RECEIVED 03 November 2025

REVISED 26 January 2026

ACCEPTED 02 February 2026

PUBLISHED 04 March 2026

CITATION

Bose A, Platt DE, Rhrissorrakrai K, Burch M,
Guzmán-Sáenz A, Haiminen N and
Parida L (2026) Remics: a redescription-
based framework for multi-
omics analysis.
Front. Cell Dev. Biol. 14:1738010.
doi: 10.3389/fcell.2026.1738010

COPYRIGHT

© 2026 Bose, Platt, Rhrissorrakrai, Burch,
Guzmán-Sáenz, Haiminen and Parida.
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.

Remics: a redescription-based framework for multi-omics analysis

Aritra Bose^{1*}, Daniel E. Platt¹, Kahn Rhrissorrakrai¹, Myson Burch¹,
Aldo Guzmán-Sáenz¹, Niina Haiminen² and Laxmi Parida¹

¹IBM T.J. Watson Research Center, Yorktown Heights, NY, United States, ²DAIN Studios, Helsinki, Finland

Complex diseases such as cancer are characterized by their intricate etiology, arising from several molecular mechanisms that span multiple omic layers. To obtain insights on disease subtypes, associated biomarkers, and improve prognostic modeling, it is essential to integrate and interpret multi-omics data in a biologically meaningful way. We introduce *Remics*, a redescription-based framework for multi-omics integration inspired by higher-order statistical representations. *Remics* leverages higher-order cumulants to identify redescrptions, which are sets of multi-omics features that jointly capture equivalent biological variation across modalities. These feature groups are further analyzed through network representations, multi-omics risk scoring, and biomarker discovery to reveal molecular interactions underlying disease mechanisms. We applied *Remics* on simulated data as well as multi-omics data of six different cancer types from The Cancer Genome Atlas. We demonstrate that redescription-based integration uncovers functionally coherent cross-omics feature associations and compare them with state-of-the-art approaches. Our results highlight the potential of higher-order multi-omics statistical analysis to advance precision medicine through improved interpretability and discovery of novel molecular relationships.

KEYWORDS

biomarker discovery, data mining, disease prediction, genetic epidemiology, multi-omics, networks, statistics

1 Introduction

Since understanding the etiology and pathogenesis of a disease provides the necessary basis for targeted treatment and prevention, the modern challenge for complex disease genetics is to discover how biomarkers from different omics platforms, such as genome, transcriptome, proteome, metabolome, etc., can impact complex disease phenotypes in concert. It is through characterizing and leveraging their interactions that we can elucidate the molecular mechanisms of complex diseases by discovering the associated biomarkers and predict disease outcome and its subtypes. As multi-omics approaches often inherit the challenges from single-omic analysis, this integration is typically fraught with challenges ranging from platform diversity and intrinsic heterogeneity within single omics (Subramanian et al., 2020) to missing data across different omics leading to sparsity (Song et al., 2020), to the varying dimensions for each omics, e.g., genomics and transcriptomics data often being magnitudes larger than metabolomics or proteomics, etc.

Many multi-omics data integration and analysis tools exist which try to address the above problems and derive some means of actionable biological insight. Some integration strategies use feature selection to reduce dimensions or perform early integration of all the omic profiles together in one large matrix (Picard et al., 2021). Recently, deep learning methods have been implemented to reduce dimensionality of the multi-omics matrix after early integration to extract embeddings associated with disease outcome (Chaudhary et al., 2018), while other methods use regularized multiple kernel learning to reduce dimensionality (Speicher and Pfeifer, 2015). However, these

approaches often suffer from lack of interpretability. Network based methods have been extensively developed for multi-omics data, including fusion methods such as Similarity Network Fusion (Wang et al., 2014) and multi-layer networks (Lee et al., 2020). These network based approaches use clustering to build upon a patient similarity matrix often applied to partial data sets (Rappoport and Shamir, 2019) and do not provide interactions between multi-omics features for affected patients. Some methods create new common latent representations from multi-omics data with matrix factorizations (Lee and Seung, 1999; Argelaguet et al., 2018; Rohart et al., 2017). One such approach employs multi-omics factor analysis (MOFA), a generalization of principal component analysis (PCA), to integrate multi-omics in an unsupervised manner. However, most of these methods are based on linear multivariate methods that do not take higher-order interactions between the multi-omics variables into account and is targeted towards biomarker discovery and classifying disease subtypes or outcome rather than interactions between features.

Here, we propose Remics, a redescription-based multi-omics analysis tool, which performs redescription mining, leveraging higher-order correlations between different omics features to obtain ensemble meta-features that represents the redescription groups of the individual features from single-omic profiles. Redescription mining (Parida and Ramakrishnan, 2005) is a conceptual clustering method which redefines sets of objects from multiple datasets, situates knowledge gained from one dataset in the context of others, and harnesses high-level abstractions in the form of meta-features thereby recovering subtle cryptic features in the data and exposing novel patterns therein. These meta-features can be used for learning association between multi-omics biomarkers. Remics has two primary components for downstream analyses, after computing the redescription groups: (1) cumulant-based network analysis (CuNA), and (2) cumulant-based risk scores (CuRES). CuNA computes a network by projecting these higher-dimensional interactions and analyzes it to identify hidden interactions in the data, subsequently discovering biomarkers. It also provides an interactive visualizer, which can be used to investigate the network for motifs, clusters, and to identify interactions between multi-omics variables. CuRES computes a single-value estimate of risk of a trait or disease per individual from the redescription groups of multi-omics variables. We applied Remics to simulated data representing multi-omics with varying degrees of correlations among them. To demonstrate how Remics integrates multi-omics features into biologically informative redescription groups, predicts disease status, and discovers latent interactions between variables, we applied it on six cancer datasets with transcriptomic, epigenetic, microRNA (miRNA), and clinical data. We compared Remics with conceptually similar state-of-the-art multi-omics integration methods, such as SNF that uses a fused network of samples across multi-omics data to integrate them with fused variables (Wang et al., 2014), and MOFA, which uses factor analysis to infer a low-dimensional representation across multiple data modalities, capturing global sources of variability (Argelaguet et al., 2018). Multi-omics variables are often correlated at higher-order (any order greater than two) than pairwise, when integrating more than two single-omic datasets. Moreover, due to underlying biological similarities of samples, some outcomes are better explained with multi-way interactions between single-omic variables. Remics enables integration of multi-omics data leveraging higher-order interactions between the variables to provide an informative and interpretable framework of analyzing multi-omics data.

2 Materials and methods

2.1 Redescription groups

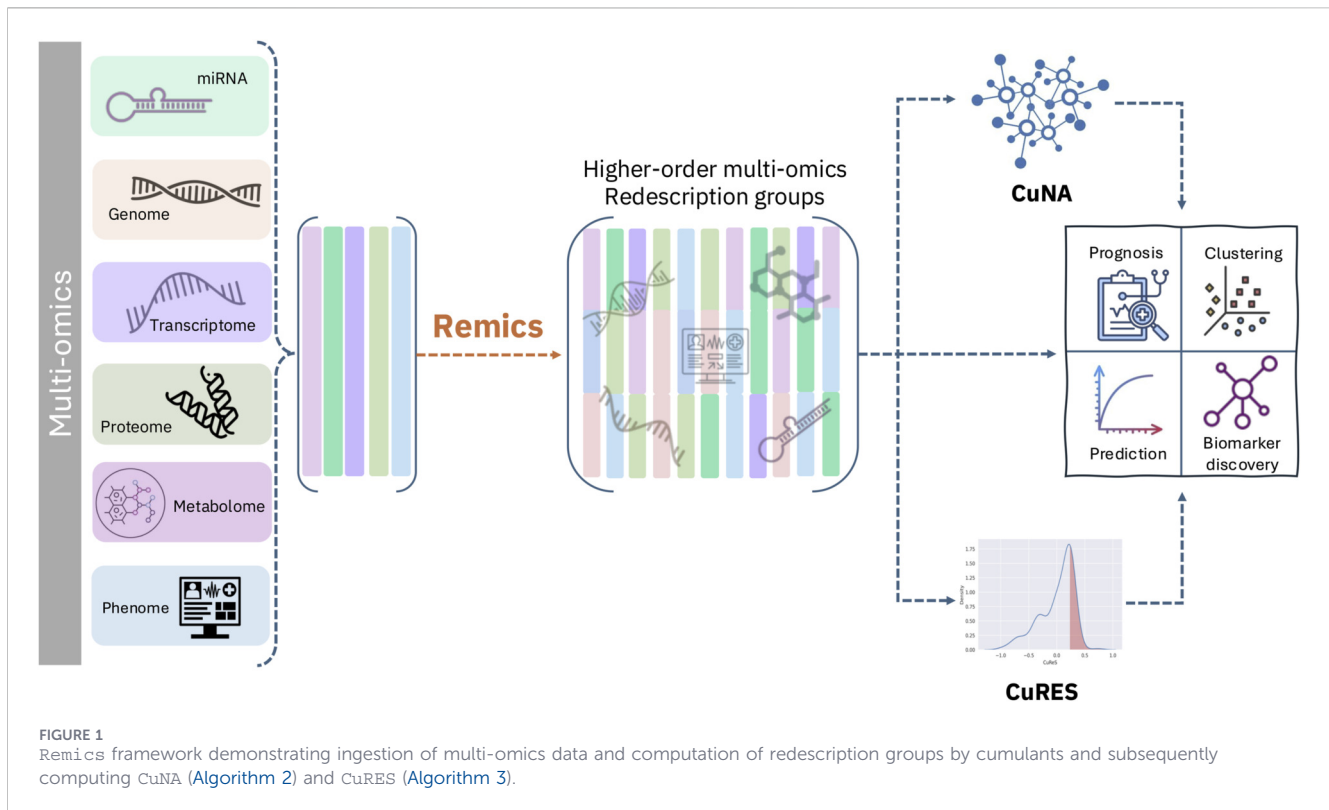
In a dataset with \mathcal{S} samples and \mathcal{F} features, $s \in \mathcal{S}$ are described by a list of features $f_i(s)$, $i \in \mathcal{F}$. Let there be a group \mathcal{G}_i associated with each feature, $f_i(s) \in \mathcal{G}_i$, where $\mathcal{G}_i \in \mathbb{R}$. Examples of features in \mathcal{F} can include genes, genomic markers, clinical variables, or other variables in multi-omics data. For a given $g_i \in \mathcal{G}_i$, the set of subjects that have that value is $f_i^{-1}(g_i) \subseteq \mathcal{S}$. These features can be binary or continuous variables, where continuous variables can be converted to binary according to a threshold, such as the mean $f_i(\mathcal{S})$, mapped to 1 if $f_i(s) \geq m(f_i(\mathcal{S}))$. Patterns in the data can be described in terms of conjunctions $i \wedge j$ for $i, j \in \mathcal{F}$ such that $f_{i \wedge j}^{-1}(g_i, g_j) = f_i^{-1}(g_i) \cap f_j^{-1}(g_j)$ for binary g_i, g_j . This definition can be extended to include not only i, j but also to any combinations of conjunctions subject to the logical algebra of \wedge (e.g., $(i \wedge j) \wedge (i \wedge k) = i \wedge j \wedge k$ for $i, j, k \in \mathcal{F}$ subject to values g_i, g_j, g_k). In other words, if there are three features that interact with each other, such as genes PARP1 and BRCA family of genes, which are strongly related in many cancers causing tumor cell death, this interaction will be represented as $f_{PARP1 \wedge BRCA1 \wedge BRCA2}^{-1}(PARP1 = 1, BRCA1 = 1, BRCA2 = 1)$. Such combinations of conjunctions i that have more or less members $f_i^{-1}(a)$ than expected by chance are called *patterns*. Binomial and other tests of the significance of patterns can be dominated by lower-order correlations among the variables in a pattern.

Definition 1: Two distinct patterns that yield the same subsets of subjects, e.g., $f_i^{-1}(g) = f_j^{-1}(g)$, are called “redescriptions”.

If conjunctions yield a form such as $A \cap B = B$, then it may be deduced that BA , and the conditions yielding A and B satisfy $b \Rightarrow a$ (see [Supplementary Note](#) for details). In other words, redescrptions can reveal logical relationships among features. Such relationships may reflect the underlying biological pathways reflected in these connected phenotype patterns. Therefore, each of these patterns i specifies a phenotype, which may be associated with genotypes or other omic data using standard methods. Redescription groups are often generated combinatorially by measuring significant associations among features. However, if those groups are selected from highly correlated variables, then it may be difficult to extract distinct interactions among these variables. We solve this problem by testing redescription groups using Fisher permutation test. This identifies whether one factor significantly affects the relationships between other factors (Karisani et al., 2022).

2.1.1 Redescriptions via cumulants

An intuitive way to find redescription groups is to compute cumulants (Parida and Ramakrishnan, 2005; Platt D. et al., 2024; Karisani et al., 2022; Bose et al., 2021; Bose et al., 2023; Bose et al., 2026), which identify logical relationships between the features defining the patterns. Multiple patterns capture the same set of samples, and thus find the redescription groups (Definition 1). Since most of the biomarkers in large biobanks are strongly correlated, we need to factor out those strong lower-order correlations from higher order associations marking distinct groups of individuals differentiating sub-types of the disease. Cumulants, in simple terms, are measures of the interaction of random



variables in a probability distribution. If $f(X)$ is a function of any random variable X with outcome $\{x_i\}$, then its expectation is given as $\mathbb{E}[f(X)] = \sum_i p_i f(x_i)$. The first order cumulant is defined as the mean of X , defined as $\sum_i p_i x_i$ and thus the higher order moments $\langle X^d \rangle = \sum_i p_i x_i^d$. This is formalized as a moment generating function,

$$M(\lambda) = \sum_{i=0}^{\infty} \frac{\lambda^i}{i!} \langle X^d \rangle$$

Thus, the moments of X can be generated as the terms of coefficients of the Taylor expansion of $M(\lambda)$. The cumulant generating function is nothing but the logarithm of $M(\lambda)$, defined as $\log M(\lambda)$. The first four orders of cumulants are known as mean, variance, skewness, and kurtosis (a detailed introduction to cumulants are in the [Supplementary Note](#)).

2.2 Remics

Remics, or redescription-based multi-omics analysis, computes redescription groups via cumulants, here calculated using the Julia package `Cumulants.jl`, to capture higher-order interactions between multi-omics variables using multidimensional tensors ([Domino et al., 2018](#)). Remics takes as input a real-valued matrix in which rows correspond to samples and columns represent molecular features from one or more omics layers (e.g., transcriptomic, epigenomic, proteomic, metabolomic, etc.). Multi-omics data may be integrated using early fusion by concatenating feature matrices across modalities, or using late fusion by supplying a preprocessed composite representation. While both strategies are supported, early fusion facilitates more direct interpretability of the resulting redescription groups. The input matrix is normalized prior to computing higher-

order cumulants; alternatively, users may provide appropriately normalized data. Remics makes no modality-specific modeling assumptions beyond this normalization step.

Input: A matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, where m is the number of individuals and n is the number of multi-omics features; order of computation, d

Output: Set \mathcal{C}' of statistically significant redescription groups, $|\mathcal{C}'| = k$; matrix $\mathbf{M} \in \mathbb{R}^{n \times k}$, with each sample having k -dimensional vectors representing redescription meta-features

1: $\tilde{\mathbf{A}} = \mathbf{A} - \mathbf{1}_m \mu^T$, where $\mu_j = \frac{1}{m} \sum_{i=1}^m \mathbf{A}_{ij}$, $j = 1, 2, \dots, n$

2: Compute G 's ([Supplementary Equation 2](#) in [Supplementary Note](#)) to identify higher-order cumulants from $\tilde{\mathbf{A}}$ for order d

3: Perform Fisher permutation tests and obtain $\mathcal{C} \in \mathbb{R}^{m \times n^d}$ along with their parameters of statistical significance, $z = \frac{c - \mu}{\sigma}$, $p = P(Z \geq z)$, where μ : expected value or mean under null hypothesis, and σ : standard deviation

4: Obtain $\mathcal{C}' = \{c_i \in \mathcal{C} \mid p_i < 0.05\}$ of statistically significant redescription groups of size k

5: Obtain the cumulant loadings for \mathcal{C}' from G ([Supplementary Equation 2](#) in [Supplementary Note](#)) in matrix $\mathbf{M} \in \mathbb{R}^{n \times k}$

6: Perform network analysis using CuNA ([Algorithm 2](#)) using \mathcal{C}'

7: Compute risk score estimates per sample using CuRES ([Algorithm 3](#)) using \mathbf{M}

Algorithm 1. Remics: Redescription-based multi-omics analysis.

The cumulants lead to meta-features which are a combination of multi-omics features. It uses this latent interaction space between variables to extract meaningful combinations between the different omics features and analyze them (Figure 1, details in Algorithm 1).

Input: Set \mathcal{C}' of statistically significant redescription groups, $|\mathcal{C}'| = k$

Output: (i) \mathcal{K} Communities of interactions between the multi-omics variables; (ii) ranking of nodes, \mathbf{V} sorted by their relative importance in the network; (iii) p-values, p , and weights, w , of the edge interactions;

- 1: **FOR** all pairs of features (f_i, f_j) in g redescription groups ($g \in \mathcal{C}'$):
- 2: s_{ij} : number of times (f_i, f_j) appear together in g
- 3: Build a network, $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ where vertices are features f_i and f_j , $f_i \in \mathbf{V}$ and $|\mathbf{V}| = n$ and edge $e_{ij} \in \mathbf{E}$ with weight $w(e_{ij}) = s_{i,j}$.
- 4: **FOR** all g groups of features:
- 5: **FOR** all (f_i, f_j) pair of $\binom{k}{2}$ features:
- 6: Build a contingency table $N_{i,j} = n_{i,j}, n_{*,j}, n_{i,*}, n_{*,*}$.
- 7: Obtain p-value $p_{i,j}$ Fisher's exact test on $N_{i,j}$
- 8: **IF** $p_{i,j} < 0.05$
- 9: $\mathbb{E}[U]e_{i,j}$
- 10: $\mathbf{V} \cup (f_i, f_j)$
- 11: **END IF**
- 12: **END FOR**
- 13: **END FOR**
- 14: Perform community detection using any method of choice (Yang et al., 2016) and obtain \mathcal{K} communities.
- 15: Obtain the ranking of nodes in the network $\{f_1, f_2, \dots, f_n\}$ using centrality analysis aggregation.
- 16: Obtain $(p, w) \forall e \in \mathbf{E}$.

Algorithm 2. CuNA: Cumulant-based Network Analysis.

2.3 CuNA

Cumulant-based network analysis (CuNA) computes a network with statistically significant connections between any pair of features in $\mathbf{M} \in \mathbb{R}^{n \times k}$ (from Algorithm 1). It analyzes the network by extracting communities, ranks the features based on their relative importance in the network, and obtains insights on molecular underpinnings of clustered multi-omics features. Manual inspection of the entire network can be challenging when the network is dense. To aid the parsing of results obtained from CuNA analysis, we developed an interactive web-tool to display relevant collections of sub-graphs or communities and highlight their common edges allowing the user to query the network. An outline and details of this algorithm are present in Algorithm 2 with further details on the ranking procedure and the visualizer in the Supplementary Note.

2.3.1 Parameter selection and robustness

CuNA employs a user-defined p – value threshold to control edge inclusion, reflecting the problem-specific nature of multi-omics

association structure. We recommend evaluating thresholds over a broad range (e.g., $p \in [10^{-2}, 10^{-16}]$ with step sizes of 10^{-2} or 10^{-4}) and assessing network stability by examining node rankings, community assignments, and global network statistics across this range. While edge density varies with stringency, high-centrality nodes and major communities remain stable over wide range of thresholds, indicating that CuNA captures higher-order interaction structure rather than threshold-specific artifacts. CuNA uses Fisher's exact test, whose conservative behavior in sparse, high-dimensional settings provide additional protection against false-positive edge inflation.

Input: A matrix $\mathbf{M} \in \mathbb{R}^{m \times k}$, where n is the number of individuals and k is the number of statistically significant redescription groups; outcome, $\mathbf{y} \in \mathbb{R}^m$.

Output: A vector $\mathbf{s} \in \mathbb{R}^n$ representing CuRES.

- 1: Split \mathbf{M} into \mathbf{M}_{tr} , training and \mathbf{M}_{te} , testing data randomly.
- 2: Solve $\mathbf{y} = \mathbf{M}_{tr}\beta$ where $\beta \in \mathbb{R}^k$ and obtain the effect size, $\hat{\beta}$.
- 3: $\mathbf{s} = \sum_i^k \mathbf{M}_{te_i} \hat{\beta}_i$ where \mathbf{M}_{te_i} is a column of the matrix \mathbf{M}_{te} representing the i^{th} redescription group.

Algorithm 3. CuRES: Cumulant-based Risk Scores.

2.4 CuRES

Remics has capabilities to compute an individual assessment of risk for a trait or disease from multi-omics data, called CuRES, from the redescription groups. It takes significant groups, in the form of meta-features and their corresponding cumulant loadings per individual, and splits the data into train and test sets. It fits a generalized linear model on the training data, learns the effect sizes, and then computes an aggregate sum of the meta-feature values per individuals in the test set weighted by the learned effect sizes. This creates a vector of scores per individual that is significantly associated with the target trait or the incidence of the disease. In its mathematical form, the estimated CuRES, $\hat{\mathbf{S}}$ is obtained as the sum across k meta-features, weighted by their weights or coefficient of the linear model, $\hat{\beta}_j$.

$$\hat{\mathbf{S}} = \sum_{j=1}^k \mathbf{M}_j \hat{\beta}_j$$

CuRES can be generalized to any trait or disease, providing a predisposition or risk estimate per sample for that trait computed from a holistic multi-omics perspective (details in Algorithm 3).

2.5 Data

2.5.1 Simulation studies

We designed a multi-omics simulator integrating phenotypes, genotypes, and gene expression levels (Platt D. E. et al., 2024). To handle the integration of different omics data we started with a multivariate distribution

$$f(x)d^d x = \sqrt{\frac{\det(A)}{(2\pi)^d}} \exp\left(-\frac{1}{2}(x-\mu)^T A(x-\mu)\right) d^d x$$

Components of x were identified as phenotypic (binary, which may include environmental conditions as well), single nucleotide polymorphisms (SNPs) (pairs of binary alleles, one for each of the chromosome pairs), or gene expression (float). Covariances A^{-1} were specified in terms of $A = \sigma \text{cor}(x, x^T) \sigma$ where the σ is a diagonal matrix with values representing the spread of the variates, and $\text{cor}(x, x^T)$ is specified to yield correlations among phenotypes, alleles between each pair of chromosomes representing Hardy-Weinberg disequilibrium, and among gene expression levels reflecting co-regulation among pathways (more details in [Supplementary Material](#)). We simulated three different scenarios for 1,000 samples and 30 features (10 phenotypes, 10 SNPs, and 10 genes with varying expression levels) to demonstrate Remics' ability to identify genotype-phenotype interactions with the highest Pearson correlation coefficient (r^2) and its robustness in the presence of false positives while correcting for spurious associations. These scenarios with varying correlations were: (i) an extreme case where only a few features among the genes, SNPs, and phenotypes were highly correlated with each other ([Figure 3](#)); (ii) an average case where many of the features were moderately correlated with each other ([Supplementary Figures S1, S2](#)); (iii) a sanity check with completely uncorrelated features, therefore, the resulting correlation matrix being equal to an identity matrix.

2.5.2 TCGA multi-omics datasets

We utilized six processed, benchmark multi-omics datasets from Rappoport et al. ([Rappoport and Shamir, 2018](#)) that were derived from The Cancer Genome Atlas (TCGA). Each dataset, which contains miRNA expression, gene expression, and DNA methylation data, represents primary tumor samples from a different cancer: acute myeloid leukemia (AML), colon adenocarcinoma (COAD), glioblastoma multiforme (GBM), kidney renal cell carcinoma (KIRC), ovarian serous cystadenocarcinoma (OV), and sarcoma (SARC). Data was downloaded from http://acgt.cs.tau.ac.il/multi_omic_benchmark/download.html. In summary, these pre-processed datasets excluded patients and features with >20% missing values with remaining missing values imputed using k-nearest neighbor imputation, as well as filtering methylation data for the top 5,000 features by variance. For each dataset, we performed orthogonal matching pursuit (OMP) ([Tropp and Gilbert, 2007](#)) using *scikit-learn's* `OrthogonalMatchingPursuit` function to select 387 gene expression features, 95 DNA methylation features, and 20 miRNA features (preserving their proportions from the quality controlled data) for all downstream analyses. We processed the "survival" outcome in the data, to create three binary outcomes, namely, "1-year", "3-year", and "5-year" survival, to perform downstream prediction and network analysis. The number of overlapping samples across all the cancer types, as well as their number of patients dead or surviving are shown in [Figure 2](#).

2.6 Analysis

2.6.1 Classification

We performed classification analysis using CuRES on the cumulant loadings for those significant cumulants observed in C'

in [Algorithm 1](#). We used different thresholds of significance to filter our results for the best performing threshold by evaluating the results from 12 p -values over the range $p = [0.01, 10^{-24}]$ decremented by 10^{-2} . We used LogisticRegression classifier with L_1 penalty from *scikit-learn* ([Pedregosa et al., 2011](#)) package in Python for each classification task. We used age and gender as confounding variables for the regression analysis. We performed five-fold cross validation with hyperparameter tuning using the `GridSearchCV` function.

2.6.2 Networks

To test the significance of CuNA, we evaluated the interactions over the same range of p -values as in the classification analysis. The width of the edges corresponds to the number of times a pair of nodes appeared together in the redescription groups and reflects their pairwise affinity. The nodes were ranked in the order of their importance computed as an aggregate score of mean of the ranks in different network centrality measures (see [Supplementary Note](#) for more details). We used the *networkx* package ([Hagberg et al., 2008](#)) in Python to perform all network analyses including community detection using greedy modularity method.

3 Results

3.1 Simulation studies

3.1.1 Network analysis

We applied Remics on different simulation scenarios varying from easy to complex interactions between genes, SNPs, and phenotypes.

In the first scenario of 11 variables, only a few interactions such as (*Gene0, SNP0*), (*Gene0, SNP2*), and (*Pheno0, Pheno1*) had correlation r^2 of 0.9, 0.8, and 0.6, respectively. We found the CuNA-projected higher-order interactions captured these highly correlated variables ([Figure 3](#)) within three communities from the network, mimicking the spiked-in correlations: {*Gene0, SNP2, Pheno2*}, {*SNP0*}, and {*Pheno0, Pheno1*}. Increasing the complexity of these interactions by involving more samples (varying from 100 to 1,000) as well as more variables (varying from 10 to 30), we found that CuNA accurately found the interacting variables and the communities reflected the clusters of the highly correlated variables together ([Supplementary Figures S1, S2](#)). Thus, CuNA captures the communities accurately as reflected in the network ([Figure 3](#)) as well as the original correlations that were input to [Algorithm 1](#).

3.1.2 Prediction with CuRES

We assigned the phenotype *Pheno0* as the target variable and considered the rest of the 10 variables as part of the data matrix and computed CuRES ([Supplementary Figure S3](#)). We observed that the net reclassification index (NRI), which is computed as the difference in classification accuracy when including CuRES as a variable in the data matrix versus the original data matrix without CuRES, was 1% with a prediction accuracy (measured by the F_1 score) of 84.7% on held out data after five-fold cross validation with a logistic regression

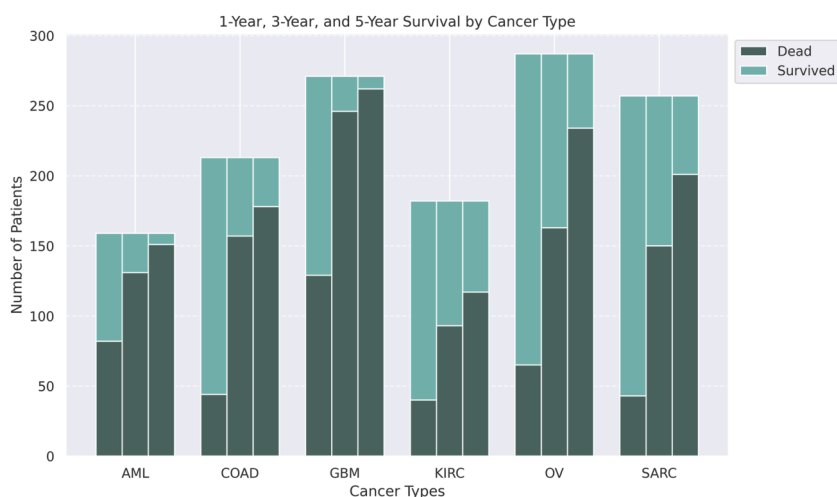


FIGURE 2 Stacked barplots of the number of patients along with their status information (dead or survived) for 1-, 3-, or 5-year survival (left to right) for six cancers, AML, COAD, GBM, KIRC, OV, and SARC.

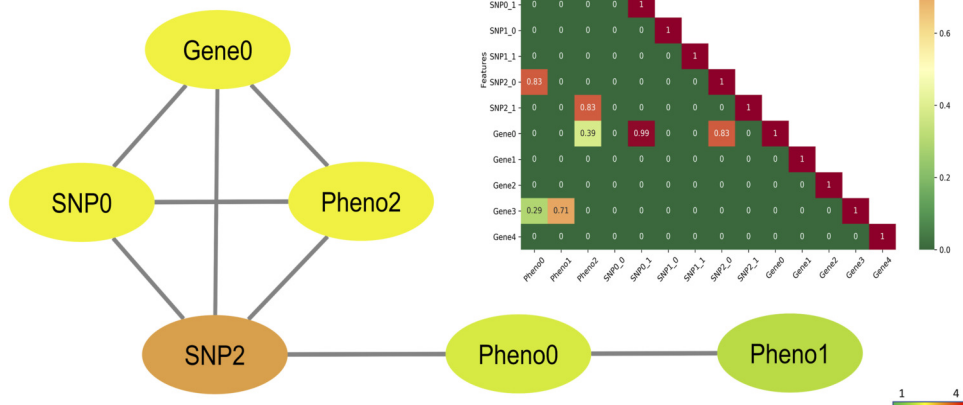


FIGURE 3 CuNA network of the simulated variables with few highly correlated features. The nodes are colored by degrees (darker colors have higher degree). The correlation matrix of the 11 variables are shown in the inset with color gradient corresponding to the feature correlation. The features are organized in the matrix in the following order: 3 phenotypes, 2 genotypes (two allele each), and 5 genes with the maximum correlation of 0.99 between allele 1 of *SNP0* and *Gene0* shown in red.

model. When extended to 30 variables and considering *Pheno5* as the target variable, we found that CuRES improved prediction performance by as much as 16% with perfect classification. We note that the effect CuRES has on the NRI varies with the correlation between the simulated variables and number of samples.

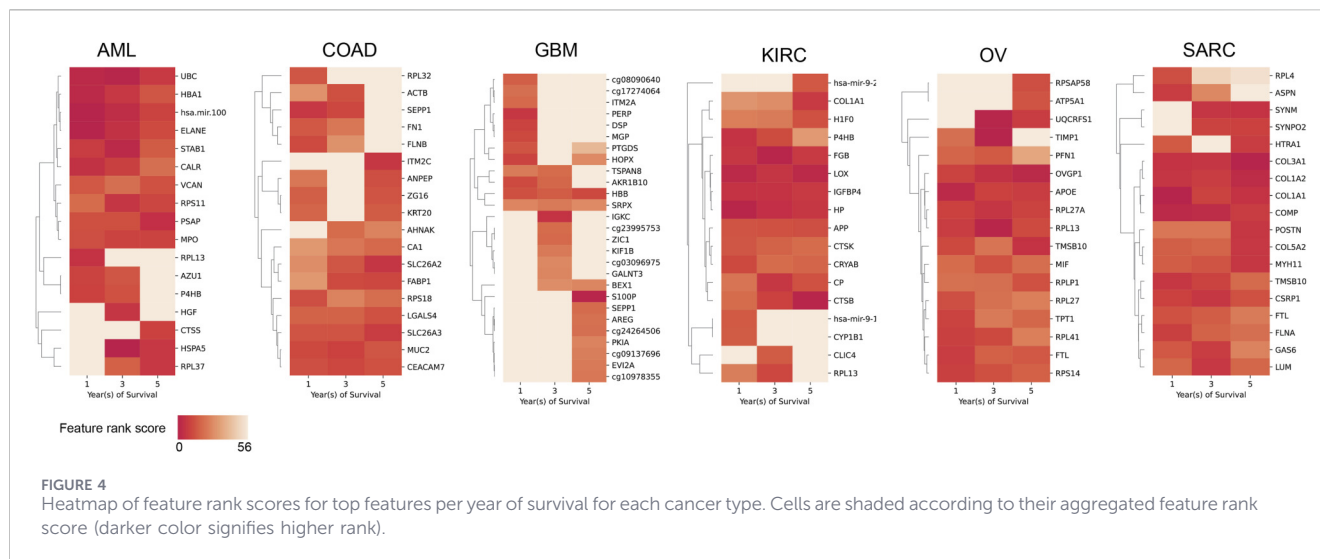
3.2 Analyzing TCGA multi-omics cancer data

We analyzed six cancer types from TCGA using Remics on the 500 top selected features and computed redescription groups of order 3. For each cancer type, we performed an early integration of

the selected transcriptomic, DNA methylation, and miRNA features, after normalization, for each outcome (“1-year”, “3-year”, and “5-year”) based on the survival information. We observed a significant expansion of the feature set for each cancer, with the number of significant ($p < 0.01$) redescription groups yielding up to approximately 700,000 meta-features (Supplementary Figure S4).

3.2.1 Network analysis

We applied CuNA (Algorithm 2) after deriving the statistically significant redescription groups from Remics. We filtered edges



for statistical significance ($p < 0.01$), and obtained varying number of vertices, represented as multi-omics features (AML: 236; COAD: 279; GBM: 375; KIRC: 259; OV: 338; SARC: 319) and edges (Supplementary Figure S5). We obtained the highest-ranked features as well as the four top-ranked edges based on their weight for each cancer type from the networks using our network centrality-based ranking algorithm (Figure 4; Table 1; see Supplementary Table S1 for details).

We performed community detection on the network generated by CuNA (Algorithm 2) and obtained communities of multi-omics features. To visualize how the communities interact within and between each other, we projected the networks into our interactive visualizer and selected community subgroups to reveal interactions between ribosomal protein gene network in OV, enriched transcription factor NFE2L2 in AML, DNA methylation interactions in GBM, etc. (Supplementary Figures S6–S8). We evaluated the detected communities for enrichment of biological pathways using the MSigDB Hallmark collection (v2024.1) (Liberzon et al., 2015). Over-represented pathways were identified in each community for every survival year (Supplementary Figure S9). This analysis revealed both well-established and emerging pathway associations that may point to new avenues of investigation. For instance, the *MYC Targets* pathway was significantly enriched across multiple time points in AML and OV, consistent with prior reports of its involvement in these cancers (Ohanian et al., 2019; Call et al., 2020; Ju et al., 2018). Pathways linked to cell cycle regulation—such as E2F targets, G2M checkpoint, and epithelial-mesenchymal transition—were also recurrently enriched across several cancer types and survival years. Additionally, we observed enrichment of hypoxia-related pathways in COAD communities, aligning with growing evidence of hypoxia’s contribution to colorectal cancer progression (Fletcher et al., 2022).

3.2.2 Disease prediction with CuRES

For each cancer type, we performed early integration for multi-omics variables after normalization and used it to serve as the baseline for our experiments. We computed CuRES (Algorithm

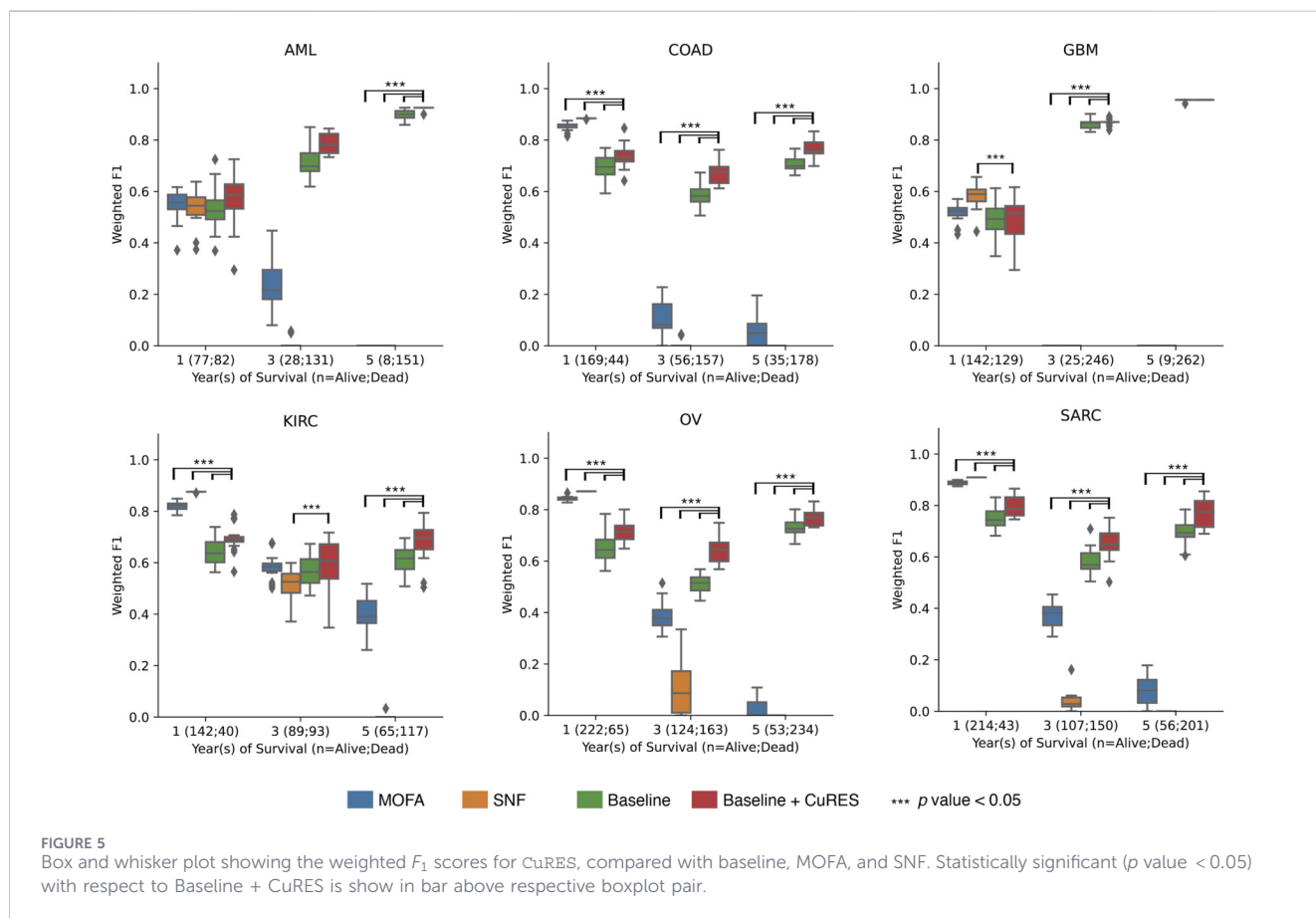
3) on the three survival outcomes and on the statistically significant redescription groups obtained from Algorithm 1. We added the CuRES vector to the baseline to check for a change in outcome prediction. To compare with state-of-the-art multi-omics integration methods MOFA (Argelaguet et al., 2018) and SNF (Wang et al., 2014), we applied them on integrated multi-omics data and compared the weighted F_1 score. To ensure a fair comparison, CuRES, MOFA, and SNF, all were applied to the same training data and evaluated on the same held-out test data. We performed experiments with bootstrapping ($n = 20$) to obtain the confidence intervals for the weighted F_1 score. We observed CuRES, when considered as a covariate with the integrated multi-omics data, consistently improves upon the baseline across all the cancers. In most cases CuRES performs better than MOFA and SNF for predicting the “3-year” and “5-year” outcomes (Figure 5). MOFA and SNF both perform better than CuRES when predicting the “1-year” outcome for most cancers where there exists a heavy class imbalance between the number of dead to the number of surviving patients. CuRES had a mean net reclassification index (NRI) of approximately 4% (with baseline), -7% (with MOFA), -10% (with SNF) for the “1-year” outcome; 6% (with baseline), 42% (with MOFA), 59% (with SNF) for the “3-year” outcome; and 4% (with baseline), 71% (with MOFA), 81% (with SNF) for the “5-year” outcome (Figure 6).

3.3 Complexity analysis

At the heart of Remics is the computation of cumulants or higher-order moments between features. The computational complexity and resource usage increases exponentially by a factor of n^d , where d is the order of the cumulants and n is the set of variables. Thus, overall computational complexity becomes $\mathcal{O}(mn^d)$ where m is the number of samples (Domino et al., 2018). Furthermore, the space requirement for computing higher-order cumulants is more challenging. We performed an in-depth complexity analysis using the TCGA data with 1,000 samples and observed a requirement of 100 GB memory to compute sixth-order cumulants with only 50 features (Supplementary Figure S10). Thus, the computation is intractable for very large sets of variables.

TABLE 1 Top ranked nodes (multi-omics features) and edges (interactions between features) for each cancer type across three outcomes using the network-centrality based ranking algorithm and the edge weight, respectively.

Cancer type	Multi-omics features (nodes)	Interactions (edges)
AML	{UBC, hsa-mir-100, ELANE, HBA1, STAB1}	(ELANE, STAB1), (hsa-mir-100, STAB1), (HSPA5, CTSD), (FTL, UBC)
COAD	{MUC2, SLC26A3, CEACAM7, SLC26A2, FABP1, AZU1}	(TAGLN, ACTG2), (TAGLN, FLNA), (DES, ACTG2)
GBM	{SRPX, HBB}	(AKR1B10, HBB), (AKR1B10, STON1)
KIRC	{LOX, HP, FGB, IGFBP4, CTSB, CP}	(LOX, FGB),(FGB, CP), (COL3A1, COL1A1)
OV	{OVGP1, APOE, FTL, RPL13}	(DLK1, MEST), (DLK1, hsa-mir-891a), (OVGP1, APOE)
SARC	{COL3A1, COMP, COL1A1, COL1A2}	(APOD, MPZ), (PYGM, HSP90B1),(LAMP1, GAS6)

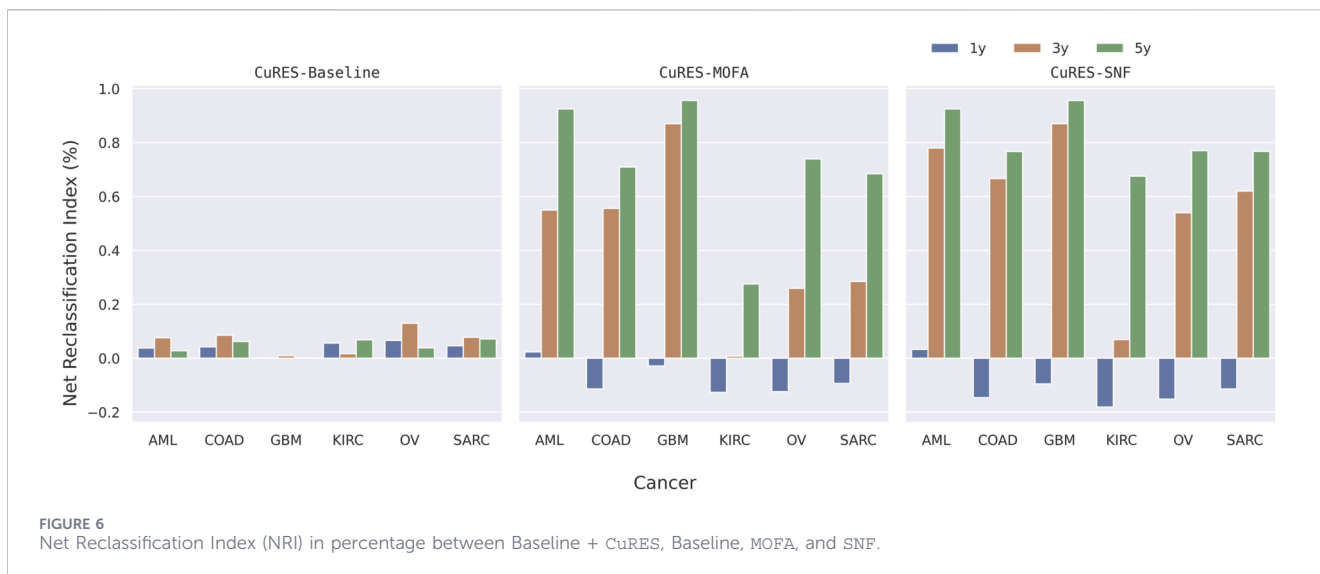


Although, in practice, it is rare to compute the cumulants beyond fifth order, and the time and memory overhead are significantly less when computing third order cumulants with 100 or fewer features. For the multi-omics analyses done here, the highest memory requirement for Remics was approximately 94 GB, taking 45 min to compute third order cumulants for 271 samples of GBM and 500 features.

4 Discussion

Remics, introduced here, is a topology driven, interpretable, integrative multi-omics analysis that supports tasks such as network

analysis, biomarker discovery, and disease prediction. Rather than assuming direct regulatory or mechanistic relationship between molecular layers, Remics identifies sets of features that exhibit statistically equivalent variation across samples, capturing higher-order cross-omic structure in a data-driven manner. It builds on the concept of redescription mining (Parida and Ramakrishnan, 2005; Ramakrishnan and Zaki, 2009; Karisani et al., 2022; Platt D. et al., 2024) to uncover multi-omics feature groups that jointly describe shared biological heterogeneity. Remics offers both sample-level and feature-level analyses with CuRES and CuNA, respectively. CuRES provides an individual-level measure of disease liability or risk by leveraging the multi-omics structure uncovered by Remics. Rather than relying on individual molecular features,



CuRES represents each patient using redescription groups, which are interpretable multi-omics, meta-features that capture statistically equivalent variation across samples. With the increasing availability of large, EHR-linked biobanks and sequencing technologies, this framework enables a holistic and extensible summary of disease risk that integrates multi-omics and multi-modal data within a unified representation. The risk score estimates that are computed from meta-features in Remics provide a logical extension to the polygenic risk score in genomics, which has found widespread usage in research and medicine (Choi et al., 2020). However, comparing CuRES with PRS is beyond the scope of this work.

Applying CuRES to TCGA cancer cohorts across multiple survival data, we observed that CuRES outperforms established multi-omics integration methods such as MOFA and SNF, in predicting “3-year” and “5-year” outcomes. In contrast, under conditions of severe class imbalance (>3.5:1 between deceased and surviving patients), which most frequently arise in “1-year” outcome prediction, MOFA and SNF tend to achieve better weighted F_1 scores by favoring the majority class. In these settings, CuRES yields more balanced precision–recall trade-offs, reflecting its emphasis on stable, cross-omics structure rather than majority-class optimization. We therefore recommend interpreting short-term outcome predictions using class-specific metrics (e.g., recall for early mortality) in conjunction with weighted F_1 , particularly when therapeutic decision-making is time sensitive. Importantly, CuRES generalizes well to unseen, held-out data and supports transfer learning in multi-omics analysis by providing pretrained models and summary statistics for redescription groups, enabling reuse across cohorts and outcome horizons.

Complementarily, CuNA embeds significant higher-order interactions in the form of redescription groups to a network of multi-omics features. The edge weight between a pair of features in CuNA signifies the number of times they were together in higher-order redescription groups. Thus, interactions with highest weights indicate their relative importance in explaining underlying biological functions of the disease and can be seen as a pattern discovery mechanism in multi-omics data. The interactions found by CuNA were validated using the IntAct database of molecular interactions (Del Toro et al.,

2022), such as the interactions between ELANE, HSPA5, FLT, UBC, etc. in AML, which are connected by heparin, an anticoagulant in blood and is a known treatment for AML (Kovacovics et al., 2018). Another interaction that highlighted the underlying biological function was between the TAGLN and ACTG2 genes, which were connected by CTFR in IntAct. TAGLN has a long history of association with COAD (Zhou et al., 2016) and CTFR with cystic fibrosis also has associations with COAD (Scott et al., 2020). In KIRC, the interaction between LOX and COL1A1 were also independently validated in literature (Di Stefano et al., 2016).

CuNA was able to find implicit and explicit connections to the respective cancer types (see Supplementary Note for other validated interactions). For example, most of the top-ranked features for AML, such as UBC, hsa-mir-100, etc. have been shown to be associated with AML. The UBC (Ubiquitin C) ligase determines AML growth and susceptibility to histone deacetylase inhibitors (Khateb et al., 2021), while the miRNA hsa-mir-100 regulates cell differentiation and survival by targeting RBSP3, a phosphatase-like tumor suppressor in AML (Zheng et al., 2012). These connections are found in all of the cancer types studied. Loss of function mutations in MUC2 has been shown to increase colon cancer tumor progression (Hsu et al., 2017; Betge et al., 2016). SRPX is being investigated as a biomarker for GBM (Ampudia-Mesias et al., 2022). OVGPI expression has been associated with OV (Wu et al., 2016), and COL1A1-PDGFB gain of function drives tumor growth in SARC (Abbott et al., 2006).

Despite all its prowess with multi-omics data, Remics is sensitive to the early integration pitfalls of multi-omics data as discussed in this review (Subramanian et al., 2020). Most of the multi-omics methods are often plagued by this issue (Picard et al., 2021) and thus use different strategies, such as variable selection or latent space analysis, to work around the bottleneck. While Remics’ main limitation is scalability in massive datasets for orders $d \geq 3$ (Supplementary Figure S10) owing to the technical limitations of current computational devices and thus leading to the use of mitigating techniques, such as feature selection. This reduction of complexity is not a required component of Remics, but rather a concession to available computational hardware. While such

dimensionality reduction techniques are commonplace in biological data analysis, we have endeavored to perform an unbiased feature selection of sufficient size to, some degree, mitigate concerns of skewing results towards known cancer driver genes or pathways. As computing advances such dimensionality reduction methods may become obsolete, for example, by using approximate cumulants from tensor decomposition techniques (Morton and Lim, 2009; Domino et al., 2018) or developing quantum computing algorithms for computing cumulants (Bose et al., 2026) or tensor decomposition (Burch et al., 2025), thereby fulfilling the promise of faster, scalable computation to higher-orders, thus allowing the full promise of Remics to be realized. Furthermore, as cumulants are widely used in many fields of research such as economics, physics, etc., a speed-up on cumulant computation would have wider implications beyond healthcare and biology.

5 Conclusion

Associations between multi-omics variables can be complex and often confounded by environmental factors. We propose a framework, Remics, to identify associations with more granularity than a standard case-control association study while performing tasks such as biomarker discovery and prediction of traits in complex diseases. We demonstrate that Remics captures true associations by validating it on simulated data as obtained from the multi-omics simulator (Platt D. E. et al., 2024). We demonstrated CuNA's application in multiple diseases such as AML, COAD, GBM, KIRC, OV, and SARC, where it was able to find functionally relevant biomarkers, predict outcomes, and perform better than the state-of-the-art methods such as MOFA and SNF. Lastly, we show that the CuRES module can be a useful predictor of the disease state and help to understand an individual's risk of a disease based on multi-omics and multi-modal variables ranging from imaging, transcriptomics, proteomics, metabolomics, etc., rather than just on genomics. Remics provides an exciting opportunity to decode phenotypic and genotypic diversity and discover biomarkers associated with various manifestations of complex diseases, paving the way for accelerated personalized medicine.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: <https://portal.gdc.cancer.gov/>.

Ethics statement

Ethical approval was not required for the studies involving humans because we used publicly available data which does not require ethics committee approval. The studies were conducted in accordance with the local legislation and institutional requirements. The human samples used in this study were acquired from a by-product of routine care or industry. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

AB: Conceptualization, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Writing – original draft, Writing – review and editing. DP: Conceptualization, Investigation, Methodology, Writing – review and editing. KR: Data curation, Formal Analysis, Methodology, Software, Writing – review and editing. MB: Software, Validation, Writing – review and editing. AG: Software, Visualization, Writing – review and editing. NH: Data curation, Resources, Writing – review and editing. LP: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review and editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This work has been supported by IBM Research. The funder was not involved in the study design, collection, analysis, interpretation of data, the writing of this article, or the decision to submit it for publication.

Conflict of interest

Authors AB, DP, KR, MB, AG, and LP were employed by IBM Research at the time of writing the article. Author NH was employed by the company DAIN Studios.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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.

Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fcell.2026.1738010/full#supplementary-material>

References

- Abbott, J. J., Erickson-Johnson, M., Wang, X., Nascimento, A. G., and Oliveira, A. M. (2006). Gains of *coll1a1-pdgfrb* genomic copies occur in fibrosarcomatous transformation of dermatofibrosarcoma protuberans. *Mod. Pathol.* 19 (11), 1512–1518. doi:10.1038/modpathol.3800695
- Ampudia-Mesias, E., El-Hadad, S., Cameron, C. S., Wöhrer, A., Ströbel, T., Saydam, N., et al. (2022). *Srpx* emerges as a potential tumor marker in the extracellular vesicles of glioblastoma. *Cancers* 14 (8), 1984. doi:10.3390/cancers14081984
- Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., et al. (2018). Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Mol. Systems Biology* 14 (6), e8124. doi:10.15252/msb.20178124
- Betje, J., Schneider, N. I., Harbaum, L., Pollheimer, M. J., Lindtner, R. A., Kornprat, P., et al. (2016). *Muc1, muc2, muc5ac, and muc6* in colorectal cancer: expression profiles and clinical significance. *Virchows Arch.* 469, 255–265. doi:10.1007/s00428-016-1970-5
- Bose, A., Platt, D. E., Haiminen, N., and Parida, L. (2021). in *Cuna: Cumulant-Based Network Analysis of Genotype-Phenotype Associations in Parkinson's Disease*. medRxiv, 2021–2108.
- Bose, A., Platt, D. E., Haiminen, N., and Parida, L. (2023). *Identifying Therapeutic Biomarkers Associated With Complex Diseases*. uS Patent App. 17/453,221.
- Bose, A., Rhrissorakrai, K., Utro, F., Parida, L. Quantum for Healthcare Life Sciences Consortium (2026). Advancing single-cell omics and cell-based therapeutics with quantum computing. *Nat. Rev. Mol. Cell Biol.*, 1–15. doi:10.1038/s41580-025-00918-0
- Burch, M., Zhang, J., Idumah, G., Doga, H., Lartey, R., Yehia, L., et al. (2025). Towards quantum tensor decomposition in biomedical applications. *arXiv Preprint arXiv:2502.13140*. doi:10.48550/arXiv.2502.13140
- Call, S. G., Duren, R. P., Panigrahi, A. K., Nguyen, L., Freire, P. R., Grimm, S. L., et al. (2020). Targeting oncogenic super enhancers in myc-dependent aml using a small molecule activator of nr4a nuclear receptors. *Sci. Rep.* 10 (1), 2851. doi:10.1038/s41598-020-59469-3
- Chaudhary, K., Poirion, O. B., Lu, L., and Garmire, L. X. (2018). Deep learning-based multi-omics integration robustly predicts survival in liver cancer using deep learning to predict liver cancer prognosis. *Clin. Cancer Res.* 24 (6), 1248–1259. doi:10.1158/1078-0432.CCR-17-0853
- Choi, S. W., Mak, T. S. H., and O'Reilly, P. F. (2020). Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protocols* 15 (9), 2759–2772. doi:10.1038/s41596-020-0353-1
- Del Toro, N., Shrivastava, A., Ragueneau, E., Meldal, B., Combe, C., Barrera, E., et al. (2022). The intact database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Research* 50 (D1), D648–D653. doi:10.1093/nar/gkab1006
- Di Stefano, V., Torsello, B., Bianchi, C., Cifola, I., Mangano, E., Bovo, G., et al. (2016). Major action of endogenous lysyl oxidase in clear cell renal cell carcinoma progression and collagen stiffness revealed by primary cell cultures. *Am. J. Pathology* 186 (9), 2473–2485. doi:10.1016/j.ajpath.2016.05.019
- Domino, K., Gawron, P., and Pawela, L. (2018). Efficient computation of higher-order cumulant tensors. *SIAM J. Sci. Comput.* 40 (3), A1590–A1610. doi:10.1137/17m1149365
- Fletcher, T., Thompson, A. J., Ashrafian, H., and Darzi, A. (2022). The measurement and modification of hypoxia in colorectal cancer: overlooked but not forgotten. *Gastroenterol. Rep. (Oxf)* 10, goac042. doi:10.1093/gastro/goac042
- Hagberg, A., Swart, P. J., and Schult, D. A. (2008). *Exploring network structure, dynamics, and function using networkx*. Tech. rep. Los Alamos, NM (United States): Los Alamos National Laboratory LANL.
- Hsu, H. P., Lai, M. D., Lee, J. C., Yen, M. C., Weng, T. Y., Chen, W. C., et al. (2017). Mucin 2 silencing promotes colon cancer metastasis through interleukin-6 signaling. *Sci. Reports* 7 (1), 5823. doi:10.1038/s41598-017-04952-7
- Jung, M., Russell, A. J., Kennedy, C., Gifford, A. J., Australian Ovarian Cancer Study, G., Mallitt, K. A., et al. (2018). Clinical importance of myc family oncogene aberrations in epithelial ovarian cancer. *JNCI Cancer Spectr.* 2 (3), pky047. doi:10.1093/jncics/pky047
- Karisani, N., Platt, D. E., Basu, S., and Parida, L. (2022). Topology and redescription detect multiple alternative biological pathways from clinical phenotypes. *Exp. Biol. Med.* 247 (22), 2015–2024. doi:10.1177/15353702221126671
- Khateb, A., Deshpande, A., Feng, Y., Finlay, D., Lee, J. S., Lazar, I., et al. (2021). The ubiquitin ligase *rnf5* determines acute myeloid leukemia growth and susceptibility to histone deacetylase inhibitors. *Nat. Commun.* 12 (1), 5397. doi:10.1038/s41467-021-25664-7
- Kovacs, T. J., Mims, A., Salama, M. E., Pantin, J., Rao, N., Kosak, K. M., et al. (2018). Combination of the low anticoagulant heparin *cx-01* with chemotherapy for the treatment of acute myeloid leukemia. *Blood Advances* 2 (4), 381–389. doi:10.1182/bloodadvances.2017013391
- Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* 401 (6755), 788–791. doi:10.1038/44565
- Lee, B., Zhang, S., Poleksic, A., and Xie, L. (2020). Heterogeneous multi-layered network model for omics data integration and analysis. *Front. Genetics* 10, 1381. doi:10.3389/fgene.2019.01381
- Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell Systems* 1 (6), 417–425. doi:10.1016/j.cels.2015.12.004
- Morton, J., and Lim, L. H. (2009). *Principal cumulant component analysis*. preprint.
- Ohanian, M., Rozovski, U., Kanagal-Shamanna, R., Abruzzo, L. V., Loghavi, S., Kadia, T., et al. (2019). Myc protein expression is an important prognostic factor in acute myeloid leukemia. *Leukemia and Lymphoma* 60 (1), 37–48. doi:10.1080/10428194.2018.1464158
- Parida, L., and Ramakrishnan, N. (2005). Redescription mining: structure theory and algorithms. *AAAI* 5, 837–844. doi:10.5555/1619410.1619467
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. doi:10.5555/1953048.2078195
- Picard, M., Scott-Boyer, M. P., Bodein, A., Périn, O., and Droit, A. (2021). Integration strategies of multi-omics data for machine learning analysis. *Comput. Struct. Biotechnol. J.* 19, 3735–3746. doi:10.1016/j.csbj.2021.06.030
- Platt, D., Bose, A., Rhrissorakrai, K., Levovitz, C., and Parida, L. (2024). Epidemiological topology data analysis links severe covid-19 to raas and hyperlipidemia associated metabolic syndrome conditions. *Bioinformatics* 40 (Suppl. ment_1), i199–i207. doi:10.1093/bioinformatics/btae235
- Platt, D. E., Bose, A., Rhrissorakrai, K., Saenz, A. G., Haiminen, N., and Parida, L. (2024). *Create Synthetic Patient Data Using a Generative Adversarial Network Having a Multivariate Gaussian Generative Model*. uS Patent App. 17/930,477.
- Ramakrishnan, N., and Zaki, M. J. (2009). “Redescription mining and applications in bioinformatics,” in *Biological Data Mining* (Chapman and Hall/CRC), 581–606.
- Rappoport, N., and Shamir, R. (2018). Multi-omic and multi-view clustering algorithms: rapid and cancer benchmark. *Nucleic Acids Res.* 46 (20), 10546–10562. doi:10.1093/nar/gky889
- Rappoport, N., and Shamir, R. (2019). Nemo: cancer subtyping by integration of partial multi-omic data. *Bioinformatics* 35 (18), 3348–3356. doi:10.1093/bioinformatics/btz058
- Rohart, F., Gautier, B., Singh, A., and Lê Cao, K. A. (2017). Mixomics: an R package for omics feature selection and multiple data integration. *PLoS Computational Biology* 13 (11), e1005752. doi:10.1371/journal.pcbi.1005752
- Scott, P., Anderson, K., Singhanian, M., and Cormier, R. (2020). Cystic fibrosis, *cfr*, and colorectal cancer. *Int. Journal Molecular Sciences* 21 (8), 2891. doi:10.3390/ijms21082891
- Song, M., Greenbaum, J., Luttrell IV, J., Zhou, W., Wu, C., Shen, H., et al. (2020). A review of integrative imputation for multi-omics datasets. *Front. Genetics* 11, 570255. doi:10.3389/fgene.2020.570255
- Speicher, N. K., and Pfeifer, N. (2015). Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics* 31 (12), i268–i275. doi:10.1093/bioinformatics/btv244
- Subramanian, I., Verma, S., Kumar, S., Jere, A., and Anamika, K. (2020). Multi-omics data integration, interpretation, and its application. *Bioinforma. Biology Insights* 14, 1177932219899051. doi:10.1177/1177932219899051
- Tropp, J. A., and Gilbert, A. C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory* 53 (12), 4655–4666. doi:10.1109/TIT.2007.909108
- Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., et al. (2014). Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* 11 (3), 333–337. doi:10.1038/nmeth.2810
- Wu, R., Zhai, Y., Kuick, R., Karnezis, A. N., Garcia, P., Naseem, A., et al. (2016). Impact of oviductal versus ovarian epithelial cell of origin on ovarian endometrioid carcinoma phenotype in the mouse. *J. Pathology* 240 (3), 341–351. doi:10.1002/path.4783
- Yang, Z., Algesheimer, R., and Tessone, C. J. (2016). A comparative analysis of community detection algorithms on artificial networks. *Sci. Reports* 6 (1), 1–18. doi:10.1038/srep30750
- Zheng, Y., Zhang, H., Zhang, X., Feng, D., Luo, X., Zeng, C., et al. (2012). *Mir-100* regulates cell differentiation and survival by targeting *rbps3*, a phosphatase-like tumor suppressor in acute myeloid leukemia. *Oncogene* 31 (1), 80–92. doi:10.1038/nc.2011.208
- Zhou, H. m., Fang, Y. y., Weinberger, P. M., Ding, L. l., Cowell, J. K., Hudson, F. Z., et al. (2016). Transgelin increases metastatic potential of colorectal cancer cells *in vivo* and alters expression of genes involved in cell motility. *BMC Cancer* 16, 1–11. doi:10.1186/s12885-016-2105-8