# FRONTIERS IN COMPUTATIONAL NEUROSCIENCE – EDITORS' PICK 2021

EDITED BY: Si Wu and Misha Tsodyks

PUBLISHED IN: Frontiers in Computational Neuroscience

frontiers Research Topics

# frontiers



ISSN 1664-8714
ISBN 978-2-88974-133-5
DOI 10.3389/978-2-88974-133-5

## 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

# FRONTIERS IN COMPUTATIONAL NEUROSCIENCE – EDITORS' PICK 2021

Topic Editors:
**Si Wu,** Peking University, China
**Misha Tsodyks,** Weizmann Institute of Science, Israel

**Citation:** Wu, S., Tsodyks, M., eds. (2022). Frontiers in Computational Neuroscience – Editors' Pick 2021. Lausanne: Frontiers Media SA.
doi: 10.3389/978-2-88974-133-5

# Table of Contents

frontiers in Computational Neuroscience

METHODS
published: 05 April 2018
doi: 10.3389/fncom.2018.00024

# Unsupervised Feature Learning With Winner-Takes-All Based STDP

*Paul Ferré*[1,2]*, *Franck Mamalet*[2] *and Simon J. Thorpe*[1]

[1] *Centre National de la Recherche Scientifique, UMR-5549, Toulouse, France,* [2] *Brainchip SAS, Balma, France*

We present a novel strategy for unsupervised feature learning in image applications inspired by the Spike-Timing-Dependent-Plasticity (STDP) biological learning rule. We show equivalence between rank order coding Leaky-Integrate-and-Fire neurons and ReLU artificial neurons when applied to non-temporal data. We apply this to images using rank-order coding, which allows us to perform a full network simulation with a single feed-forward pass using GPU hardware. Next we introduce a binary STDP learning rule compatible with training on batches of images. Two mechanisms to stabilize the training are also presented : a Winner-Takes-All (WTA) framework which selects the most relevant patches to learn from along the spatial dimensions, and a simple feature-wise normalization as homeostatic process. This learning process allows us to train multi-layer architectures of convolutional sparse features. We apply our method to extract features from the MNIST, ETH80, CIFAR-10, and STL-10 datasets and show that these features are relevant for classification. We finally compare these results with several other state of the art unsupervised learning methods.

**Keywords: Spike-Timing-Dependent-Pasticity, neural network, unsupervised learning, winner-takes-all, vision**

OPEN ACCESS

***Edited by:***
*Guenther Palm,*
*Universität Ulm, Germany*

***Reviewed by:***
*Michael Beyeler,*
*University of Washington,*
*United States*
*Stefan Duffner,*
*UMR5205 Laboratoire d'Informatique en Image et Systèmes d'Information (LIRIS), France*

***Correspondence:**
*Paul Ferré*
*paul.ferre@cnrs.fr*

***Received:*** *18 May 2017*
***Accepted:*** *20 March 2018*
***Published:*** *05 April 2018*

***Citation:***
*Ferré P, Mamalet F and Thorpe SJ (2018) Unsupervised Feature Learning With Winner-Takes-All Based STDP. Front. Comput. Neurosci. 12:24. doi: 10.3389/fncom.2018.00024*

## 1. INTRODUCTION

Unsupervised pre-training methods help to overcome difficulties encountered with current neural network based supervised algorithms. Such difficulties include : the requirement for a large amount of labeled data, vanishing gradients during back-propagation and the hyper-parameters tuning phase. Unsupervised feature learning may be used to provide initialized weights to the final supervised network, often more relevant than random ones (Bengio et al., 2007). Using pre-trained weights tends to speed up network convergence, and may also increase slightly the overall classification performance of the supervised network, especially when the amount of labeled examples is small (Rasmus et al., 2015).

Unsupervised learning methods have recently regained interest due to new methods such as Generative Adverserial Networks (Goodfellow et al., 2014; Salimans et al., 2016), Ladder networks (Rasmus et al., 2015), and Variational Autoencoders (Kingma and Welling, 2013). These methods reach state of the art performances, either using top layer features as inputs for a classifier or within a semi-supervised learning framework. As they rely on gradient descent methods to learn the representations for their respective tasks, computations are done with 32-bits floating point values. Even with dedicated hardware such as GPUs and the use of 16-bits half-floats type (Gupta et al., 2015), floating point arithmetic remains time and power consuming for large datasets. Several works are addressing this problem by reducing the resolution of weights, activations and gradients during inference and learning phases (Stromatias et al., 2015; Esser et al., 2016; Deng et al., 2017)

and have shown small to zero loss of accuracy with such supervised methods. Nevertheless, learning features both with unsupervised methods and lower precision remains a challenge.

On the other hand, Spiking Neural Networks (SNNs) propagate information between neurons using spikes, which can be encoded as binary values. Moreover, SNNs often use an unsupervised Hebbian learning scheme, Spike-Timing-Dependent-Plasticity (STDP), to capture representations from data. STDP uses differences of spikes times between pre and post-synaptic neurons to update the synaptic weights. This learning rule is able to capture repetitive patterns in the temporal input data (Masquelier and Thorpe, 2007). SNNs with STDP may only require fully feed-forward propagation to learn, making them good candidates to perform learning faster than backpropagation methods.

Our contribution is three-fold. First, we demonstrate that Leaky Integrate and Fire neurons act as artificial neurons (perceptrons) for temporally-static data such as images. This allows the model to infer temporal information while none were given as input. Secondly, we develop a winner-takes-all (WTA) framework which ensure a balanced competition between our excitatory neuron population. Third, we develop a computationally-efficient and nearly parameter-less STDP learning rule for temporally static-data with binary weight updates.

# 2. RELATED WORK

## 2.1. Spiking Neural Networks

### 2.1.1. Leaky-Integrate-and-Fire Model

Spiking neural networks are widely used in the neuroscience community to build biologically plausible models of neuron populations in the brain. These models have been designed to reproduce information propagation and temporal dynamics observable in cortical layers. As many models exists, from the most simple to the most realistic, we will focus on the Leaky-Integrate-and-Fire model (LIF), a simple and fast model of a spiking neuron.

LIF neurons are asynchronous units receiving input signals called spikes from pre-synaptic cells. Each spike $x_i$ is modulated by the weight $w_i$ of the corresponding synapse and added to the membrane potential $u$. In a synchronous formalism, at each time step, the update of the membrane potential at time $t$ can be expressed as follow:

$$\mathcal{T}\frac{\delta u(t)}{\delta t} = -(u(t) - u_{res}) + \sum_{i=1}^{n} w_i x_{i,t} \tag{1}$$

Where $\mathcal{T}$ is the time constant of the neuron, $n$ the number of afferent cells and $u_{res}$ is the reset potential (which we also consider as the initial potential at $t_0 = 0$).

When $u$ reaches a certain threshold $T$, the neuron emits a spike to its axons and resets its potential to its initial value $u_{res}$.

This type of network has proven to be energy-efficient Gamrat et al. (2015) on analog devices due to its asynchronous and sparse characteristics. Even on digital synchronous devices, spikes can be encoded as binary variables, therefore carrying maximum information over the minimum memory unit.

### 2.1.2. Rank Order Coding Network

A model which fits the criteria of processing speed and adaptation to images data is the rank order coding SNN (Thorpe et al., 2001). This type of network processes the information with single-step feed-forward information propagation by means of the spike latencies. One strong hypothesis for this type of network is the possibility to compute information with only one spike per neuron, which has been demonstrated in rapid visual categorization tasks (Thorpe et al., 1996). Implementations of such networks have proven to be efficient for simple categorization tasks like frontal-face detection on images (Van Rullen et al., 1998; Delorme and Thorpe, 2001).

The visual-detection software engine SpikeNet Thorpe et al. (2004) is based on rank order coding networks and is used in industrial applications including face processing for interior security, intrusion detection in airports and casino games monitoring. Also, it is able to learn new objects with a single image, encoding objects with only the first firing spikes.

The rank order model SpikeNet is based on a several layers architecture of LIF neurons, all sharing the time constant $\mathcal{T}$, the reset potential $u_{res}$ and the spiking threshold $T$. During learning, only the first time of spike of each neuron is used to learn a new object. During inference, the network only needs to know if a neuron has spiked or not, hence allowing the use of a binary representation.

## 2.2. Learning With Spiking Neural Networks

### 2.2.1. Deep Neural Networks Conversion

The computational advantages of SNNs led some researchers to convert fully learned deep neural networks into SNNs (Diehl et al., 2015, 2016), in order to give SNNs the inference performance of back-propagation trained neural networks.

However, deep neural networks use the back-propagation algorithm to learn the parameters, which remains a computationally heavy algorithm, and requires enormous amounts of labeled data. Also, while some researches hypothesize that the brain could implement back-propagation (Bengio et al., 2015), the biological structures which could support such error transmission process remain to be discovered. Finally, unsupervised learning within DNNs remains a challenge, whereas the brain may learn most of its representations through unsupervised learning (Turk-Browne et al., 2009). Suffering from both its computational cost and its lack of biological plausibility, back-propagation may not be the best learning algorithm to take advantage of SNNs capabilities.

On the other hand, researches in neuroscience have developed models of unsupervised learning in the brain based on SNNs. One of the most popular model is the STDP.

### 2.2.2. Spike Timing Dependent Plasticity

Spike-Timing-Dependent-Plasticity is a biological learning rule which uses the spike timing of pre and post-synaptic neurons to update the values of the synapses. This learning rule is said to be Hebbian ("What fires together wires together").

Synaptic weights between two neurons updated as a function of the timing difference between a pair or a triplet of pre and post-synaptic spikes. Long-Term Potentiation (LTP) or a Long-Term Depression (LTD) are triggered depending on whether a presynaptic spike occurs before or after a post-synaptic spike, respectively.

Formulated two decades ago by Markram et al. (1997), STDP has gained interest in the neurocomputation community as it allows SNN to be used for unsupervised representation learning (Kempter et al., 2001; Rao and Sejnowski, 2001; Masquelier and Thorpe, 2007; Nessler et al., 2009). The features learnt in low-level layers have also been shown to be relevant for classification tasks combined with additional supervision processes in the top layers (Beyeler et al., 2013; Mozafari et al., 2017). As such STDP may be the main unsupervised learning mechanisms in biological neural networks, and shows nearly equivalent mathematical properties to machine learning approaches such as auto-encoders (Burbank, 2015) and non-negative matrix factorization (Carlson et al., 2013; Beyeler et al., in review).

We first consider the basic STDP pair-based rule from Kempter et al. (2001). Each time a post synaptic neuron spikes, one computes the timing difference $\Delta t = t_{pre} - t_{post}$ (relative to each presynaptic spike) and updates each synapse $w$ as follows:

$$\Delta w = \begin{cases} A^+ . e^{\frac{\Delta t}{\mathcal{T}_+}} \text{ if } \Delta t < 0 \\ A^- . e^{\frac{\Delta t}{\mathcal{T}_-}} \text{ otherwise} \end{cases} \quad (2)$$

where $A_+ > 0, A_- < 0$, and $\mathcal{T}_+, \mathcal{T}_- > 0$. The top and bottom terms in this equation are respectively the LTP and LTD terms.

This update rule can be made highly computationally efficient by removing the exponential terms $e^{\frac{\Delta t}{\mathcal{T}}}$, resulting in a simple linear time-dependent update rule.

Parameters $A_+$ and $A_-$ must be tuned on order to regularize weight updates during the learning process. However in practice, tuning these parameters is a tedious task. In order to avoid weight divergences, networks trained with STDP learning rule should also implement stability processes such as refractory periods, homoeostasis with weight normalization or inhibition. Weight regularization may also be implemented directly by reformulating the learning rule equations. For instance in Masquelier and Thorpe (2007), the exponential term in Equation (2) is replaced by a process which guaranties that the weights remain in the range [0...1] :

$$\Delta w = \begin{cases} A_+ . w . (1 - w) \text{ if } \Delta t < 0 \\ A_- . w . (1 - w) \text{ otherwise} \end{cases} \quad (3)$$

Note that in Equation (3), the amplitude of the update is independent from the absolute time difference between pre and post-synaptic spikes, which only works if pairs of spikes belongs to the same finite time window. In Masquelier and Thorpe (2007) this is guaranteed by the whole propagation schemes, which is applied on image data and rely on a single feedforward propagation step taking into account only one spike per neuron. Thus the maximum time difference between pre and post-synaptic spikes is bounded in this case.

## 2.3. Regulation Mechanisms in Neural Networks

### 2.3.1. WTA as Sparsity Constrain in Deep Neural Networks

Winner-takes-all (WTA) mechanisms are an interesting property of biological neural networks which allow a fast analysis of objects in exploration tasks. Following de Almeida et al. (2009), gamma inhibitory oscillations perform a WTA mechanism independent from the absolute activation level. They may select the principle neurons firing during a stimulation, thus allowing, e.g., the tuning of narrow orientation filters in V1.

WTA has been used in deep neural networks in Makhzani and Frey (2015) as a sparsity constraint in autoencoders. Instead of using noise or specific loss functions in order to impose activity sparsity in autoencoder methods, the authors propose an activity-driven regularization technique based on a WTA operator, as defined by Equation (4).

$$WTA(X, d) = \begin{cases} X_j \text{ if } |X_j| = \underset{k \in d}{max}(|X_k|) \\ 0 \text{ otherwise} \end{cases} \quad (4)$$

where $X$ is a multidimensional matrix and $d$ is a set of given dimensions of $X$.

After definition of a convolutional architecture, each layer is trained in a greedy layer-wise manner with representation from the previous layer as input. To train a convolutional layer, a WTA layer and a deconvolution layer are placed on top of it. The WTA layer applies the WTA operator on the spatial dimensions of the convolutional output batch and retains only the $n_p$% first activities of each neuron. This way for a given layer with $N$ representations map per batch and $C$ output channels, only $N.n_p.C$ activities are kept at their initial values, all the others activation values being zeroed. Then the deconvolutional layer attempts to reconstruct the input batch.

While this method demonstrates the potential usefulness of WTA mechanisms in neural networks, it still relies on computationally heavy backpropagation methods to update the weights of the network.

### 2.3.2. Homosynaptic and Heterosynaptic Homeostasis

In their original formulation, Hebbian-type learning rule (STDP, Oja rule , BCM rule) does not have any regulation process. The absence of regulation in synaptic weights may impact negatively the way a network learns. Hebbian learning allows the synaptic weights to grow indefinitely, which can lead to abnormally high spiking activity and neurons to always win the competitions induced by inhibitory circuits.

To avoid such issues, two types of homeostasis have been formulated.

Homosynaptic homeostasis acts on a single synapse and is depends on its respective inputs and outputs activity only. This homeostatic process can be modeled with a self-regulatory term in the Hebbian rule as in Masquelier and Thorpe (2007) or as a synaptic scaling rule depending on the activity driven by the synapse as in Carlson et al. (2013).

Heterosynaptic homeostasis is a convenient way to regulate the synaptic strength of a network. The model of such homeostasis takes into account all the synapses connected to a given neuron, all the synapses in a layer (like the L2 loss weight decay in deep learning) or at the network scale. Biological plausibility of such process is still discussed. Nevertheless, some evidences of heterosynaptic homeostasis have been observed in the brain to compensate runaway dynamics of synaptic strength introduced by Hebbian learning (Royer and Paré, 2003; Chistiakova et al., 2014). It then plays an important role in the regulation of spiking activity in the brain and is complementary to homosynaptic plasticity.

## 2.4. Neural Networks and Image Processing

Image processing with neural networks is performed with multiple layers of spatial operations (like convolutions, pooling, and non-linearities), giving the name Deep Convolutional Neural Networks to these methods. Their layer architecture is directly inspired from the biological processes of the visual cortex, in particular from the well known HMAX model (Riesenhuber and Poggio, 1999), except that the layers' weights are learnt with back-propagation. Deep CNN models use a single-step forward propagation to perform a given task. Even if convolutions on large maps may be computationally heavy, all the computations are done through only one pass in each layer. One remaining advantage of CNNs is their ability to learn from raw data, such as pixels for images or waveforms for audio.

On the other hand, since SNNs use spikes to transmit information to the upper layers, they need to perform neuron potential updates at each time step. Hence, applying such networks with a convolutional architecture requires heavy computations once for each time step. However, spikes and synaptic weights may be set to a very low bit-resolution (down to 1 bit) to reduce this computational cost Thorpe et al. (2004). Also, STDP is known to learn new representations with a few iterations Masquelier et al. (2009), theoretically reducing the number of epochs required to converge.

# 3. CONTRIBUTION

Our goal here is to apply STDP in a single-step feed-forward formalism directly from raw data, which should be beneficial in the cases where training times and data labeling are issues. Thus we may select a neural model which combines the advantages of each formalism in order to reduce the computational cost during both training and inference.

## 3.1. Feedforward Network Architecture

### 3.1.1. Neural Dynamics

Here, we will consider the neural dynamics of a spiking LIF network in presence of image data. Neural updates in the temporal domain in such neural architecture are as defined by Equation (1).

Since a single image is a static snapshot of visual information, all the $x_{i,t}$ are considered constant over time. Hence $\sum_{i=1}^{n} w_i.x_{i,t}$ is also constant over time under the assumption of static synaptic weights during the processing of the current image.

Let us define $v_{in} = \sum_{i=1}^{n} w_i.x_{i,t}, \forall t$ the total input signal to the neuron. Let us also determine $u(t_0 = 0) = u_{res}$ as an initial condition. As $v_{in}$ is constant over time, we can solve the differential equation of the LIF neuron, which gives:

$$\begin{aligned} \mathcal{T}\frac{\delta u(t)}{\delta t} &= -(u(t) - u_{res}) + v_{in} \\ \Rightarrow \quad u(t) &= -v_{in}.e^{\frac{-t}{\mathcal{T}}} + u_{res} + v_{in} \qquad \forall t > 0 \end{aligned} \tag{5}$$

The precise first spike-time of a neuron given its spiking threshold $T$ is given by :

$$t_s = -\mathcal{T}.log(1 + \frac{u_{res} - T}{v_{in}}) \tag{6}$$

Since Equation (6) decreases monotonically wrt. $v_{in}$, we can recover the intensity-latency equivalence. The relative order of spike-times is also known since $v_{in,1} > v_{in,2} \rightarrow t_{s,1} < t_{s,2}$.

### 3.1.2. Equivalence With Artificial Neuron With ReLU Activation

Thus from Equation (6), for each neuron we can determine the existence of a first spike, along with its precise timing. Hence, since we are only concerned with the relative times of first spikes across neurons, one can replace the computation at each time-step by a single-step forward propagation given the input intensity of each neuron.

The single-step forward propagation correspond to LIF integration when $t \rightarrow \infty$. As we are first looking for the existence of any $t_s$ such that $u(t_s) > T$:

$$\begin{aligned} \lim_{t\to\infty} u(t) - T &= \lim_{t\to\infty} -v_{in}.e^{\frac{-t_s}{\mathcal{T}}} + u_{res} + v_{in} - T \\ &= u_{res} + v_{in} - T \end{aligned} \tag{7}$$

Having $\quad v_{in} = \sum_{i=1}^{n} w_i.x_i \quad$ and $\quad b = u_{res} - T,$

$$\lim_{t\to\infty} u(t) - T = b + \sum_{i=1}^{n} w_i.x_i \tag{8}$$

which is the basic expression of the weighted sum of a perceptron with bias.Also, $t_s$ exists if and only if $b + \sum_{i=1}^{n} w_i.x_i > 0$, which shows the equivalence between LIF neurons with constant input at infinity and the artificial neuron with rectifier activation function (ReLU).

This demonstration can be generalized to local receptive fields with weight sharing, and thus we propose to replace the time-step computation of LIF neurons, by common GPU optimized routines of deep learning such as 2D convolutions and ReLU non-linearity. This allows us to obtain in a single-step all the first times of spikes -inversely ordered by their activation level- and nullified if no spike would be emitted in an infinite time. Moreover, these different operations are compatible with mini-batch learning. Hence, our model is also capable of processing

several images in parallel, which is an uncommon feature in STDP-based networks.

#### 3.1.3. Winner-Takes-All Mechanisms

Following the biological evidence of the existence of WTA mechanisms in visual search tasks (de Almeida et al., 2009) and the code sparsity learned with such processes (Makhzani and Frey, 2015), we may take advantage of WTA to match the most repetitive patterns in a given set of images. Also, having to learn only these selected regions should drastically decrease the number of computations required for the learning phase (compared to dense approaches in deep learning and SNN simulations). Inspired by this biological mechanism, we propose to use three WTA steps as sparsifying layers in our convolutional SNN architecture.

The first WTA step is performed on feature neighborhood with a max-pooling layer on the convolution matrix with kernel size $k_{pool} >= k_{conv}$ and stride $s_{pool} = k_{conv}$. This acts as a lateral inhibition, avoiding the selection of two spikes from different kernels in the same region.

Next we perform a WTA step with the WTA operation (Equation 4) on the channel axis for each image (keeping at each pooled pixel, the neuron that spikes first). This forces each kernel to learn from different input patches.

The third WTA step is performed with WTA operation on spatial axes as in Makhzani and Frey (2015). This forces the neuron to learn from the most correlated patch value in the input image.

The WTA operation (Equation 4) is not to be confused with the Maxout operation from Goodfellow et al. (2013) and the max pooling operation, since these latter squeeze the dimensions on which they are applied, while the WTA operation preserves them.

Then we extract the indexes of the selected outputs along with their sign and their corresponding input patch. Extracted input patches are organized in $k$ subsets, each subset corresponding to one output channel. These matrices will be refered to as follow :

- $Y_k$ : matrices of selected outputs, of dimension $(m_k, c_{out})$
- $X_k$ : matrices of selected patches, of dimension $(m_k, c_{in} \times h_{in} \times w_{in})$
- $W$ : matrices of filters, of dimension $(c_{in} \times h_{in} \times w_{in}, c_{out})$

with $m_k$ the number of selected indexes and patches for neuron $k \in [1...c_{out}]$, $c_out$ the number of channels (or neurons) of the output layer, and $c_in, h_in, w_in$ are the receptive field size (resp. channel, height and width). Note that at most one output is selected per channel and per image, $m_k \leq N$.

The WTA in our model has two main advantages. First, it allows the network to learn faster on only a few regions of the input image. Second, classical learning frameworks use the mean of weights gradient matrix to update the synaptic parameters. By limiting the influence of averaging on the gradient matrix, synaptic weights are updated according to the most extreme values of the input, which allow the network to learn sparse features.

Note that the network is able to propagate relative temporal information through multiple layer, even though presented inputs lack this type of data. It is also able to extract regions which are relevant to learn in terms of information maximization. The full processing chain for propagation and WTA is shown in **Figure 1**.

### 3.2. Binary Hebbian Learning

#### 3.2.1. Simplifying the STDP Rule

Taking inspiration from the STDP learning rule, we propose a Hebbian correlation rule which follows the relative activations of input and output vectors.

Considering the input patch value $x_{n,i} \in X_n, n \in [1...m_k], i \in [1...c_{in} \times h_{in} \times w_{in}]$, the corresponding weight value $w_{k,i}$, the selected output value $y_k \in Y_k$ and a heuristically defined threshold $T_l$, the learning rule is described in Equation (9).

$$\Delta w_{k,i} = \begin{cases} sign(x_{n,i}).sign(y_k) \text{ if } |x_{n,i}| > T_l \\ -sign(w_{k,i}) \text{ otherwise} \end{cases} \tag{9}$$

The learning rule is effectively Hebbian as shown in the next paragraph and can be implemented with lightweight operations such as thresholding and bit-wise arithmetic.

Also, considering our starting hypotheses, where we limit to one the number of spikes per neuron during a full propagation phase for each image, it is guaranteed that, for any pair of pre and post-synaptic neuron, the choice of LTP or LTD exist and is unique for each image presentation. These hypotheses are similar to the ones in Masquelier and Thorpe (2007), where these conditions simulates a single wave of spikes within a range of 30 ms.

#### 3.2.2. Equivalence to Hebbian Learning in Spiking Networks

In this section we show the Hebbian behavior of this learning rule. For this, we first focus on the "all positive case" $(x, y, w \in R+)$ and will explain in the next section the extension to symmetrical neurons.

In the case of "all positive," the Equation (9) can be rewritten as Equation (10).

$$\Delta w_{k,i} = \begin{cases} 1 \text{ if } x_{k,i} > u(t_{post}) \\ -1 \text{ otherwise} \end{cases} \tag{10}$$

This rule tends to increase the weights when the input activity is greater than a threshold (here the post-synaptic neuron firing threshold), and decreases it otherwise.

Equation (10) is equivalent to the pair-based STDP rule presented in Equation (2) removing the exponential term and using $A_+ = 1$ and $A_- = -1$.

#### 3.2.3. Extension to Symmetric Neurons

We have demonstrated that the proposed learning rule is effectively Hebbian in the case where $x, w, y \in \mathbb{R}_+$. Our learning rule also takes into account negative values of $x, w, y$. In biological networks models, negative values do not seem to make much

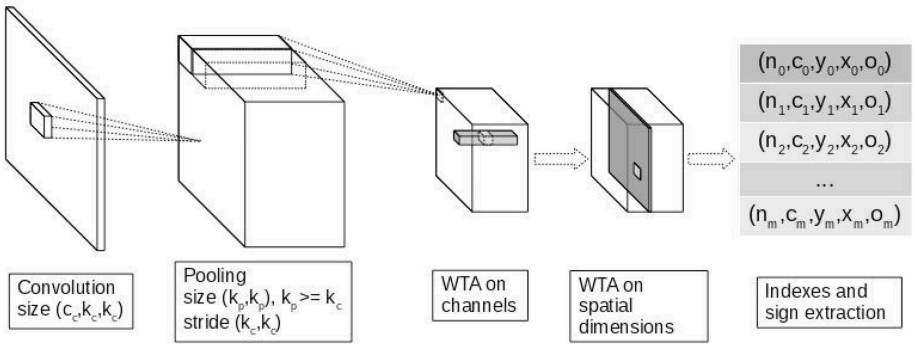


**FIGURE 1 |** Processing chain for the region WTA.

**TABLE 1 |** Weight update given x, y, and w following the proposed learning rule (Equation 9).

| | $x < -T$ | $-T < x < T$ | $x > T$ |
|---|---|---|---|
| $y > 0$ | $-1$ | $-sign(w)$ | $+1$ |
| $y < 0$ | $+1$ | $-sign(w)$ | $-1$ |

sense since firing rates and synaptic conductance are expressed in units defined only in $\mathbb{R}_+$.

Nevertheless, negative values are used in many spiking networks models in the very first layer of visual features. For instance, ON-centered-OFF-surround and OFF-centered-ON-surround filters (also known as *Mexican hat* filters) are often used to pre-process an image in order to simulate retinal cells extracting gradients. These two filters are symmetric with respect to the origin. Hence a common computational optimization is to apply only one of the two filters over the image, separating negative and positive resulting values as OFF and ON activities, respectively.

We extend this computational trick to neurons in any neural layer under the hypothesis that negative values for $x, w, y$ corresponds to activities and weights of synaptically symmetric neurons. For a neuron with constant input activity $X$ and synaptic weights $W$ of size $n$, we can express its output activity $y = \sum_{i=1}^{n} X_i \times W_i$. If $y < 0$, we can convert it to a positive value using the synaptically opposite weights $\sum_{i=1}^{n} X_i \times -W_i = -y$.

Under the hypothesis of the existence of a pair-wise competition between neurons with symmetric weights (for instance with inhibition), this computational trick remains biologically plausible.

Considering now the proposed learning rule, the weights update given $x, y$, and $w$ is shown in **Table 1**. In this table, the first spikes ($|x| > T$) will induce an update of the weight to increase the $|y|$ ($\Delta w = sign(y).sign(x)$). Meanwhile, the weights corresponding to the last spike will be reduced ($\Delta w = -sign(w)$).

With this framework the choice of the parameter $T_l$ is critical. Thanks to the WTA mechanism developed, the selection of a neuron for learning is performed disregarding its firing threshold $T$, set to zero in practice. Hence contrary to Masquelier and Thorpe (2007), we cannot rely on the precise firing threshold of the neuron. In order to approximate this threshold, we developed two strategies described in the next paragraphs. These strategies are made adaptative such that the learning rule can be invariant to contrast variation. Also the adaptative behavior of this threshold avoids to tune an additional parameter in the model.

### 3.2.4. Hard Percentile Threshold

The first strategy applied follows the STDP learning rule, which fixes a time constant for LTP and LTD. In our framework this is implemented as a percentile of the input activity to map their influence in the spike. For each input vector $x_n \in X_k \forall k$ , we compute the patch threshold $T_l$ as the minimum value in the local $p_{n\%}$ percentile. $p_{n\%}$ is manually set and global for all the patches.

$$\Delta w_{k,i} = \begin{cases} -sign(w_{k,i}) \text{ if } |x_{n,i}| \leq p_{n\%} \\ sign(x_{n,i}).sign(y_k) \text{ otherwise} \end{cases} \tag{11}$$

However, we have seen experimentally that the threshold tuning may be cumbersome. As it regulates the sparsity of the synaptic weight matrix, fixing the sparsity manually may lead to unsatisfying results. Also, getting the percentiles uses the index-sorting operation which is time consuming.

### 3.2.5. Average Correlation Threshold

We propose a second strategy which relies on the computation of an adaptative threshold between LTP and LTD. For each input vector $x_n \in X_k \forall k$ we compute the sign correlated input activation as $\hat{x_{n,i}} = x_{n,i}.sign(w_k).sign(y_k)$. Next we compute the threshold $T_l$ as the mean of $\hat{x_n}$. Then we apply the learning rule in Equation (9).

With this strategy, the learning rule is also equivalent to Equation (12), which is straightforward to implement since it avoids conditional branching.

$$\Delta w_{k,i} = sign(x_{n,i}.sign(y_k).sign(w_{k,i}) - T_l).sign(w_{k,i}) \tag{12}$$

Using the mean sign corrected input activation as a threshold, the model is able to be invariant to local contrasts. It also requires the

calculation of the mean and a thresholding, two operations that are much faster than sorting. Finally, the adaptative behavior of such a threshold automate the sparsity of synaptic weights.

#### 3.2.6. Computing Updates From a Batch of Images

Since our method allows the propagation of several images at the same time through mini-batch, we can also adapt our learning rule when batches of images are presented. Since biological visual systems never deal with batches of dozen images at once, the following proposal is a computational trick to accelerate the learning times, not a model of any existing biological feature.

When all the update vectors have been computed, the weight update vector for the current batch is obtained through the binarization of the sum of all the update vector for the corresponding kernel. We finally modulate the update vector with a learning rate $\lambda$.

$$U_{n,i} = \sum_{k=1}^{m_k} \Delta w_{k,i} \tag{13}$$

$$\Delta W_{k,i} = \begin{cases} -1 \text{ if } U_{n,i} \leq 0 \\ 1 \text{ otherwise} \end{cases} \tag{14}$$

$$W_{k,i} = W_{k,i} + \lambda . \Delta W_{k,i} \tag{15}$$

#### 3.2.7. Weight Normalization Through Simple Statistics

Since each update step adds $+\lambda$ or $-\lambda$ to the weights, a regularization mechanism is required to avoid the weights growing indefinitely. Also we want to maintain a fair competition between neurons of the same layer, thus the total energy of the weights should be the same for all the neurons.

We propose a simple model of heterosynaptic homeostasis in order to regulate the weights of each neuron.We chose to normalize the weights of each neuron $k$ by mean centering and standardization by variance. Hence, after each update phase, the normalization is done as follows :

$$W_k = \frac{W_k - \mu(W_k)}{\sigma^2(W_k)} \tag{16}$$

This way, even neurons which did not learn a lot during the previous epochs can win a competition against the others. In practice, we set $\lambda$ in an order of magnitude of $10^{-1}$ and halved it after each epoch. Given the order of magnitude of $\lambda$ and the unit variance of $W_k$, we know that ninety-five percent of the weights belongs to the interval $[-1.5...1.5]$. In fact, only a few batches of images are necessary to modify the influence of a given afferent. Two neurons responding to a similar pattern can thus diverge and specialize on different patterns in less than a dozen training batches.

As a detail, if the WTA region selected is small, some neurons may learn parts of patterns already learned by an other one. Since $\sigma^2(W_k) = 1$ and most of the weights are equal to zero, the values of the remaining weights would grow very large. This can end up in multiple neurons learning almost identical patterns. We have observed that clipping weights after normalization between the range $[-2...2]$ prevents this situation.

### 3.3. Multi-layer Architectures With Binary STDP

This proposed approach is able to learn a multi-layer convolutional architecture as defined by the user. It does not require a greedy layer-wise training, all the convolutional layers can be trained in parallel. We can optionally apply a non-linearity, a downsampling operation or a normalization after each convolution layer.

Once all the features layers have learned, the whole features architecture can process images as a classical convolutional neural network in order to obtain the new representations.

## 4. EXPERIMENTS AND RESULTS

### 4.1. Method

The proposed method learns, unsupervised, convolutional features from image data. In order to validate our approach, we evaluated the learnt features on four different classification datasets : MNIST, ETH80, CIFAR10, and STL10. Architectures and hyper-parameters were tuned separately for each dataset, details being given in the relevant sections.

The overall evaluation method remains the same for each dataset. The proposed framework will be used to learn one or several convolutional layer with the simplified STDP. In order to show the faster convergence of features with our method, we will only train these layer with a subset of the full training dataset with very few epochs.

Once the features are learnt, we show qualitatively the learnt features for each dataset. To quantitatively demonstrate their relevance, we use the extracted features as input to a supervised classifier. Although as state of the art classification are deep learning systems, we use a simple Multi-Layer Perceptron (MLP) with zero, one, or two hidden layers (depending on the dataset) taking as inputs the learnt features with the proposed solution.

For all the experiments, we started with a lightweight network architecture (the simplest available in the literature if available), and incrementally added complexity until further additions stopped improving performance. The classifier on top of the network starts as linear dense layer with as many neurons as the number of classes, and is complexified with intermediate layers as the architectural-tuning goes on.

We compare our results with other state of the art unsupervised feature learning methods specific for each dataset.

### 4.2. MNIST

The MNIST dataset contains 60,000 training images and 10,000 testing images of size $28 \times 28$ containing handwritten digits from 0 to 9. MNIST digits are written in white on a black background, hence pixel values are distributed across two modes. Considering the data distribution and the limited number of classes, MNIST may be considered as an easy classification task for current state-of-the-art methods. As a matter of fact, neural based methods do not need deep architectures in order to perform well on this dataset. Light-weight architectures can be defined in order to explore issues with the developed method. Once the method has satisfying results on MNIST, more complex datasets may be tackled.


**FIGURE 2 |** Architecture of the network in the MNIST experiment.

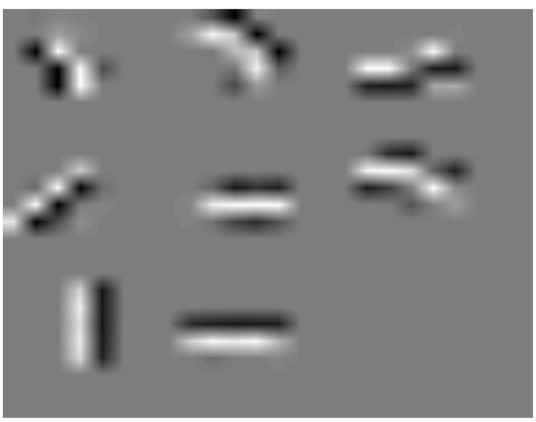

**FIGURE 3 |** Eight 5 × 5 features learned from MNIST dataset on raw images.

To perform classification on this dataset, we defined a lightweight convolutional architecture of features close to LeNet LeCun et al. (1998), presented in **Figure 2**. Since achieving high classification accuracy on MNIST is easy with a high number of neurons per layer, the number of neurons per layer was kept as low as possible in order to actually verify the relevance of the features.

Unsupervised learning was performed over only 5,000 random images from the dataset for 5 epochs, which only represents 25,000 image presentations. A visualization of the learnt features is shown in **Figure 3**.

Once the features were learnt, we used a two-hidden layers MLP to perform classification over the whole transformed training set. The learnt features and classifier were then run on all the testing set images in order to get the test error rate.

Classification performances are reported in **Table 2**. While the best methods in the state-of-the-art reach up to 99.77% accuracy, we did not report these results since these approaches use supervised learning with data augmentation, which is outwith the scope of this paper. All the reported results were obtained without data augmentation and using unsupervised feature learning.

**TABLE 2 |** MNIST accuracy.

| Method | Accuracy (%) |
|---|---|
| SDNN (Kheradpisheh et al., 2016) | 98.40 |
| Two layer SNN (Diehl and Cook, 2015) | 95.00 |
| PCA-Net (Chan et al., 2014) | 98.94 |
| Our method | 98.49 |

Our approach performs as well as SDNN since they are structurally close, reaching state-of-the-art performance without fine-tuning and data-augmentation. While PCA-Net has better performance, learning was done on twice the number of samples we used. Doubling the number of samples to match the same number used for PCA-Net (10,000) did not improve the performance of our method.

## 4.3. ETH80

The ETH80 (Leibe and Schiele, 2003) contains 3,280 color images of eight different object categories (apple, car, cow, cup,dog, horse, pear, tomato). Each category contains 10 different object instances taken from 41 points of view. This dataset is interesting since the number of available images is limited and contains a lot of variability in 3D rotations. It allows us to evaluate the generalization potential of the features and their robustness to changes in viewpoint.

As the number of samples is restrained here, we performed both unsupervised and supervised learning on half the dataset (1,640 images chosen randomly). The other half was used as the test set.

We compare our approach to the classical HMAX model and to Kheradpisheh et al. (2016). The architectures for unsupervised and supervised part are shown in **Figure 4**. Learning visual features becomes more and more difficult with the proposed method as we add convolutional layers on top of the network. Since ETH80 images are large (96 × 96), we apply pooling with a stride of 4 in order to quickly reduce the dimensions over the hierarchy.

Results are reported in **Table 3**. While our approach does not reach the same performance as Kheradpisheh et al. (2016), it is able to learn features relevant for a classification task with multiple points of view of different objects.

## 4.4. CIFAR-10

The CIFAR-10 dataset (Krizhevsky, 2009) is a dataset for classification of natural images from 10 classes (airplane, automobile, bird, cat,deer, dog, frog, horse, ship, and truck). The dataset is split into three with 60,000 training, 10,000 validation, and 10,000 testing images. Images are a subset of the 80 million tiny images dataset (Torralba et al., 2008). All the images are 32 × 32 pixels size with three color channels (RGB).

This dataset is quite challenging, since it contains many variations of objects with natural backgrounds, in low resolution.

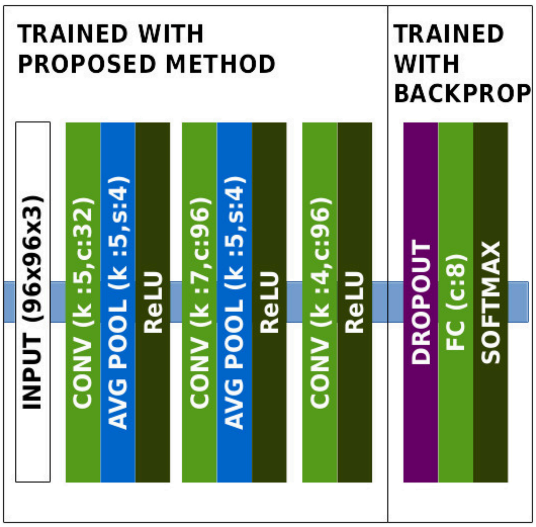


**FIGURE 4 |** Architecture of the network in the ETH80 experiment.

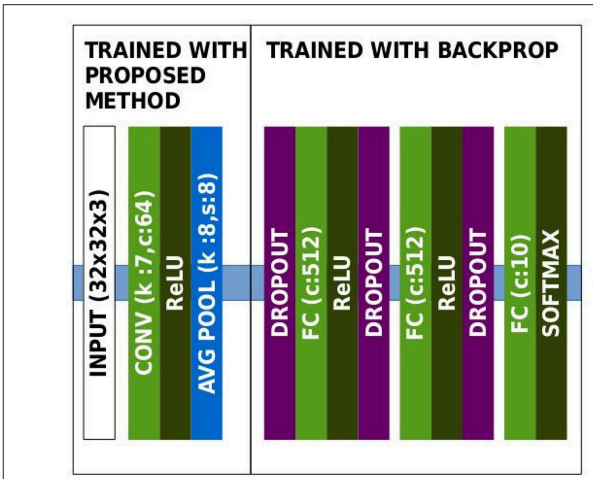


**FIGURE 5 |** Architecture of the network in the CIFAR-10 experiment.

**TABLE 3 |** ETH80 results.

| Method | Accuracy (%) |
|---|---|
| HMAX (Riesenhuber and Poggio, 1999) | 69.0 |
| SDNN (Kheradpisheh et al., 2016) | 82.8 |
| Our method | 75.2 |

Hence in order to tackle this dataset, algorithms must be able to find relevant information in noisy data.

The architecture used for this dataset is given in **Figure 5**. Learnt features are shown in **Figure 6A**. We observe that the features are similar to oriented-gabor features, which is consistent with the results of other unsupervised methods such as *k*-means and RBM. Also the weights distribution displayed in **Figure 6B** contains a majority of values close to zero, showing the sparsity of the features. Performances obtained on CIFAR-10, along with other methods evaluation, are shown in **Table 4**.

As a performance baseline, we also trained the MLP with the same architecture but keeping the convolutional layer's weights randomly initialized and frozen. The increase of 17% of classification rate proves the usefulness of the features learnt with our method in the classification process.

Only a few works related to SNNs have been benchmarked on CIFAR-10. Cao et al. (2015) and Hunsberger and Eliasmith (2015) rely on convolutional to SNN conversion to perform supervised learning on the dataset. Panda and Roy (2016) built a convolutional feature hierarchy on the principle of auto-encoders with SNNs, and classified the top level representations with an MLP.

Also, some works unrelated to SNNs are worth comparing here. Coates et al. (2011) benchmarked four unsupervised feature learning methods (k-means, triangle k-means, RBM, and sparse auto-encoders) with only one layer. Results from the PCA-Net approach are also included.

Our approach reached good performance given the lightweight architectures and the limited number of samples. It outperforms the CNN with 64 random filters, confirming the relevance of the learnt features for classification, and also the Triangle K-means approach with 100 features. Empirically however, training with more samples without increasing the number of features does not improve the performance.

Also, due to the low resolution of CIFAR-10 images, we tried to add a second convolutional layer. The learnt filters in this new layer were very redundant and led to the same performance observed with only one layer. Further investigations might explore ways to force layers above the first to learn more sparse features.

## 4.5. STL-10

STL-10 is a dataset dedicated to unsupervised feature learning. Images were taken from the ImageNet dataset. The training set contains 5,000 images labeled over the same ten classes as CIFAR-10. An unlabeled training set of 100,000 images is also provided. Unlabeled images may contain objects from other classes of ImageNet (like bear, monkeys, trains...). The testing set contains 8,000 images (800 per class). All images are in RGB format with a resolution of 96 × 96.

We applied the same architecture as for the CIFAR-10 dataset, except the average pooling layer was done over 24 × 24 sized windows (in order to have the same 4 × 4 output dimension). As before, we limited the number of samples during the unsupervised learning step to 5,000.

While some works related to SNNs or STDP have been benchmarked on CIFAR-10, we were not able to find any using the STL-10 dataset. Hence our approach may be the first biologically inspired method trying to tackle this dataset.

Our approach reaches 60.1% accuracy on STL-10, which is above the lower-bound performance on this dataset. Performances obtained by other unsupervised methods range between 58 and 74%.

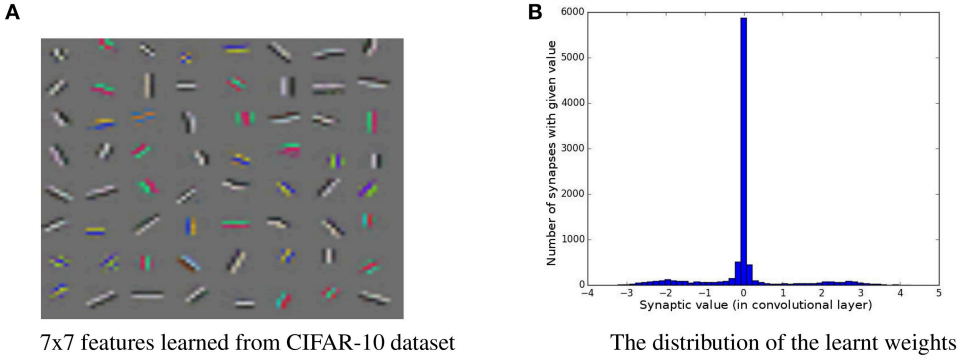


**FIGURE 6 | (A)** Sixty-four filters of size 7 × 7 learned with our method on the CIFAR-10 dataset. **(B)** The weights distribution of the network's first layer trained on CIFAR-10.

**TABLE 4 |** CIFAR-10 results.

| Method | Unsupervised | Training samples | Accuracy (%) |
|---|---|---|---|
| Triangle k-means (1,600 features) (Coates et al., 2011) | Yes | 50,000 | 79.6 |
| Triangle k-means (100 features) (Coates et al., 2011) | Yes | 50,000 | 55.5 |
| PCA-Net (Chan et al., 2014) | Yes | 50,000 | 78.67 |
| LIF CNN (Hunsberger and Eliasmith, 2015) | No | 50,000 | 82.95 |
| Regenerative Learning (Panda and Roy, 2016) | Yes | 20,000 | 70.6 |
| Our method (64 features) | Yes | 5,000 | 71.2 |
| CNN random frozen filters | No | 50,000 | 55.3 |

## 5. DISCUSSION

The proposed approach is able to train lightweight convolutional architectures based on LIF neurons which can be used as a feature extractor prior to a supervised classification method. These networks achieve average levels of performance on four image classification datasets. While the performances are not as impressive as the ones obtained with fully supervised learning methods, where features are learnt specifically for the classification task, interesting characteristics emerge from this model.

By showing the equivalence between rank-order LIF neurons and perceptrons with ReLU activation, we were able to borrow computationally efficient concepts from both neuroscience and machine learning literature while remaining biologically plausible enough to allow the conversion of network trained this way to be converted into SNN.

Binary STDP along with WTA and synaptic normalization reduces drastically the process of parameters tuning compared to other STDP approaches. LIF neurons require the tuning of their respective time constant. STDP also requires four parameters to be tuned : the time constants $\mathcal{T}_+$ and $\mathcal{T}_-$ as well as the LTP and LTD factors $A_+$ and $A_-$ for each layer. Our model of binary STDP on the other hand only needs to set its learning rate $\lambda$, set globally for the whole architecture.

Another advantage over other STDP approaches is the ability to train the network with multiple images in parallel. While this ability is biologically implausible, it can become handy in order to accelerate the training phase thanks to the intrinsic parallel optimization provided by GPU. Also, the equivalence between LIF neurons and perceptrons with ReLU activation in presence of images allows us to perform the full propagation phase of a SNN in one shot, and to apply our STDP rule without the need of interpolation precise timing information from the image. Other approaches using SNNs with STDP requires the interpolation of temporal information from the image (Masquelier and Thorpe, 2007; Kheradpisheh et al., 2016), with gabor filters for instance, in order to generate spike trains. This way, STDP can be applied to learn the correlations between spike timings.

From a deep learning point of view, the main interest of our model resides in the proposal of a backpropagation-free training procedure for the first layers. As the backward pass in deep neural networks implies computationally heavy deconvolutions to compute the gradients of the parameters, any prior on visual modelization which can avoid a backpropagation over the whole network may help to reduce the computational overhead of this step. The LIF-ReLU equivalence demonstrated allows a convolutional network to take advantage of the inherent characteristic of STDP to quickly find repeating pattern in an input signal (Masquelier and Thorpe, 2007; Masquelier et al., 2009; Nessler et al., 2009).

With the WTA scheme proposed, we made the assumption that relevant visual information resides in the most contrasted patches. It also imposes the neurons to learn a sparse code with the combination of neighburhood and channel-wise inhibition. Such hard-coded WTA led to first layers features very similar to the gabor-like receptive-fields of LGN and V1. Quantitatively, the performances obtained on classification tasks allows us to conclude on the relevance of this learning process on such task. However it is still far from optimality considering the supervised learning methods (Graham, 2014; Hunsberger and Eliasmith, 2015) and human-level performances. The main drawback of our method is the difficulty to train more than one or two convolutional layers with. Since spatial inhibitions are critical in our WTA scheme to achieve feature sparseness, we suspect that

the input width and height of one layer must be large enough to make the competition between neurons effective. Other competition schemes less dependent on the spatial dimension have to be explored in order to train deeper architectures with the proposed framework.

Also our binary variant of STDP rule shows the ability to train neurons with very low precision updates. Gradients used to be coded on floating-point variables ranging from 32 bits as these encoding schemes had the better trade-off between numerical precision and efficiency on CPU and GPU hardware. Gupta et al. (2015) showed the possibility to perform gradient descent with only 16-bits floating-point resolution, a feature implemented since then in NVidia Pascal and AMD RX Vega GPUs. Studies on gradient quantization (Zhou et al., 2016; Deng et al., 2017) showed promising results reducing the precision down to 2 bits without penalizing significantly the performances. The main advantage of such reduction in resolution is two-fold : the lowest the resolution, the fastest the computations (under the condition hardware has sufficient dedicated compute units) and the fastest the memory transfers. Seide et al. (2014) accelerated learning speed by a factor 50 quantizing the weight updates gradients on 1 bit, enabling a very fast transfer between the 8 GPU of the considered cluster. The binary STDP learning rule proposed here may fit this goal. Further quantization on activations and weights (even if the distributions obtained on MNIST and CIFAR-10 seem to converge to three modes) are to be studied in such framework in order to bring massive acceleration thanks to this biologically inspired method.

In order to better understand the implication of the binary STDP learning rule from a machine learning point of view, studies on the equivalence to state-of-the art methods should be performed as in Hyvärinen et al. (2004) and Carlson et al. (2013). Further mathematical analysis may help us understanding better the limits and potentials of our approach in order to combine it with other approaches. The literature in machine learning and neuroscience (accurately summarized in Marblestone et al., 2016) shows that it is unlikely that only one objective function or algorithm may be responsible for all the learning capabilities of the brain. Considered combinations include supervised approach with backpropagation compatible models such as Esser et al. (2015), reinforcement learning methods (Mnih et al., 2013; Mozafari et al., 2017), as well as other unsupervised strategies such as auto-encoders and GANs.

Finally, the binary STDP along with WTA and normalization has been shown to be successful at learning in an unsupervised manner low level visual features from image data. Extension of this learning framework on temporal data is envisaged. The roles of neural oscillations in the brain are still studied, and their place in attention-demanding tasks (Dugué et al., 2015; McLelland and VanRullen, 2016) is still under debate. Nevertheless, oscillation processes like the theta-gamma model (McLelland and VanRullen, 2016) shows interesting information segmentation abilities, and may be incorporated in a network of spiking or recurrent artificial neurons (such as GRU and LTSM) as a more hard-coded WTA scheme to evaluate their impact during learning.

## AUTHOR CONTRIBUTIONS

PF, FM, and ST: Designed the study; PF and FM: Analyzed the data; PF: Wrote the manuscript; PF, FM, and ST: Revised the manuscript, approved the final version, and agreed to be accountable for all aspects of the work.

## FUNDING

This work was supported by the Centre National de la Recherche Scientifique (CNRS), the Agence Nationale Recherche Technologie (ANRT) and Brainchip SAS, a Brainchip Holdings Ltd company.

## ACKNOWLEDGMENTS

We would like to thank Timothée Masquelier, Saeed Reza Kheradpisheh, Douglas McLelland, Christophe Garcia, and Stefan Dufner for their advice on the method and the manuscript.

## REFERENCES

Bengio, Y., Lamblin, P., Popovici, D., and Larochelle, H. (2007). "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems 19*, eds B. Schölkopf, J. C. Platt, T. Hoffman (Montreal, QC: MIT Press), 153–160.

Bengio, Y., Lee, D., Bornschein, J., and Lin, Z. (2015). Towards biologically plausible deep learning. *arXiv:1502.04156*.

Beyeler, M., Dutt, N. D., and Krichmar, J. L. (2013). Categorization and decision-making in a neurobiologically plausible spiking network using a STDP-like learning rule. *Neural Netw.* 48(Suppl. C), 109–124. doi: 10.1016/j.neunet.2013.07.012

Burbank, K. S. (2015). Mirrored STDP implements autoencoder learning in a network of spiking neurons. *PLoS Comput. Biol.* 11:e1004566. doi: 10.1371/journal.pcbi.1004566

Cao, Y., Chen, Y., and Khosla, D. (2015). Spiking deep convolutional neural networks for energy-efficient object recognition. *Int. J. Comp. Vis.* 113, 54–66. doi: 10.1007/s11263-014-0788-3

Carlson, K. D., Richert, M., Dutt, N., and Krichmar, J. L. (2013). "Biologically plausible models of homeostasis and stdp: stability and learning in spiking neural networks," in *Neural Networks (IJCNN), The 2013 International Joint Conference on IEEE* (Dallas, TX), 1–8.

Chan, T. H. , Jia, K., Gao, S., Lu, J., Zeng, Z., and Ma, Y. (2014). PCANet: A simple deep learning baseline for image classification. *arXiv:1404.3606*.

Chistiakova, M., Bannon, N. M., Bazhenov, M., and Volgushev, M. (2014). Heterosynaptic plasticity: multiple mechanisms and multiple roles. *Neuroscientist* 20, 483–498. doi: 10.1177/1073858414529829

Coates, A., Lee, H., and Ng, A. (2011). "An analysis of single-layer networks in unsupervised feature learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Vol 15, JMLR Workshop and Conference Proceedings (JMLR W&CP)* (Fort Lauderdale, FL), 215–223.

de Almeida, L., Idiart, M., and Lisman, J. E. (2009). A second function of gamma frequency oscillations: an E%-max winner-take-all mechanism selects which cells fire. *J. Neurosci.* 29, 7497–7503. doi: 10.1523/JNEUROSCI.6044-08.2009

Delorme, A., and Thorpe, S. J. (2001). Face identification using one spike per neuron: resistance to image degradations. *Neural Netw.* 14, 795–803. doi: 10.1016/S0893-6080(01)00049-1

Deng, L., Jiao, P., Pei, J., Wu, Z., and Li, G. (2017). Gated XNOR networks: deep neural networks with ternary weights and activations under a Unified Discretization Framework. *arXiv:1705.09283*.

Diehl, P. U., and Cook, M. (2015). Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Front. Comput. Neurosci.* 9:99. doi: 10.3389/fncom.2015.00099

Diehl, P. U., Neil, D., Binas, J., Cook, M., Liu, S.-C., and Pfeiffer, M. (2015). "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *2015 International Joint Conference on Neural Networks (IJCNN) IEEE*, 1–8.

Diehl, P. U., Zarrella, G., Cassidy, A., Pedroni, B. U., and Neftci, E. (2016). Conversion of artificial recurrent neural networks to spiking neural networks for low-power neuromorphic hardware. *arXiv:1601.04187*.

Dugué, L., McLelland, D., Lajous, M., and VanRullen, R. (2015). Attention searches nonuniformly in space and in time. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15214–15219. doi: 10.1073/pnas.1511331112

Esser, S. K., Appuswamy, R., Merolla, P., Arthur, J. V., and Modha, D. S. (2015). "Backpropagation for energy-efficient neuromorphic computing," in *Advances in Neural Information Processing Systems 28,* eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Montreal, QC: Curran Associates, Inc.), 1117–1125.

Esser, S. K., Merolla, P. A., Arthur, J. V., Cassidy, A. S., Appuswamy, R., Andreopoulos, A., et al. (2016). Convolutional networks for fast, energy-efficient neuromorphic computing. *arXiv:1603.08270*.

Gamrat, C., Bichler, O., and Roclin, D. (2015). "Memristive based device arrays combined with spike based coding can enable efficient implementations of embedded neuromorphic circuits," in *IEEE International Electron Devices Meeting (IEDM)* (Washington, DC), 4.5.1–4.5.7.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). Generative adversarial nets. in *Advances in Neural Information Processing Systems*, 2672–2680.

Goodfellow, I. J., Warde-farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013). "Maxout Networks," in *ICML*, (Atlanta, GA).

Graham, B. (2014). Fractional max-pooling. *arXiv:1412.6071*.

Gupta, S., Agrawal, A., Gopalakrishnan, K., and Narayanan, P. (2015). Deep learning with Limited Numerical Precision. *arXiv:1502.02551*.

Hunsberger, E., and Eliasmith, C. (2015). Spiking deep networks with LIF neurons. *arXiv:1510.08829*.

Hyvärinen, A., Karhunen, J., and Oja, E. (2004). "Independent component analysis," *Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control*, ed John Wiley & Sons (Wiley-Blackwell).

Kempter, R., Gerstner, W., and van Hemmen, J. L. (2001). Intrinsic stabilization of output rates by spike-based hebbian learning. *Neural Comput.* 13, 2709–2741. doi: 10.1162/089976601317098501

Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J., and Masquelier, T. (2016). STDP-based spiking deep neural networks for object recognition. *arXiv:1611.01421*.

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *arXiv:1312.6114*.

Krizhevsky, A. (2009). *Learning Multiple Layers of Features from Tiny Images*. Computer Science Department, University of Toronto, Technical Report.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). "Gradient-based learning applied to document recognition," *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Leibe, B., and Schiele, B. (2003). "Analyzing appearance and con tour based methods for object categorization," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Madison, WI), 409–415.

Makhzani, A., and Frey, B. J. (2015). "Winner-take-all autoencoders," in *Advances in Neural Information Processing Systems 28*, eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Montreal, QC: MIT Press), 2791–2799.

Marblestone, A. H., Wayne, G., and Kording, K. P. (2016). Towards an integration of deep learning and neuroscience. *arXiv:1606.03813*.

Markram, H., Lübke, J., Frotscher, M., and Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic aps and EPSPs. *Science* 275, 213–215.

Masquelier, T., Guyonneau, R., and Thorpe, S. J. (2009). Competitive STDP-based spike pattern learning. *Neural Comput.* 21, 1259–1276. doi: 10.1162/neco.2008.06-08-804

Masquelier, T., and Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Comput. Biol.* 3:e31. doi: 10.1371/journal.pcbi.0030031

McLelland, D., and VanRullen, R. (2016). Theta-gamma coding meets communication-through-coherence: neuronal oscillatory multiplexing theories reconciled. *PLoS Comput. Biol.* 12:e1005162. doi: 10.1371/journal.pcbi.1005162

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. A. (2013). Playing atari with deep reinforcement learning. *arXiv:1312.5602*.

Mozafari, M., Kheradpisheh, S. R., Masquelier, T., Nowzari-Dalini, A., and Ganjtabesh, M. (2017). First-spike based visual categorization using reward-modulated STDP. *arXiv:1705.09132*.

Nessler, B., Pfeiffer, M., and Maass, W. (2009). "STDP enables spiking neurons to detect hidden causes of their inputs," in *Advances in Neural Information Processing Systems 22*, eds Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Curran Associates, Inc.), 1357–1365. Available online at: http://papers.nips.cc/paper/3744-stdp-enables-spiking-neurons-to-detect-hidden-causes-of-their-inputs.pdf

Panda, P., and Roy, K. (2016). Unsupervised regenerative learning of hierarchical features in spiking deep networks for object recognition. *arXiv:1602.01510*.

Rao, R. P., and Sejnowski, T. J. (2001). Spike-timing-dependent hebbian plasticity as temporal difference learning. *Neural Comput.* 13, 2221–2237. doi: 10.1162/089976601750541787

Rasmus, A., Valpola, H., Honkala, M., Berglund, M., and Raiko, T. (2015). Semi-supervised learning with ladder network. *arXiv:1507.02672*.

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025.

Royer, S., and Paré, D. (2003). Conservation of total synaptic weight through balanced synaptic depression and potentiation. *Nature* 422, 518–522. doi: 10.1038/nature01530

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. *arXiv:1606.03498*.

Seide, F., Fu, H., Droppo, J., Li, G., and Yu, D. (2014). "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech DNNs," in *INTERSPEECH*, (Singapore).

Stromatias, E., Neil, D., Pfeiffer, M., Galluppi, F., Furber, S. B., and Liu, S.-C. (2015). Robustness of spiking Deep Belief Networks to noise and reduced bit precision of neuro-inspired hardware platforms. *Front. Neurosci.* 9:222. doi: 10.3389/fnins.2015.00222

Thorpe, S., Delorme, A., and Van Rullen, R. (2001). Spike-based strategies for rapid processing. *Neural Netw.* 14, 715–725. doi: 10.1016/S0893-6080(01)00083-1

Thorpe, S., Fize, D., and Marlot, C. (1996). Speed of processing in the human visual system. *Nature* 381:520.

Thorpe, S. J., Guyonneau, R., Guilbaud, N., Allegraud, J.-M., and VanRullen, R. (2004). Spikenet: real-time visual processing with one spike per neuron. *Neurocomputing* 58–60, 857–864. doi: 10.1016/j.neucom.2004.01.138

Torralba, A., Fergus, R., and Freeman, W. T. (2008). 80 million tiny images: a large data set for nonparametric object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 1958–1970. doi: 10.1109/TPAMI.2008.128

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., and Johnson, M. K. (2009). Neural evidence of statistical learning: efficient detection of visual regularities without awareness. *J. Cogn. Neurosci.* 21, 1934–1945. doi: 10.1162/jocn.2009.21131

Van Rullen, R., Gautrais, J., Delorme, A., and Thorpe, S. (1998). Face processing using one spike per neurone. *Biosystems* 48, 229–239.

Zhou, S., Ni, Z., Zhou, X., Wen, H., Wu, Y., and Zou, Y. (2016). Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv:1606.06160*.

**Conflict of Interest Statement:** 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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 19 July 2018
doi: 10.3389/fncom.2018.00056

# Modern Machine Learning as a Benchmark for Fitting Neural Responses

*Ari S. Benjamin*[1]*, *Hugo L. Fernandes*[2], *Tucker Tomlinson*[3], *Pavan Ramkumar*[2,4], *Chris VerSteeg*[5], *Raeed H. Chowdhury*[3,5], *Lee E. Miller*[2,3,5] and *Konrad P. Kording*[1,6]

[1] Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, United States, [2] Department of Physical Medicine and Rehabilitation, Rehabilitation Institute of Chicago, Northwestern University, Chicago, IL, United States, [3] Department of Physiology, Northwestern University, Chicago, IL, United States, [4] Department of Neurobiology, Northwestern University, Evanston, IL, United States, [5] Department of Biomedical Engineering, Northwestern University, Evanston, IL, United States, [6] Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, United States

Neuroscience has long focused on finding encoding models that effectively ask "what predicts neural spiking?" and generalized linear models (GLMs) are a typical approach. It is often unknown how much of explainable neural activity is captured, or missed, when fitting a model. Here we compared the predictive performance of simple models to three leading machine learning methods: feedforward neural networks, gradient boosted trees (using XGBoost), and stacked ensembles that combine the predictions of several methods. We predicted spike counts in macaque motor (M1) and somatosensory (S1) cortices from standard representations of reaching kinematics, and in rat hippocampal cells from open field location and orientation. Of these methods, XGBoost and the ensemble consistently produced more accurate spike rate predictions and were less sensitive to the preprocessing of features. These methods can thus be applied quickly to detect if feature sets relate to neural activity in a manner not captured by simpler methods. Encoding models built with a machine learning approach accurately predict spike rates and can offer meaningful benchmarks for simpler models.

**Keywords: encoding models, neural coding, tuning curves, machine learning, generalized linear model, GLM, spike prediction**

OPEN ACCESS

***Edited by:***
*Yoram Burak,*
*Hebrew University of Jerusalem, Israel*

***Reviewed by:***
*Tatyana Sharpee,*
*Salk Institute for Biological Studies, United States*
*Jonas Kubilius,*
*KU Leuven, Belgium and Massachusetts Institute of Technology, United States*

***Correspondence:**
*Ari S. Benjamin*
*aarrii@seas.upenn.edu*

***Received:*** *05 October 2017*
***Accepted:*** *29 June 2018*
***Published:*** *19 July 2018*

***Citation:***
*Benjamin AS, Fernandes HL, Tomlinson T, Ramkumar P, VerSteeg C, Chowdhury RH, Miller LE and Kording KP (2018) Modern Machine Learning as a Benchmark for Fitting Neural Responses. Front. Comput. Neurosci. 12:56. doi: 10.3389/fncom.2018.00056*

## INTRODUCTION

A central tool of neuroscience is the tuning curve, which maps aspects of external stimuli to neural responses. The tuning curve can be used to determine what information a neuron encodes in its spikes. For a tuning curve to be meaningful it is important that it accurately describes the neural response. Often, however, methods are chosen for simplicity but not evaluated for their relative accuracy. Since inaccurate methods may systematically miss aspects of the neural response, any choice of predictive method should be compared with accurate benchmark methods.

A popular predictive model for neural data is the Generalized Linear Model (GLM) (Nelder and Baker, 1972; Simoncelli et al., 2004; Truccolo et al., 2005; Wu et al., 2006; Gerwinn et al., 2010). The GLM performs a nonlinear operation upon a linear combination of the input features, which are often called external covariates. Typical covariates are stimulus features, movement vectors, or the animal's location, and may include covariate history or spike history. In the absence of history terms, the GLM is also referred to as a linear-nonlinear Poisson (LN) cascade. The nonlinear

operation is usually held fixed, though it can be learned (Chichilnisky, 2001; Paninski et al., 2004a), and the linear weights of the combined inputs are chosen to maximize the agreement between the model fit and the neural recordings. This optimization problem of weight selection is convex, allowing a global optimum, and can be solved with efficient algorithms (Paninski, 2004). The assumption of Poisson firing statistics can often be loosened (Pillow et al., 2005), as well, allowing the modeling of a broad range of neural responses. Due to its ease of use, perceived interpretability, and flexibility, the GLM has become a popular model of neural spiking.

When using a GLM, it is important to check that the method's assumptions about the data are correct. The GLM's central assumption is that the inputs relate linearly to the log firing rate, or generally some monotonic function of the firing rate. It thus cannot learn arbitrary multi-dimensional functions of the inputs. When the nonlinearity is different than assumed, it is likely that the optimal weight on one input will depend on the values of other inputs. In this case the GLM will only partially represent the neural response, will poorly predict activity, and may not be reproducible on other datasets. This drawback has been noted before, and indeed the GLM has been shown to miss nonlinearity in numerous circumstances (Butts et al., 2011; Freeman et al., 2015; Heitman et al., 2016; McIntosh et al., 2016). However, GLMs are still commonly applied without comparison to other methods. To test if the linearity assumption is valid, it is sufficient to test if other nonlinear methods predict activity more accurately from the same features. Many extensions have been proposed that introduce a specific form of nonlinearity (McFarland et al., 2013; Theis et al., 2013; Latimer et al., 2014; Williamson et al., 2015; Maheswaranathan et al., 2017), but these methods ask specific research questions and are not intended as general benchmarks. What is needed is are nonlinear methods that are universally applicable to new data.

Machine learning (ML) methods for regression have improved dramatically since the invention of the GLM. Many ML methods require little feature engineering (i.e., pre-transformations the features) and do not need to assume linearity. These methods are thus ideal candidates for benchmark methods. The ML approach is now quite standardized and robust across many domains of data. As exemplified by winning solutions on Kaggle, an ML competition website (Kaggle Winner's Blog, 2016), the usual approach is to fit several top performing methods, and then to ensemble these models together. These methods are now relatively easy to implement in a few lines of code in a scripting language such as Python, and are enabled by well-supported machine learning packages, such as scikit-learn (Pedregosa et al., 2011), Keras (Chollet, 2015), Tensorflow (Abadi et al., 2016), and XGBoost (Chen and Guestrin, 2016). The greatly increased predictive power of modern ML methods is now very accessible and could help to benchmark and improve the state of the art in encoding models across neuroscience.

In order to investigate the feasibility of ML as a benchmark approach, we applied several ML methods, including artificial neural networks, gradient boosted trees, and ensembles to the task of predicting spike rates, and evaluated their performance alongside a GLM. We compared the methods on data from three separate brain areas. These areas differed greatly in the effect size of covariates and in their typical spike rates, and thus served to evaluate the strengths of these methods across different conditions. In each area we found that the ensemble of methods could more accurately predict spiking than the GLM with typical feature choices. The use of an ML benchmark thus made clear that tuning curves built for these features with a GLM would not capture the full nature of neural activity. We provide our implementing code at https://github.com/KordingLab/spykesML so that all neuroscientists may easily test and compare ML to their own methods on other datasets.

# MATERIALS AND METHODS

## Data

We tested our methods at predicting spike rates for neurons in the macaque primary motor cortex, the macaque primary somatosensory cortex, and the rat hippocampus. All animal use procedures were approved by the institutional animal care and use committees at Northwestern University and conform to the principles outlined in the Guide for the Care and Use of Laboratory Animals (National Institutes of Health publication no. 86-23, revised 1985). Data presented here were previously recorded for use with multiple analyses. Procedures were designed to minimize animal suffering and reduce the number used.

The macaque motor cortex data consisted of previously published electrophysiological recordings from 82 neurons in the primary motor cortex (M1) (Stevenson et al., 2011). The neurons were sorted from recordings made during a two-dimensional center-out reaching task with eight targets. In this task the monkey grasped the handle of a planar manipulandum that controlled a cursor on a computer screen and simultaneously measured the hand location and velocity (**Figure 1**). After training, an electrode array was implanted in the arm area of area 4 on the precentral gyrus. Spikes were discriminated using offline sorter (Plexon, Inc), counted and collected in 50-ms bins. The neural recordings used here were taken in a single session lasting around 13 min.

The macaque primary somatosensory cortex (S1) data was recorded during a two-dimensional random-pursuit reaching task and was previously unpublished. In this task, the monkey gripped the handle of the same manipulandum. The monkey was rewarded for bringing the cursor to a series of randomly positioned targets appearing on the screen. After training, an electrode array was implanted in the arm area of area 2 on the post-central gyrus, which receives a mix of cutaneous and proprioceptive afferents. Spikes were processed as for M1. The data used for this publication derives from a single recording session lasting 51 min.

As with M1 (described in results), we processed the hand position, velocity, and acceleration accompanying the S1 recordings in an attempt to obtain linearized features. The features $(x, y, \dot{x}, \dot{y})$ were found to be the most successful for the GLM. Since cells in the arm area of S1 have been shown to have approximately sinusoidal tuning curves relating to movement direction (Prud'homme and Kalaska, 1994), we also tested the

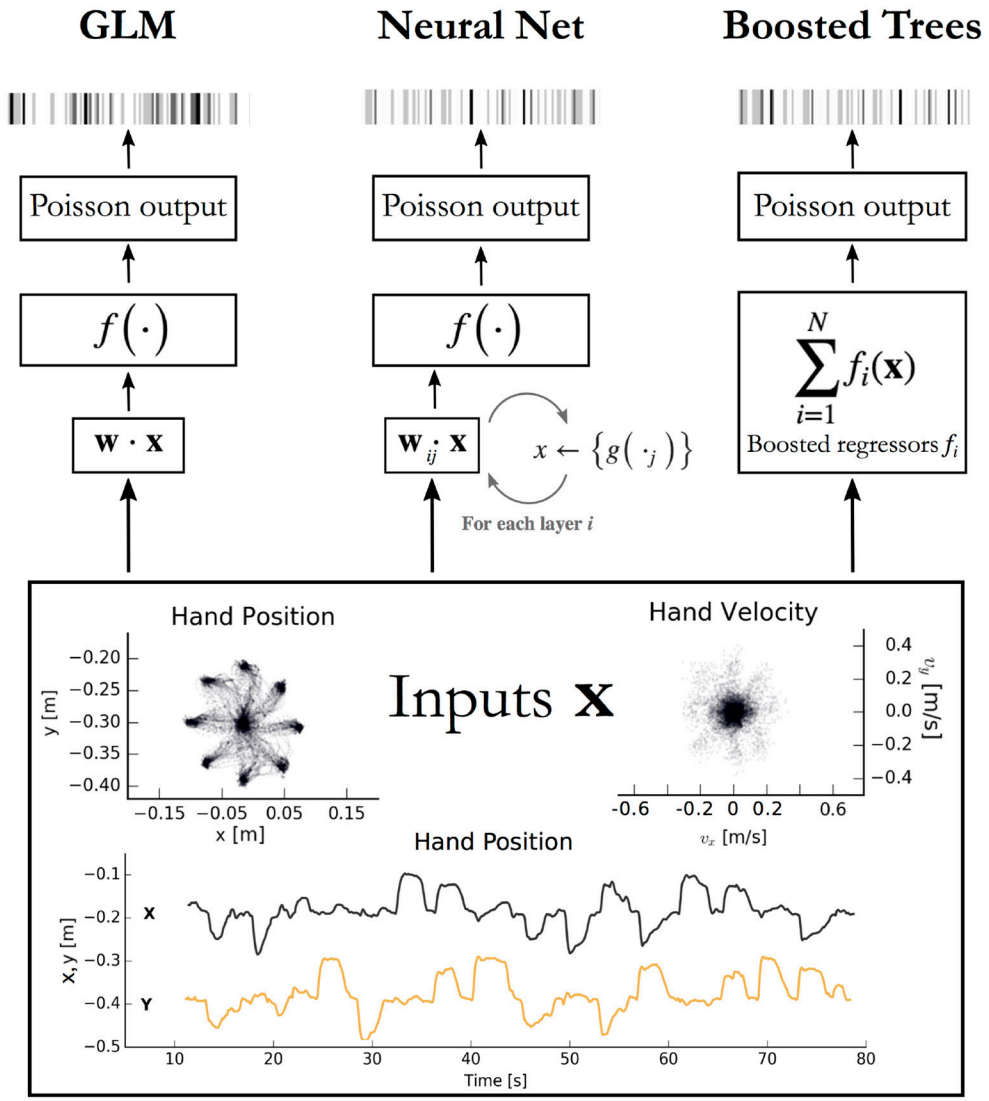


**FIGURE 1 |** Encoding models aim to predict spikes, top, from input data, bottom. The inputs displayed are the position and velocity signals from the M1 dataset (Stevenson et al., 2011) but could represent any set of external covariates. The GLM takes a linear combination of the inputs, applies an exponential function $f$, and produces a Poisson spike probability that can be used to generate spikes **(Left)**. The feedforward neural network **(Center)** does the same when the number of hidden layers $i = 0$. With $i \geq 1$ hidden layers, the process repeats; each of the $j$ nodes in layer $i$ computes a nonlinear function $g$ of a linear combination of the previous layer. The vector of outputs from all $j$ nodes is then fed as input to the nodes in the next layer, or to the final exponential $f$ on the final iteration. Boosted trees **(Right)** return the sum of N functions of the original inputs. Each of the $f_i$ is built to minimize the residual error of the sum of the previous $f_{0:i-1}$.

same feature transformations as were performed for M1 but did not observe any increase in predictive power.

The third dataset consists of recordings from 58 neurons in the CA1 region of the rat dorsal hippocampus during a single 93 min free foraging experiment, previously published and made available online by the authors (Mizuseki et al., 2009a,b). Position data from two head-mounted LEDs provided position and heading direction inputs. Here we binned inputs and spikes from this experiment into 50 ms bins. Since many neurons in the dorsal hippocampus are responsive to the location of the rat, we processed the 2D position data into a list of squared distances from a $5 \times 5$ grid of place fields that tile the workspace. Each position feature thus has the form

$$p_{ij} = \frac{1}{2}\left(x(t) - \mu_{ij}\right)^T \Sigma_{ij}^{-1}\left(x(t) - \mu_{ij}\right),$$

where $\mu_{ij}$ is the center of place field $i, j \leq 5$ and $\Sigma_{ij}$ is a covariance matrix chosen for the uniformity of tiling. An exponentiated linear combination of the $p_{ij}$ (as is performed in the GLM) evaluates to a single Gaussian centered anywhere between the place fields. The inclusion of the $p_{ij}$ as features thus transforms the standard representation of cell-specific place fields (Brown et al., 1998) into the mathematical formulation of a GLM. The

final set of features included the $p_{ij}$ as well as the rat speed and head orientation.

## Treatment of Spike and Covariate History

We slightly modified our data preparation methods for spike rate prediction when spike and covariate history terms were included as regressors (**Figure 6**). To construct spike and covariate history filters, we convolved 10 raised cosine bases (built as in Pillow et al., 2008) with binned spikes and covariates. The longest temporal basis included times up to 250 ms before the time bin being predicted. This process resulted in 120 total covariates per sample (10 current covariates, 100 covariate temporal filters, and 10 spike history filters). We predicted spike rates in 5 ms bins (rather than 50 ms) to allow for modeling of more precise time-dependent phenomena, such as refractory effects. The cross-validation scheme also differs from the main analysis of this paper, as using randomly selected splits of the data would result in the appearance in the test set of samples that were in history terms of training sets, potentially resulting in overfitting. We thus employed a cross-validation routine to split the data continuously in time, assuring that no test set sample has appeared in any form in training sets.

## Generalized Linear Model

The Poisson GLM is a multivariate regression model that describes the instantaneous firing rate as a nonlinear function of a linear combination of input features (see e.g., Schwartz et al., 2006; Aljadeff et al., 2016 for review, Pillow et al., 2008; Fernandes et al., 2014; Ramkumar et al., 2016 for usage). Here, we took the form of the nonlinearity to be exponential, as is common in previous applications of GLMs to similar data (Saleh et al., 2012). It should be noted that it is also possible to learn arbitrary link functions through histogram methods (Chichilnisky, 2001; Paninski et al., 2004a). We approximate neural activity as a Poisson process, in which the probability of firing in any instant is independent of firing history. The general form of the GLM is depicted in **Figure 1**. We implemented the GLM using elastic-net regularization, using the open-source Python package pyglmnet (Ramkumar et al., 2017). The regularization path was optimized separately on a single neuron in each dataset on a validation set not used for scoring.

## Neural Network

Neural networks are well-known for their success at supervised learning tasks. More comprehensive reviews can be found elsewhere (Schmidhuber, 2015). Here, we implemented a simple feedforward neural network and, for the analysis with history terms, an LSTM, a recurrent neural network architecture that allows the modeling of time dependencies on multiple time-scales (Gers et al., 2000).

We point out that a feedforward neural network with no hidden layers is equivalent in mathematical form to a GLM (**Figure 1**). For multilayer networks, one can write each hidden layer of $n$ nodes as simply $n$ GLMs, each taking the output of the previous layer as inputs (noting that the weights of each are chosen to maximize only the final objective function, and that the intermediate nonlinearities need not be the same as the output nonlinearity). A feedforward neural network can be seen as a generalization, or repeated application of a GLM.

The networks were implemented with the open-source neural network library Keras, running Theano as the backend (Chollet, 2015; Team et al., 2016). The feedforward network contained two hidden layers, dense connections, rectified linear activation, and a final exponentiation. To help avoid overfitting, we allowed dropout on the first layer, included batch normalization, and allowed elastic-net regularization upon the weights (but not the bias term) of the network (Srivastava et al., 2014). The networks were trained to maximize the Poisson likelihood of the neural response. We optimized over the number of nodes in the first and second hidden layers, the dropout rate, and the regularization parameters for the feedforward neural network, and for the number of epochs, units, dropout rate, and batch size for the LSTM. Optimization was performed on only a subset of the data from a single neuron in each dataset, using Bayesian optimization (Snoek et al., 2012) in an open-source Python implementation (BayesianOptimization, 2016).

## Gradient Boosted Trees

A popular method in many machine learning competitions is that of gradient boosted trees. Here we describe the general operation of XGBoost, an open-source implementation that is efficient and highly scalable, works on sparse data, and easy to implement out-of-the-box (Chen and Guestrin, 2016).

XGBoost trains many sequential models to minimize the residual error of the sum of previous model. Each model is a decision tree, or more specifically a classification and regression tree (CART) (Friedman, 2001). Training a decision tree amounts to determining a series of rule-based splits on the input to classify output. The CART algorithm generalizes this to regression by taking continuously-valued weights on each of the leaves of the decision tree.

For any predictive model $\hat{y}^{(1)} = f_1(\mathbf{x_i})$ and true response $y_i$, we can define a loss function $l\left(\hat{y}^{(1)},\ y_i\right)$ between the prediction and the response. The objective to be minimized during training is then simply the sum of the loss over each training example $i$, plus some regularizing function $\Omega$ that biases toward simple models.

$$L = \sum_i l(\hat{y}_i^{(1)}, y_i) + \Omega(f_1)$$

After minimizing $L$ for a single tree, XGBoost constructs a second tree $f_2(\mathbf{x_i})$ that approximates the residual. The objective to be minimized is thus the total loss $L$ between the true response $y_i$ and the sum of the predictions given by the first tree and the one to be trained.

$$L = \sum_i l(\hat{y}_i^{(1)} + f_2(\mathbf{x}_i), y_i) + \Omega(f_2)$$

This process is continued sequentially for a predetermined number of trees, each trained to approximate the residual of the sum of previous trees. In this manner XGBoost is designed to progressively decrease the total loss with each additional tree. At

the end of training, new predictions are given by the sum of the outputs of all trees.

$$\hat{y} = \sum_{k=1}^{N} f_k(\mathbf{x})$$

In practice, it is simpler to choose the functions $f_k$ via gradient boosting, which minimizes a second order approximation of the loss function (Friedman et al., 2000).

XGBoost offers several additional parameters to optimize performance and prevent overfitting. Many of these describe the training criteria for each tree. We optimized some of these parameters for a single neuron in each dataset using Bayesian optimization (again over a validation set different from the final test set). These parameters included the number of trees to train, the maximum depth of each decision tree, and the minimum weight allowed on each decision leaf, the data subsampling ratio, and the minimum gain required to create a new decision branch.

## Random Forests

We implement random forests here to increase the power of the ensemble (see below); their performance alone is displayed in Supplementary Figure 1. It should be noted that the Scikit-learn implementation currently only minimizes the mean-squared error of the output, which is not properly applicable to Poisson processes and may cause poor performance. Despite this drawback their presence still improves the ensemble scores. Random forests train multiple parallel decision trees on the features-to-spikes regression problem (not sequentially on the remaining residual, as in XGBoost) and averages their outputs (Ho, 1998). The variance on each decision tree is increased by training on a sample of the data drawn with replacement (i.e., bootstrapped inputs) and by choosing new splits using only a random subset of the available features. Random forests are implemented in Scikit-learn (Pedregosa et al., 2011).

## Ensemble Method

It is a common machine learning practice to create ensembles of several trained models. Different algorithms may learn different characteristics of the data, make different types of errors, or generalize differently to new examples. Ensemble methods allow for the successes of different algorithms to be combined. Here we implemented *stacking*, in which the output of several models is taken as the input set of a new model (Wolpert, 1992). After training the GLM, neural network, random forest, and XGBoost on the features of each dataset, we trained an additional instance of XGBoost using the spike rate predictions of the previous methods as input. The outputs of this "second stage" XGBoost are the predictions of the ensemble.

## Scoring and Cross-Validation

Each of the three methods was scored with the Poisson pseudo-$R^2$ score, a scoring function applicable to Poisson processes (Cameron and Windmeijer, 1997). Note that a standard $R^2$ score assumes Gaussian noise and cannot be applied here. The pseudo-$R^2$ was calculated as one minus the ratio of the deviances of the predicted output $\hat{y}$ to the mean firing rate $\overline{y}$.

$$R^2_M = 1 - \frac{D\left(\hat{y}\right)}{D\left(\overline{y}\right)}$$

We can gain intuition into the pseudo-$R^2$ score by writing out the deviances in terms of log likelihoods $L()$, and combining the fraction.

$$R^2_M = 1 - \frac{\log L\left(y\right) - \log L\left(\hat{y}\right)}{\log L\left(y\right) - \log L\left(\overline{y}\right)} = \frac{\log L\left(\hat{y}\right) - \log L\left(\overline{y}\right)}{\log L\left(y\right) - \log L\left(\overline{y}\right)}$$

This expression includes $L\left(y\right)$, which is the log likelihood of the "saturated model," which offers one parameter per observation and models the data perfectly. The pseudo-$R^2$ can thus be interpreted as the fraction of the maximum potential log-likelihood gain achieved by the tested model (Cameron and Windmeijer, 1997). It takes a value of 0 when the data is as likely under the tested model as the null model, and a value of 1 when the tested model perfectly describes the data. It is empirically a lower value than a standard $R^2$ when both are applicable (Domencich and McFadden, 1975). The null model can also be taken to be a model other than the mean firing rate (e.g., the GLM) to directly compare two methods, in which case we refer to the score as the "comparative pseudo-$R^2$." The comparative pseudo-$R^2$ is referred to elsewhere as the "relative pseudo-$R^2$," renamed here to avoid confusion with the difference of two standard pseudo-$R^2$ scores both measured against the mean (Fernandes et al., 2014).

We used 8-fold cross-validation (CV) when assigning a final score to the models. The input and spike data were segmented into eight equal partitions. These partitions were continuous in time when spike and covariate history were included as covariates, and otherwise were segmented randomly in time. The methods were trained on seven partitions and tested on the eighth, and this was repeated until all segments served as the test partition once. The mean of the eight scores are then recorded for the final score.

Cross-validation for ensemble methods requires extra care since the inputs for the ensemble are themselves model predictions for each data point. The training set for the ensemble must contain predictions from methods that were themselves not trained on the validation set. Otherwise, there may be a leak of information from the validation set into the training set and the validation score might be better than on a true held-out set. This rules out using simple $k$-fold CV with all methods and the ensemble trained on the same test/train splits. Instead, we used a nested CV scheme to train and score the ensemble. We create an outer $j = 8$ folds to build training and test sets for the ensemble. On each outer fold we create first-order predictions for each data point in the following manner. We first run an inner $k$-fold CV on just the training set (i.e., 7/8 of the original dataset) with each first stage method such that we obtain predictions for the whole

training set of that fold. This ensures that the ensemble's test set was never used for training any method. Finally, we build the ensemble's test set from the predictions of the first stage methods trained on the entire training set. The ensemble can then be tested on a held-out set that was never used to fit any model. The process is repeated for each of the *j* folds and the mean and variance of the *j* scores of the ensemble's predictions are recorded.

## RESULTS

We applied several machine learning methods to predict spike counts in three brain regions and compared the quality of the predictions to those of a GLM. Our primary analysis centered on neural recordings from the macaque primary motor cortex (M1) during reaching (**Figure 1**). We examined the methods' relative performance on several sets of movement features with various levels of preprocessing, including one set that included spike and covariate history terms. Analyses of data from rhesus macaque S1 and rat hippocampus indicate how these methods compare for areas other than M1. On each of the three datasets we trained a GLM and compared it to the performance of a feedforward neural network, XGBoost (a gradient boosted trees implementation), and an ensemble method. The ensemble was an additional instance of XGBoost trained on the predictions of all three methods plus a random forest regressor. The application of these methods allowed us to demonstrate the potential of a modern approach to be able to identify whether there are typically neural nonlinearities that are not captured by a GLM. The code implementing these methods can be used by any electrophysiology lab to benchmark their own encoding models.

To test that all methods work reasonably well in a trivial case, we trained each to predict spiking from a simple, well-understood feature. Some neurons in M1 have been described as responding linearly to the exponentiated cosine of movement direction relative to a preferred angle (Amirikian and Georgopulos, 2000). We therefore predicted the spiking of M1 neurons from the cosine and sine of the direction of hand movement in the reaching task. (The linear combination of a sine and cosine curve is a phase-shifted cosine, by identity, allowing the GLM to learn the proper preferred direction). We observed that each method identified a similar tuning curve (**Figure 2B**) and that the bulk of the neurons in the dataset were just as well predicted by each of the methods (**Figures 2A,C**) {though the ensemble was slightly more accurate than the GLM, with mean comparative pseudo-$R^2$ above zero, 0.06 [0.043 – 0.084], 95% bootstrapped confidence interval (CI)}. The similar performance suggested that, for the majority of neurons, an exponentiated cosine successfully approximates the response to movement direction alone, as has been previously found (Paninski et al., 2004b). All methods can in principle estimate tuning curves, and machine learning can indicate if the proper features are used.

If the form of the nonlinearity is not known, machine learning can still attain good predictive ability. To illustrate the ability of modern machine learning to find the proper nonlinearity, we performed the same analysis as above but omitted the initial cosine feature-engineering step. Trained on only the hand velocity direction, in radians, which changes discontinuously at $\pm\pi$, all methods but the GLM closely matched the predictive power they attained using the engineered feature (**Figure 3A**). The GLM failed at generating a meaningful tuning curve, which was expected since the exponentiated velocity direction is not equal to cosine tuning (**Figure 3B**). Both trends were consistent across the population of recorded neurons (**Figure 3C**). The neural net, XGBoost, and ensemble methods can learn the nonlinearity of single features without requiring manual feature transformation.

The inclusion of multiple features raises the possibility of nonlinear feature interactions that may elude a GLM. As a simple demonstration of this principle, we trained all methods on the four-dimensional set of hand position and velocity $(x, y, \dot{x}, \dot{y})$. While all methods gained predictive power relative to models using movement direction alone, the GLM failed to match the other methods (**Figures 4A,C**). If the GLM was fit alone, and no further featuring engineering been attempted, these features would have appeared to be relatively uninformative of the neural response. If nonlinear interactions exist between preselected features, machine learning methods can potentially learn these interactions and indicate if more linearly-related features exist.

While feature engineering can improve the performance of GLMs, it is not always simple to guess the optimal set of processed features. We demonstrated this by training all methods on features that have previously been successful at explaining spike rate in a similar center-out reaching task (Paninski et al., 2004a). These extra features included the sine and cosine of velocity direction (as in **Figure 2**), and the speed, radial distance of hand position, and the sine and cosine of position direction. The training set was thus 10-dimensional, though highly redundant, and was aimed at maximizing the predictive power of the GLM. Feature engineering improved the predictive power of all methods to variable degrees, with the GLM improving to the level of the neural network (**Figure 5**). XGBoost and the ensemble still predicted spike rates better than the GLM (**Figure 5C**), with the ensemble scoring on average nearly double the GLM (ratio of population means of 1.8 [1.4 – 2.2], 95% bootstrapped CI). The ensemble was significantly better than XGBoost (mean comparative pseudo-$R^2$ of 0.08 [0.055 – 0.103], 95% bootstrapped CI) and was thus consistently the best predictor. Though standard feature engineering greatly improved the GLM, the ensemble and XGBoost still could identify that neural nonlinearity was missed by the GLM.

It is important to note that the specific ordering of methods depends on features such as the amount of data available for training. We investigated this dependence for the M1 dataset by plotting the cross-validated performance as a function of the fraction of the data used for training (Supplementary Figure 3). Some neurons are best fit by the GLM when very little data is available, while other neurons are best fit by XGBoost and the ensemble for any amount of data tested. The neural network is most sensitive to training data availability. This sensitivity to the domain of data emphasizes the importance of the applied ML paradigm of evaluating (and potentially ensembling) many methods.

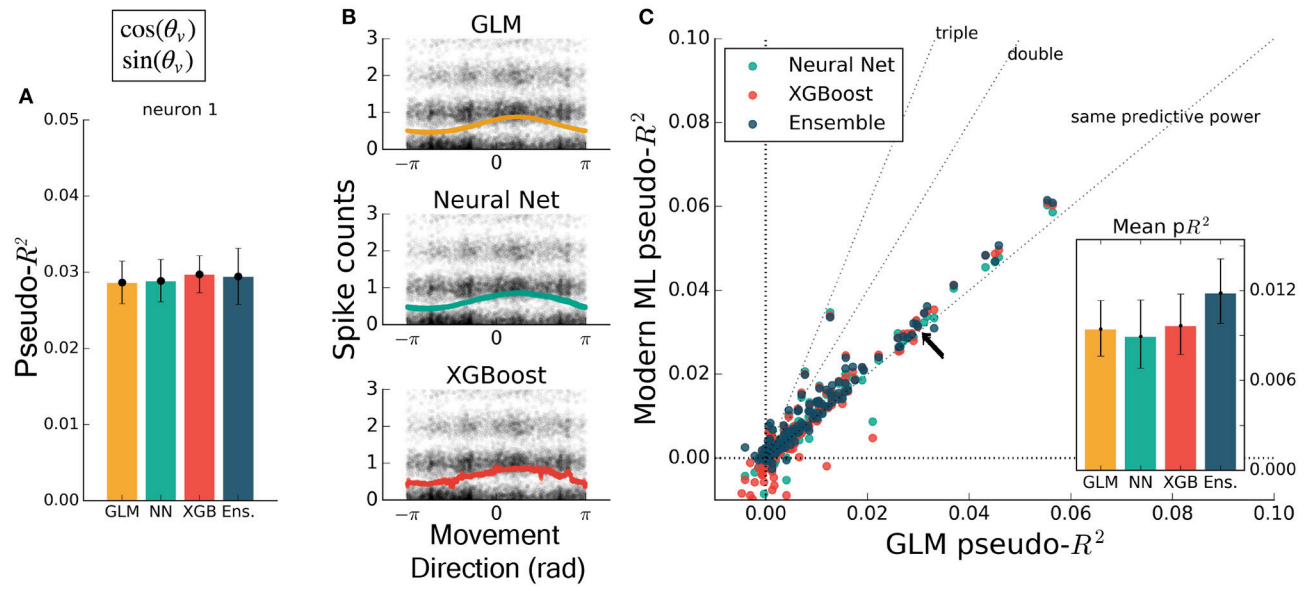


**FIGURE 2 |** Encoding models of M1 performed similarly when trained on the sine and cosine of hand velocity direction. All methods can in principle estimate tuning curves. **(A)** The pseudo-$R^2$ for an example neuron was similar for all four methods. On this figure and in **Figures 3–5** the example neuron is the same, and is not the neuron for which method hyperparameters were optimized. **(B)** We constructed tuning curves by plotting the predictions of spike rate on the validation set against movement direction. The black points are the recorded responses, to which we added y-axis jitter for visualization to better show trends in the naturally quantized levels of binned spikes. The tuning curves of the neural net and XGBoost were similar to that of the GLM. The tuning curve of the ensemble method was similar and is not shown. **(C)** Plotting the pseudo-$R^2$ of modern ML methods vs. that of the GLM indicates that the similarity of methods generalizes across neurons. The single neuron plotted at left is marked with black arrows. The mean scores, inset, indicate the overall success of the methods; error bars represent the 95% bootstrap confidence interval.

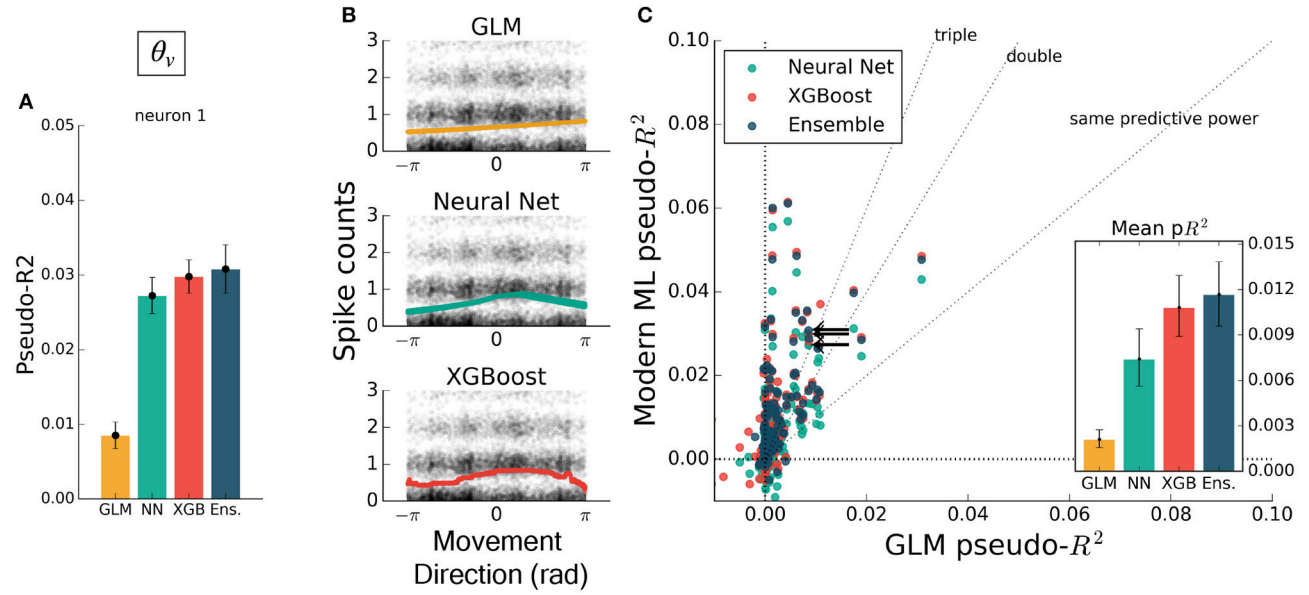


**FIGURE 3 |** Modern ML models learn the cosine nonlinearity when trained on hand velocity direction, in radians. **(A)** For the same example neuron as in **Figure 2**, the neural net and XGBoost maintained the same predictive power, while the GLM was unable to extract a relationship between direction and spike rate. **(B)** XGBoost and neural nets displayed reasonable tuning curves, while the GLM reduced to the average spiking rate (with a small slope, in this case). **(C)** Most neurons in the population were poorly fit by the GLM, while the ML methods achieved the performance levels of **Figure 2**. The ensemble performed the best of the methods tested. The single neuron plotted at left is marked with black arrows.

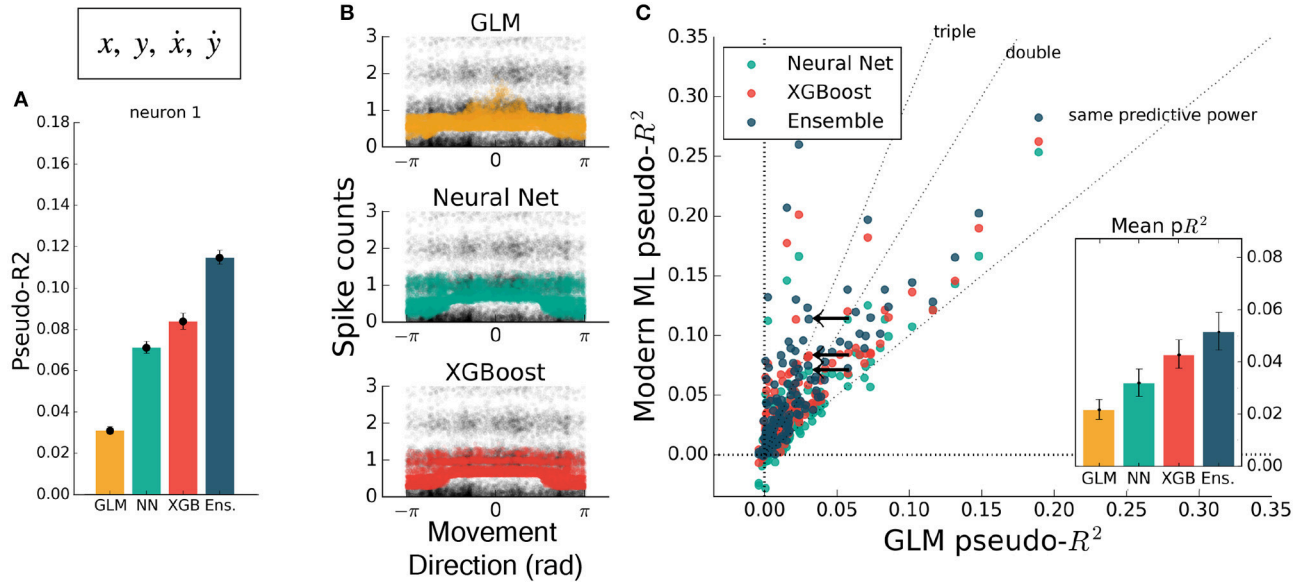


**FIGURE 4 |** Modern ML methods can learn nonlinear interactions between features. Here the methods are trained on the feature set $(x, y, \dot{x}, \dot{y})$. Note the change in axes scales from **Figures 2**, **3**. **(A)** For the same example neuron as in **Figure 3**, all methods gained a significant amount of predictive power, indicating a strong encoding of position and speed or their correlates. The GLM showed less predictive power than the other methods on this feature set. **(B)** The spike rate in black, with jitter on the y-axis, again overlaid with the predictions of the three methods plotted against velocity direction. The projection of the multidimensional tuning curve onto a 1D velocity direction dependence leaves the projected curve diffuse. **(C)** The ensemble method, neural network, and XGBoost performed consistently better than the GLM across the population. The mean pseudo-$R^2$ scores show the hierarchy of success across methods. The single neuron plotted at left is marked with black arrows.

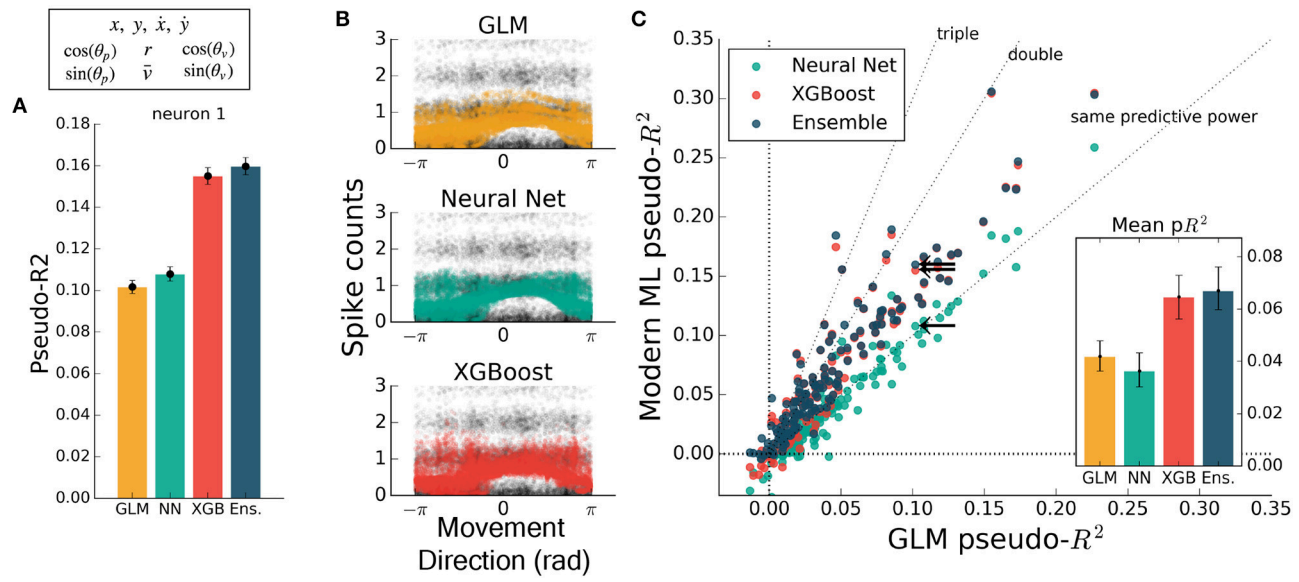


**FIGURE 5 |** Modern ML methods outperform the GLM with standard featuring engineering. For this figure, all methods were trained on the features $(x, y, \dot{x}, \dot{y})$ plus the engineered features. **(A)** For this example neuron, inclusion of the computed features increased the predictive power of the GLM to the level of the neural net. All methods increased in predictive power. **(B)** The tuning curves for the example neuron are diffuse when projected onto the movement direction, indicating a high-dimensional dependence. **(C)** Even with feature engineering, XGBoost and the ensemble consistently achieve pseudo-$R^2$ scores higher than the GLM, though the neural net does not. The neuron selected at left is marked with black arrows.

Studies employing a GLM often include activity history as a covariate when predicting spike rates, as well as past values of the covariates themselves, and it is known that this allows GLMs to model a wider range of phenomena (Weber and Pillow,

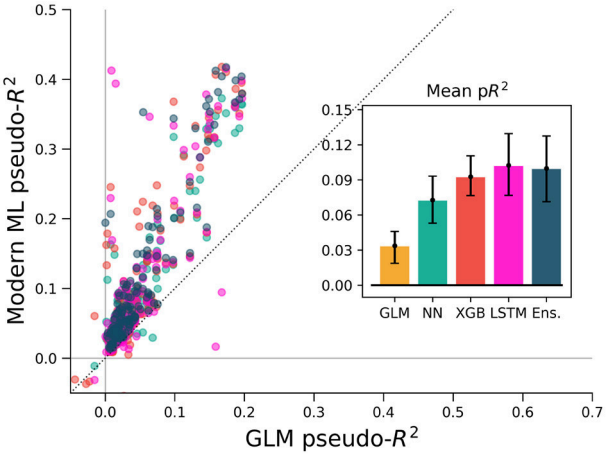


**FIGURE 6 |** ML algorithms outperform a GLM when covariate history and neuron spike history are included. The feature set of **Figure 5** (in macaque M1) was augmented with spike and covariate history terms, so that spike rate was predicted for each 5 ms time bin from the past 250 ms of covariates and neural activity. Cross-validation methods for this figure differ from other figures (see methods) and pseudo-$R^2$ scores should not be compared directly across figures. All methods outperform the GLM, indicating that the inclusion of history terms does not alone allow the GLM to capture the full nonlinear relationship between covariates and spike rate.

2016). We tested various ML methods on the M1 dataset using this history-augmented feature set to see if all methods would still explain a similar level of activity. We binned data by 5 ms (rather than 50 ms) to agree in timescale with similar studies, and built temporal filters by convolving 10 raised-cosine bases with features and spikes. We note that smaller time bins result in a sparser dataset, and thus pseudo-$R^2$ scores cannot be directly compared with other analysis in this paper. On this problem, our selected ML algorithms again outperformed the GLM (**Figure 6**). The overall best algorithm was the LSTM, which we include here as it specifically designed for modeling time series, though for most neurons XGBoost performed similarly. Thus, for M1 neurons, the GLM did not capture all predicable phenomena even when spike and covariate history were included.

To ensure that these results were not specific to the motor cortex, we extended the same analyses to primary somatosensory cortex (S1) data. We again predicted neural activity from hand movement and speed, and here without spike or covariate history terms. The ML methods outperformed the GLM for all but three of the 52 neurons, indicating that firing rates in S1 generally relate nonlinearly to hand position and velocity (**Figure 7A**). Each of the three ML methods performed similarly for each neuron. The S1 neural function was thus equally learnable by each method, which is surprising given the dissimilarity of the neural network and XGBoost algorithms. This situation would occur if learning has saturated near ground truth, though this cannot be proven definitively to be the case. It is at least clear from the underperformance of the GLM that the relationship of S1 activity to these covariates is nonlinear beyond the assumptions of the GLM.

We asked if the same trends of performance would hold for the rat hippocampus dataset, which was characterized by very low mean firing rates but strong effect sizes. All methods were trained on a list of squared distances to a grid of place fields and on and the rat head orientation, as described in methods. Far more even than the neocortical data, neurons were described much better by XGBoost and the ensemble method than by the GLM (**Figure 7B**). Many neurons shifted from being completely unpredictable by the GLM (pseudo-$R^2$ near zero) to very predictable by XGBoost and the ensemble (pseudo-$R^2$ above 0.2). These neurons thus have responses that do not correlate with firing in any one Gaussian place field. We note that the neural network performed poorly, likely due to the very low firing rates of most hippocampal cells (Supplementary Figure 2). The median spike rate of the 58 neurons in the dataset was just 0.2 spikes/s, and it was only on the four neurons with rates above 1 spikes/s that the neural network achieved pseudo-$R^2$ scores comparable to the GLM. The relative success of XGBoost was interesting given the failure of the neural network, and supported the general observation that boosted trees can work well with smaller and sparser datasets than those that neural networks generally require (Supplementary Figure 3). Thus for hippocampal cells, a method leveraging decision trees such as XGBoost or the ensemble is able to capture more structure in the neural response and thus demonstrate a deficiency of the parameterization of the GLM.

## DISCUSSION

We analyzed the ability of various machine learning techniques at the task of predicting binned spike counts in three brain regions. We found that of the tested ML methods, XGBoost and the ensemble routinely predicted spike counts more accurately than did the GLM, which is a popular method for neural data. Feedforward neural networks did not always outperform the GLM and were often worse than XGBoost and the ensemble. Machine learning methods, especially LSTMs, also outperformed GLMs when covariate and spike history were included as inputs. The ML methods performed comparably well with and without feature engineering, even for the very low spike rates of the hippocampus dataset. These findings indicate that a standard ML approach can serve as a reliable benchmark to test if data meets the assumptions of a GLM. Furthermore, it may be quite common that standard ML outperforms GLMs given standard feature choices.

When a GLM fails to explain data as well as more expressive, nonlinear methods, the current parameterization of inputs must relate to the data with a different nonlinearity than is assumed by the GLM. Such situations have been identified several times in the literature (Butts et al., 2011; Freeman et al., 2015; Heitman et al., 2016; McIntosh et al., 2016). This unaccounted nonlinearity may produce feature weights that do not reflect true feature importance. A GLM will incorrectly predict no dependence on feature $x$ whatsoever, for example, in the extreme case when the neural response to some feature $x$ does not correlate with $\exp(x)$. The only way to ensure that feature weights can be reliably

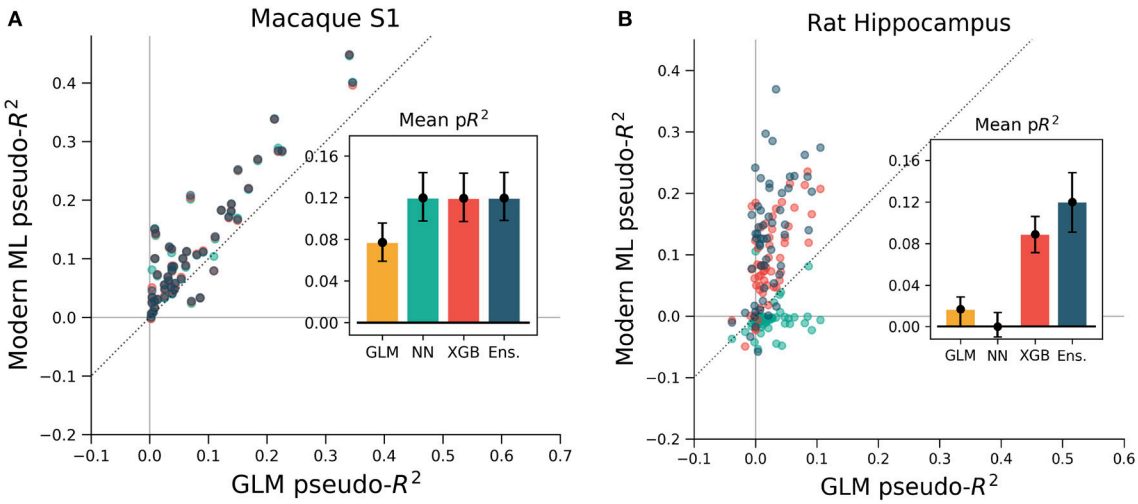


**FIGURE 7 |** XGBoost and the ensemble method predicted the activity of neurons in S1 and in hippocampus better than a GLM. The diagonal dotted line in both plots is the line of equal predictive power with the GLM. **(A)** All methods outperform the GLM in the macaque S1 dataset. Interestingly, the neural network, XGBoost and the ensemble scored very similarly for each neuron in the 52 neuron dataset. **(B)** Many neurons in the rat hippocampus were described well by XGBoost and the ensemble but poorly by the GLM and the neural network. The poor neural network performance in the hippocampus was due to the low rate of firing of most neurons in the dataset (Supplementary Figure 2). Note the difference in axes; hippocampal cells are generally more predictable than those in S1.

interpreted is to find an input parameterization that maximizes the GLM's predictive power. ML methods can assist this process by indicating how much nonlinearity remains to be explained. New features can then be tested, such as those suggested by a search for maximally informative dimensions (Sharpee et al., 2004). In our analysis, then, the GLM underperforms because we have selected the suboptimal input features. It is always theoretically possible to linearize features such that a GLM obtains equal predictive power. ML methods can highlight the deficiency of features that might have otherwise seemed uncontroversial. When applying a GLM or any simple model to neural data, it is important to compare its predictive power with standard ML methods to ensure the neural response is properly understood.

There are other ways of estimating the performance of a method besides benchmark nonlinear methods. For example, if the same exact stimulus can be given many times in a row, then we can estimate neural variability without having to model how activity depends on stimulus features (Schoppe et al., 2016). This approach, however, requires that we can model how neural responses vary with repetition (Grill-Spector et al., 2006). This approach also makes it difficult to include spike history as an input, since the exact history is rarely repeated. We note that in some cases it may also be impossible to show the same stimulus multiple times, e.g., because eyes move. However, comparing these two classes of benchmark would be interesting on applications where both are feasible.

Advanced ML methods are not widely considered to be interpretable. Interpretation is not necessary for performance benchmarks, but it would be desirable to use these methods as standalone encoding models. We can better discuss this issue with a more precise definition of interpretability. Following Lipton, we make the distinction between a method's *post-hoc interpretability*, the ease of justifying its predictions, and *transparency,* the degree to which its operation and internal parameters are human-readable or easily understandable (Lipton et al., 2016). A GLM is certainly more transparent than many ML methods due to its algorithmic simplicity. Certain nonlinear extensions of the GLM have also been designed to remain transparent (McFarland et al., 2013; Theis et al., 2013; Latimer et al., 2014; Williamson et al., 2015; Maheswaranathan et al., 2017). For high-level areas, though, such as V4, the linearized features may be difficult to be interpreted themselves (Yamins et al., 2014), though it may be possible to increase the interpretability of features (Kaardal et al., 2013). A GLM is also generally more conducive to *post-hoc* interpretations, though this is also possible with modern ML methods. It is possible, for example, to visualize the aspects of stimuli that most elicit a predicted response, as has been implemented in previous applications of neural networks to spike prediction (Lau et al., 2002; Prenger et al., 2004). Various other methods exist in the literature to enable *post-hoc* explanations (McAuley and Leskovec, 2013; Simonyan et al., 2013). Here we highlight Local Interpretable Model-Agnostic Explanations (LIME), an approach that fits simple models in the vicinity of single examples to allow a local interpretation (Ribeiro et al., 2016). On problems where interpretability is important, such capabilities for *post-hoc* justifications may prove sufficient.

Not all types of interpretability are necessary for a given task, and many scientific questions can be answered based on predictive ability alone. Questions of the form, "does feature $x$ contribute to neural activity?" for example, or "is past activity necessary to explain current activity?" require no method transparency. One can simply ask whether predictive power increases with feature $x$'s inclusion or decreases upon its exclusion. Importance measures based on inclusion and

exclusion, or upon the strategy of shuffling a covariate of interest, are well-studied in statistics and machine learning (Bell and Wang, 2000; Strobl et al., 2008). Depending on the application, it may thus be worthwhile to ask not just whether different features could improve a GLM but also whether it is enough to use ML methods directly. It is possible for many questions to stay agnostic to the form of linearized features and directly use changes in predictive ability.

With ongoing progress in machine learning, many standard techniques are easy to implement and can even be automated. Ensemble methods, for example, remove the need to choose any one algorithm. Moreover, the choice of model-specific parameters is made easy by hyperparameter search methods and optimizers. We hope that this ease of use might encourage use in the neurosciences, thereby increasing the power and efficiency of studies involving neural prediction without requiring complicated, application-specific methods development (e.g., Corbett et al., 2012). Community-supported projects in automated machine learning, such as autoSklearn and auto-Weka, are quickly improving and promise to handle the entire regression workflow (Feurer et al., 2015; Kotthoff et al., 2016). Applied to neuroscience, these tools will allow researchers to gain descriptive power over current methods even with simple, out-of-the-box implementations.

Machine learning methods perform quite well and make minimal assumptions about the form of neural encoding. Models that seek to understand the form of the neural code can test if they systematically misconstrue the relationship between stimulus and response by comparing their performance to these benchmarks. Encoding models built with machine learning can thus greatly aid the construction of models that capture arbitrary nonlinearity and more accurately describe neural activity.

The code used for this publication is available at https://github.com/KordingLab/spykesML. We invite researchers to adapt it freely for future problems of neural prediction.

## AUTHOR CONTRIBUTIONS

KK and HF first conceived the project. TT, CV, and RC gathered and curated macaque data. AB prepared the manuscript and performed the analyses, for which HF and PR assisted. LM and KK supervised, and all authors assisted in editing.

## FUNDING

LM acknowledges the following grants from the National Institute of Neurological Disorders and Stroke (https://www.ninds.nih.gov/): NS074044, NS048845, NS053603, NS095251. CV recognizes support from the Biomedical Data Driven Discovery (BD3) Training Program funded through the National Institute of Health, grant number 5T32LM012203-02. KK acknowledges support from National Institute of Health (https://www.nih.gov/) grants R01NS063399, R01NS074044, and MH103910. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. RC thanks NSF GRFP DGE-1324585.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2018.00056/full#supplementary-material

## REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2016). Tensorflow: large-scale machine learning on heterogeneous distributed systems. *arXiv:*160304467[preprint].

Aljadeff, J., Lansdell, B. J., Fairhall, A. L., and Kleinfeld, D. (2016). Analysis of neuronal spike trains, deconstructed. *Neuron* 91, 221–259. doi: 10.1016/j.neuron.2016.05.039

Amirikian, B., and Georgopulos, A. P. (2000). Directional tuning profiles of motor cortical cells. *Neurosci. Res.* 36, 73–79. doi: 10.1016/S0168-0102(99)00112-1

BayesianOptimization (2016). *GitHub Repository*.

Bell, D. A., and Wang, H. (2000). A formalism for relevance and its application in feature subset selection. *Mach. Learn.* 41, 175–195. doi: 10.1023/A:1007612503587

Brown, E. N., Frank, L. M., Tang, D., Quirk, M. C., and Wilson, M. A. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *J. Neurosci.* 18, 7411–7425. doi: 10.1523/JNEUROSCI.18-18-07411.1998

Butts, D. A., Weng, C., Jin, J., Alonso, J. M., and Paninski, L. (2011). Temporal precision in the visual pathway through the interplay of excitation and stimulus-driven suppression. *J. Neurosci.* 31, 11313–11327. doi: 10.1523/JNEUROSCI.0434-11.2011

Cameron, A. C., and Windmeijer, F. A. (1997). An R-squared measure of goodness of fit for some common nonlinear regression models. *J. Econom.* 77, 329–342.

Chen, T., and Guestrin, C. (2016). Xgboost: a scalable tree boosting system. *arXiv preprint arXiv:*160302754[preprint].

Chichilnisky, E. (2001). A simple white noise analysis of neuronal light responses. *Netw. Comp. Neural Syst.* 12, 199–213. doi: 10.1080/713663221

Chollet, F. (2015). Keras. *GitHub repository*. Available Online at: https://github.com/keras-team/keras

Corbett, E. A., Perreault, E. J., and Körding, K. P. (2012). Decoding with limited neural data: a mixture of time-warped trajectory models for directional reaches. *J. Neural Eng.* 9:036002. doi: 10.1088/1741-2560/9/3/036002

Domencich, T. A., and McFadden, D. (1975). *Urban Travel Demand-A Behavioral Analysis*, Oxford: North-Holland Publishing Company Limited.

Fernandes, H. L., Stevenson, I. H., Phillips, A. N., Segraves, M. A., and Kording, K. P. (2014). Saliency and saccade encoding in the frontal eye field during natural scene search. *Cereb. Cortex* 24, 3232–3245. doi: 10.1093/cercor/bht179

Feurer, M. Klein, A., Eggensperger, K., Springenberg, J., Blum, M., and Hutter, F. (2015). "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems, Vol. 28*, eds C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Curran Associates, Inc.), 2962–2970. Available online at: http://papers.nips.cc/paper/5872-efficient-and-robust-automated-machine-learning.pdf

Freeman, J., Field, G. D., Li, P. H., Greschner, M., Gunning, D. E., Mathieson, K., et al. (2015). Mapping nonlinear receptive field structure in primate retina at single cone resolution. *Elife* 4:e05241. doi: 10.7554/eLife.05241

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Statist.* 29, 1189–1232. doi: 10.1214/aos/1013203451

Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Ann. Statist.* 28, 337–407. doi: 10.1214/aos/1016120463

Gers, F. A., Schmidhuber, J., and Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Comput.* 12, 2451–2471. doi: 10.1162/089976600300015015

Gerwinn, S., Macke, J. H., and Bethge, M. (2010). Bayesian inference for generalized linear models for spiking neurons. *Front. Comp. Neurosci.* 4:12. doi: 10.3389/fncom.2010.00012

Grill-Spector, K., Henson, R., and Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* 10, 14–23. doi: 10.1016/j.tics.2005.11.006

Heitman, A., Brackbill, N., Greschner, M., Sher, A., Litke, A. M., and Chichilnisky, E. (2016). Testing pseudo-linear models of responses to natural scenes in primate retina. *bioRxiv*:045336 [preprint]. doi: 10.1101/045336

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 832–844.

Kaardal, J., Fitzgerald, J. D., Berry, M. J., and Sharpee, T. O. (2013). Identifying functional bases for multidimensional neural computations. *Neural Comput.* 25, 1870–1890. doi: 10.1162/NECO_a_00465

Kaggle Winner's Blog (2016). Available Online at: http://blog.kaggle.com/

Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F., and Leyton-Brown, K. (2016). Auto-WEKA 2.0: automatic model selection and hyperparameter optimization in WEKA. *J. Mach. Learn. Res.* 17, 1–5. doi: 10.1145/2487575.2487629

Latimer, K. W., Chichilnisky, E., Rieke, F., and Pillow, J. W. (eds). (2014). "Inferring synaptic conductances from spike trains with a biophysically inspired point process model," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc.), 954–962.

Lau, B., Stanley, G. B., and Dan, Y. (2002). Computational subunits of visual cortical neurons revealed by artificial neural networks. *Proc. Natl. Acad. Sci. U.S.A.* 99, 8974–8979. doi: 10.1073/pnas.122173799

Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint* arXiv:1606.03490.

Maheswaranathan, N., Baccus, S. A., and Ganguli, S. (2017). Inferring hidden structure in multilayered neural circuits. *bioRxiv*: 120956 [preprint]. doi: 10.1101/120956

McAuley, J., and Leskovec, J. (eds). (2013). "Hidden factors and hidden topics: understanding rating dimensions with review text," in *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong: ACM).

McFarland, J. M., Cui, Y., and Butts, D. A. (2013). Inferring nonlinear neuronal computation based on physiologically plausible inputs. *PLoS Comput. Biol.* 9:e1003143. doi: 10.1371/journal.pcbi.1003143

McIntosh, L., Maheswaranathan, N., Nayebi, A., Ganguli, S., Baccus, S. (2016). "Deep learning models of the retinal response to natural scenes," in *Advances in Neural Information Processing Systems, Vol. 29*, eds D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc.), 1369–1377. Available online at: http://papers.nips.cc/paper/6388-deep-learning-models-of-the-retinal-response-to-natural-scenes.pdf

Mizuseki, K., Sirota, A., Pastalkova, E., and Buzsáki, G. (2009a). Multi-unit recordings from the rat hippocampus made during open field foraging. doi: 10.6080/K0Z60KZ9

Mizuseki, K., Sirota, A., Pastalkova, E., and Buzsáki, G. (2009b). Theta oscillations provide temporal windows for local circuit computation in the entorhinal-hippocampal loop. *Neuron* 64, 267–280. doi: 10.1016/j.neuron.2009.08.037

Nelder, J. A., and Baker, R. J. (1972). Generalized linear models. *Encyclop. Statist. Sci.* 135, 370–384. doi: 10.2307/2344614

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Netw. Comp. Neural Sys.* 15, 243–262. doi: 10.1088/0954-898X_15_4_002

Paninski, L., Fellows, M. R., Hatsopoulos, N. G., and Donoghue, J. P. (2004a). Spatiotemporal tuning of motor cortical neurons for hand position and velocity. *J. Neurophysiol.* 91, 515–532. doi: 10.1152/jn.00587.2002

Paninski, L., Shoham, S., Fellows, M. R., Hatsopoulos, N. G., and Donoghue, J. P. (2004b). Superlinear population encoding of dynamic hand trajectory in primary motor cortex. *J. Neurosci.* 24, 8551–8561. doi: 10.1523/JNEUROSCI.0919-04.2004

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.

Pillow, J. W., Paninski, L., Uzzell, V. J., Simoncelli, E. P., and Chichilnisky, E. (2005). Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. *J. Neurosci.* 25, 11003–11013. doi: 10.1523/JNEUROSCI.3305-05.2005

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., et al. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* 454, 995–999. doi: 10.1038/nature07140

Prenger, R., Wu, M. C., David, S. V., and Gallant, J. L. (2004). Nonlinear V1 responses to natural scenes revealed by neural network analysis. *Neural Netw.* 17, 663–679. doi: 10.1016/j.neunet.2004.03.008

Prud'homme, M. J., and Kalaska, J. F. (1994). Proprioceptive activity in primate primary somatosensory cortex during active arm reaching movements. *J. Neurophysiol.* 72, 2280–2301. doi: 10.1152/jn.1994.72.5.2280

Ramkumar, P., Jas, M., Achakulvisut, T., Idrizović, A., themantalope, Acuna, D. E., et al. (2017). *Pyglmnet 1.0.1*. (Chicago, IL).

Ramkumar, P., Lawlor, P. N., Glaser, J. I., Wood, D. K., Phillips, A. N., Segraves, M. A., et al. (2016). Feature-based attention and spatial selection in frontal eye fields during natural scene search. *J. Neurophysiol.* 116, 1328–1343. doi: 10.1152/jn.01044.2015

Ribeiro, M. T., Singh, S., and Guestrin, C. (eds). (2016). "Why should i trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, CA: ACM).

Saleh, M., Takahashi, K., and Hatsopoulos, N. G. (2012). Encoding of coordinated reach and grasp trajectories in primary motor cortex. *J. Neurosci.* 32, 1220–1232. doi: 10.1523/JNEUROSCI.2438-11.2012

Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Netw.* 61:85–117. doi: 10.1016/j.neunet.2014.09.003

Schoppe, O., Harper, N. S., Willmore, B. D., King, A. J., and Schnupp, J. W. (2016). Measuring the performance of neural models. *Front. Comput. Neurosci.* 10:10. doi: 10.3389/fncom.2016.00010

Schwartz, O., Pillow, J. W., Rust, N. C., and Simoncelli, E. P. (2006). Spike-triggered neural characterization. *J. Vis.* 6:13. doi: 10.1167/6.4.13

Sharpee, T., Rust, N. C., and Bialek, W. (2004). Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Comput.* 16, 223–250. doi: 10.1162/089976604322742010

Simoncelli, E. P., Paninski, L., Pillow, J., and Schwartz, O. (2004). "Characterization of neural responses with stochastic stimuli," in *The cognitive neurosciences, 3rd edn* ed M. Gazzaniga (MIT Press), 327–338.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv*:13126034[preprint].

Snoek, J., Larochelle, H., and Adams, R. P. (2012). "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems, Vol. 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc.), 2951–2959. Available online at: http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Stevenson, I. H., Cherian, A., London, B. M., Sachs, N. A., Lindberg, E., Reimer, J., et al. (2011). Statistical assessment of the stability of neural movement representations. *J. Neurophysiol.* 106, 764–774. doi: 10.1152/jn.00626.2010

Strobl, C., Boulesteix, A. L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics* 9:307. doi: 10.1186/1471-2105-9-307

Team, T. T. D., Al-Rfou, R., Alain, G., Almahairi, A., Angermueller, C., Bahdanau, D., et al. (2016). Theano: a Python framework for fast computation of mathematical expressions. *arXiv:* 160502688[preprint].

Theis, L., Chagas, A. M., Arnstein, D., Schwarz, C., and Bethge, M. (2013). Beyond GLMs: a generative mixture modeling approach to neural system identification. *PLoS Comput. Biol.* 9:e1003356. doi: 10.1371/journal.pcbi.1003356

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., and Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *J. Neurophysiol.* 93, 1074–1089. doi: 10.1152/jn.00697.2004

Weber, A. I., and Pillow, J. W. (2016). Capturing the dynamical repertoire of single neurons with generalized linear models. *arXiv:* 160207389[preprint].

Williamson, R. S., Sahani, M., and Pillow, J. W. (2015). The equivalence of information-theoretic and likelihood-based methods for neural dimensionality reduction. *PLoS Comput. Biol.* 11:e1004141. doi: 10.1371/journal.pcbi.1004141

Wolpert, D. H. (1992). Stacked generalization. *Neural Netw.* 5, 241–259.

Wu, M. C. K., David, S. V., and Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.* 29, 477–505. doi: 10.1146/annurev.neuro.29.051605.113024

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceed. Natl. Acad. Sci. U.S.A.* 111, 8619–8624. doi: 10.1073/pnas.1403112111

**Conflict of Interest Statement:** 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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 24 January 2019
doi: 10.3389/fncom.2019.00002

# Modeling Emotions Associated With Novelty at Variable Uncertainty Levels: A Bayesian Approach

*Hideyoshi Yanagisawa*[1]*, *Oto Kawamata*[1] and *Kazutaka Ueda*[2]

[1] Design Engineering Laboratory, Department of Mechanical Engineering, The University of Tokyo, Tokyo, Japan, [2] Creative Design Laboratory, Department of Mechanical Engineering, The University of Tokyo, Tokyo, Japan

Acceptance of novelty depends on the receiver's emotional state. This paper proposes a novel mathematical model for predicting emotions elicited by the novelty of an event under different conditions. It models two emotion dimensions, arousal and valence, and considers different uncertainty levels. A state transition from before experiencing an event to afterwards is assumed, and a Bayesian model estimates a posterior distribution as being proportional to the product of a prior distribution and a likelihood function. Our model uses Kullback-Leibler divergence of the posterior from the prior, which we termed information gain, to represent arousal levels because it corresponds to surprise, a high-arousal emotion, upon experiencing a novel event. Based on Berlyne's hedonic function, we formalized valence as a summation of reward and aversion systems that are modeled as sigmoid functions of information gain. We derived information gain as a function of prediction errors (i.e., differences between the mean of the posterior and the peak likelihood), uncertainty (i.e., variance of the prior that is proportional to prior entropy), and noise (i.e., variance of the likelihood function). This functional model predicted an interaction effect of prediction errors and uncertainty on information gain, which we termed the arousal crossover effect. This effect means that the greater the uncertainty, the greater the information gain for a small prediction error. However, for large prediction errors, greater uncertainty means a smaller information gain. To verify this effect, we conducted an experiment with participants who watched short videos in which different percussion instruments were played. We varied uncertainty levels by using familiar and unfamiliar instruments, and we varied prediction error magnitudes by including congruent or incongruent percussive sounds in the videos. Event-related potential P300 amplitudes and subjective reports of surprise in response to the percussive sounds were used as measures of arousal levels, and the findings supported the hypothesized arousal crossover effect. The concordance between our model's predictions and our experimental results suggests that Bayesian information gain can be decomposed into uncertainty and prediction errors and is a valid measure of emotional arousal. Our model's predictions of arousal may help identify positively accepted novelty.

**Keywords: novelty, emotion, information, arousal, valence, uncertainty, P300, surprise**

OPEN ACCESS

***Edited by:***
*Florentin Wörgötter, University of Göttingen, Germany*

***Reviewed by:***
*Jan Lauwereyns, Kyushu University, Japan*
*J. Michael Herrmann, University of Edinburgh, United Kingdom*

***Correspondence:**
*Hideyoshi Yanagisawa*
*hide@mech.t.u-tokyo.ac.jp*

***Received:*** *31 August 2018*
***Accepted:*** *09 January 2019*
***Published:*** *24 January 2019*

***Citation:***
*Yanagisawa H, Kawamata O and Ueda K (2019) Modeling Emotions Associated With Novelty at Variable Uncertainty Levels: A Bayesian Approach. Front. Comput. Neurosci. 13:2. doi: 10.3389/fncom.2019.00002*

# INTRODUCTION

Novelty is a factor of creativity. Acceptance of novelty, however, depends on the receiver's emotions. As the "most advanced yet acceptable" (MAYA) principle of industrial designer Raymond Loewy (1951) suggested, an extremely advanced (i.e., novel) design may not be accepted. In design aesthetics, Hekkert et al. (2003) observed experimentally that both typicality and novelty affect product design preferences in ways consistent with the MAYA principle. Berlyne (1970) suggested that novelty, which he termed a collative variable, is a source of arousal potential. According to his theory, an appropriate level of arousal potential might induce a positive hedonic response, but an extreme arousal potential might induce negative responses. Several experimental studies have supported Berlyne's theory, including studies on food preferences (Giacalone et al., 2014) and artistic preferences (Silvia, 2005). However, Berlyne's model did not mathematically formalize novelty or its effects on emotions, and biases due to factors such as one's prior knowledge and experience were not exhaustively investigated. Experiments with multiple participants are required to identify the effect of novelty on the emotional response to each target and condition. The objective of this study was to mathematically model emotions elicited by novelty in order to predict how novelty affects emotions. In doing so, we aimed to provide fundamental knowledge of how to achieve acceptable novelty. Most dimensional models of emotion incorporate dimensions for arousal (or intensity) and valence (i.e., positivity or negativity) (Russell, 1980; Lang, 1995). We therefore proposed a mathematical model incorporating arousal and valence dimensions through an information theory approach. We used this model to analyze how the uncertainty of expectations prior to a novel event and the difference between expectations and reality (i.e., prediction errors) interactively affect emotional arousal. We tested our model's predictions by conducting an experiment in which participants watched short videos of percussion instruments. In the experiment, we induced uncertainty of expectations by showing instruments of varying probable familiarity, and we used inconsistencies between the instrument shown and the sound played to model prediction errors. We evaluated participants' responses to the videos by analyzing event-related potentials (ERPs) and subjective reports of feelings of surprise.

# MODEL OF EMOTIONAL DIMENSIONS ELICITED BY A NOVEL EVENT

## Overview

Novelty provides new information. We assume that the amount of information gained from an event represents the degree of novelty. The information content of an event $x$ can be described as $I(x) = -\log p_x$, where $p_x$ is the probability of $x$. $I(x)$ is termed self-information. The self-information averaged over a probability density is termed information entropy, which Shannon et al. (1949) defined as follows:

$$H(X) = -\sum\nolimits_{x \in X} p_x \log p_x \tag{1}$$

For the continuous random variable $X$ following a probability density distribution $p(x)$, information entropy is expressed as:

$$H(X) = -\int_{-\infty}^{\infty} p(x) \log p(x) dx \tag{2}$$

Assume a state transition from before an event to afterwards. Let the probability density distribution of a continuous random variable $x$ before an event occurs, which we term the *prior*, be $q(x)$, and let the probability density distribution of $x$ after an event occurs, which we term the *posterior*, be $p(x)$. The information entropy of the prior represents the expectation of information content gained after an event occurs or the uncertainty of prior expectations. Information content gained after an event occurs corresponds to the decrement of information entropy over the posterior. Thus, the information content gained is obtained by subtracting prior self-information from posterior self-information and averaging over the posterior:

$$\left\langle -\log q(x) - (-\log p(x)) \right\rangle_p = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \equiv D_{KL}\left(p(x)||q(x)\right) \tag{3}$$

where $<q>_p$ represents the average of density $q$ over density $p$. The expression $D_{KL}\left(p||q\right)$ is the Kullback-Leibler (KL) divergence of $p$ from $q$ (Kullback and Leibler, 1951). Hereinafter, we term the KL divergence of the Bayesian posterior from the prior the *information gain*. The more novel an event is, the more information one gains. Information gain represents averaged surprise. Itti and Baldi (2009) defined the KL divergence of the Bayesian posterior from the prior as surprise and provided experimental evidence that it attracts visual attention.

Surprise is often used as a typical high-arousal emotion (Mauss and Robinson, 2009). Thus, we used the information gain as a mathematical expression of the arousal dimension of emotion. We then investigated the valence dimensions. An event with no information causes no arousal and has a neutral valence. Conversely, excessive information gain, such that one can hardly cope, should cause discomfort (i.e., a negative valence). Therefore, we hypothesized that one positively accepts a novel event providing an appropriate amount of information gain that can be coped with. Based on the arousal potential model (Berlyne, 1970), we formulated the valence as a function of information gain.

## Bayesian Model

Bayes's theorem provides a formula for updating the prior to the posterior. Recent studies have indicated that humans perform near-optimal Bayesian inference (Ma et al., 2006) in a wide variety of tasks, ranging from cue integration (Ernst and Banks, 2002; Kersten et al., 2004; Stocker and Simoncelli, 2006; Yanagisawa, 2016) to decision-making and motor control (Körding and Wolpert, 2004, 2006). Let a prior be $\pi(\theta)$ in terms of a parameter $\theta$ that one estimates. After one obtains continuous data $x \in R$ by experiencing an event, the prior $\pi(\theta)$ is updated to the posterior $\pi(\theta|x)$ according to the following formula derived from Bayes's theorem:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int_\theta f(x|\theta)\pi(\theta)d\theta} \propto f(x|\theta)\pi(\theta) \tag{4}$$

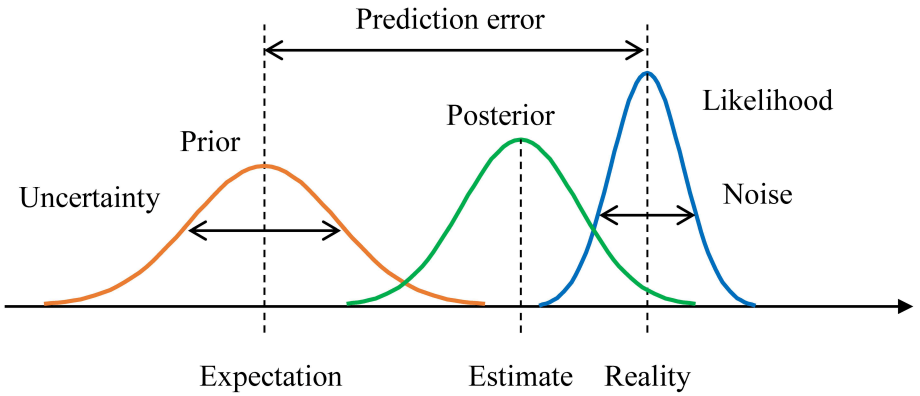


**FIGURE 1 |** Example of Bayesian inference with a prior distribution, a posterior distribution, and a likelihood function. The prediction error is the difference between the prior expectation and the peak of the likelihood function (i.e., reality). Uncertainty is the variance of the prior. Noise is the variance of the likelihood function.

where $f(x|\theta)$ is the likelihood function of $\theta$ when data $x$ are obtained. The posterior is proportional to the product of the likelihood function and the prior.

**Figure 1** shows an example of the relationships between the prior, the posterior, and the likelihood function. Neural population activity with Poisson variability can encode any Gaussian probability distribution (Ma et al., 2006). With Poisson variability, the posterior with a flat prior converges to a Gaussian distribution as the number of neurons increases. The mean of the Gaussian distribution is close to the stimulus at which the population activity peaks. The variance of the distribution is encoded as a value that is inversely proportional to the gain of the population code (i.e., the distribution's amplitude). Hence, we assume Gaussian distributions for the prior and the likelihood function. Assume one obtains $n$ samples of event $x$ and encodes them as a Gaussian posterior $N(\mu, \sigma^2)$ with a flat prior. Now assume a non-flat prior of $\mu$ that follows a Gaussian distribution $N(\eta, \tau^2)$. Using Bayes's theorem, the prior is updated to a Gaussian distribution $N(\eta_{post}, {\tau_{post}}^2)$, where:

$$Average: \eta_{post} = \frac{s_p \bar{x} + s_l \eta}{s_p + s_l}; \; Variance: {\sigma_{post}}^2 = \frac{s_p s_l}{s_p + s_l} \tag{5}$$

In these formulae, $\bar{x}$ is the mean of the data, $s_p = \tau^2$, and $s_l = \sigma^2/n$. Therefore, the prior and the posterior are represented as the following Gaussian functions, respectively:

$$\pi(\mu) = N(\mu; \eta, s_p) = \frac{1}{\sqrt{2\pi s_p}} \exp\left[-\frac{(\mu-\eta)^2}{2s_p}\right], and \tag{6}$$

$$\pi(\mu|x) = N(\mu; \eta_{post}, {\sigma_{post}}^2) = \frac{1}{\sqrt{2\pi {\sigma_{post}}^2}} \exp\left[-\frac{(\mu-\eta_{post})^2}{2{\sigma_{post}}^2}\right] \tag{7}$$

## A Functional Model of Emotional Arousal

As noted in Overview, we represented emotional arousal as information gain after experiencing an event. The information gain from the prior to the posterior $D_{KL}(\pi(\mu|x)||\pi(\mu)) \equiv G$ can be derived from formulae (2, 5, 6, and 7) as the following formula:

$$\begin{aligned} G &= \int_{-\infty}^{\infty} \pi(\mu|x) \log \frac{\pi(\mu|x)}{\pi(\mu)} d\mu \\ &= \frac{1}{2}\left\{ \frac{s_p}{(s_p+s_l)^2}\delta^2 + \log\frac{s_p+s_l}{s_l} - \frac{s_p}{s_p+s_l} \right\} \end{aligned} \tag{8}$$

where $\delta$ is the difference between the prior expectation ($\eta$) and the peak of the likelihood function ($\bar{x}$). $\delta$ represents the difference between expectations and reality, so we term $\delta$ the *prediction error* (Yanagisawa, 2016) (**Figure 1**).

Information entropy of the prior is proportional to a logarithm of $\tau$ as follows:

$$H_{prior} = -\int_{-\infty}^{\infty} \pi(\mu) \log \pi(\mu) d\mu = \log\sqrt{2\pi e} \log\tau \propto \log\tau \tag{9}$$

Thus, we term $s_p$ the *uncertainty* (Yanagisawa, 2016), and $s_l$ represents the variation of data $x$. In the case of sensory data (i.e., stimuli), the variance refers to *external noise* (Yanagisawa, 2016). From formula (7), we can regard the information gain $G$ as a function of the prediction error $\delta$ , the uncertainty $s_p$, and the external noise $s_l$:

$$G = f(\delta, s_p, s_l) \tag{10}$$

## Interaction Effect of Uncertainty and Prediction Errors on Information Gain

We analyzed how prediction errors, uncertainty, and external noise affect information gain (i.e., arousal levels). In formula (8), information gain is a quadratic function of the prediction error $\delta$ when uncertainty and external noise are fixed.

$$\begin{aligned} G &= \alpha\delta^2 + \beta, \\ \alpha &= \frac{s_p}{2(s_p+s_l)^2}, and \\ \beta &= \frac{1}{2}\left(log\frac{s_p+s_l}{s_l} - \frac{s_p}{s_p+s_l}\right) \end{aligned} \tag{11}$$

The value of $\alpha$ is always greater than zero because $s_p$ and $s_l$ are variances that are always greater than zero. Thus, the information gain is a monotonically increasing function of a prediction error. This means that the level of an arousal dimension, such as the degree of surprise, is proportional to the square of the difference between expectations and reality.

Next, we investigated the effect of uncertainty. We found that the partial derivative of the intercept $\beta$ with respect to uncertainty $s_p$ is always less than zero:

$$\frac{\partial \beta}{\partial s_p} = \frac{s_p}{2(s_p + s_l)^2} > 0 \tag{12}$$

Thus, at $\delta = 0$, the greater the uncertainty, the greater the information gain. We then investigated the case of $\delta > 0$. We compared any two information gain functions of $\delta$ using formula (10) with constant external noise between different degrees of uncertainty. If the two functions of different uncertainties have an intersection, then the information gains change as $\delta$ increases. We then assumed two information gain functions with different uncertainties, $G_1$ and $G_2$:

$$\begin{aligned} G_1 &= \alpha_1 \delta^2 + \beta_1 \; and \\ G_2 &= \alpha_2 \delta^2 + \beta_2 \end{aligned} \tag{13}$$

A condition where the two functions have an intersection is $\alpha_1\delta^2 + \beta_1 = \alpha_2\delta^2 + \beta_2$. We derived $\delta^2(\alpha_1 - \alpha_2) + (\beta_1 - \beta_2) = 0$ under $\beta_1 \neq \beta_2$. Therefore, $(\alpha_1 - \alpha_2)(\beta_1 - \beta_2) < 0$ is the condition. We found that this condition applies when the relationship between different uncertainties $s_{p1}$ and $s_{p2}$ and constant external noise $s_l$ is as follows:

$$s_{p_1} s_{p_2} > {s_l}^2 \tag{14}$$

Because the uncertainty of prediction is likely to exceed the external noise (i.e., the uncertainty of sensory stimuli), the condition in question is likely to occur. Given formula (12), the greater the uncertainty, the greater the intercept of the information gain function. As the prediction error increases, the difference in information gains between the two functions changes such that lower uncertainty tends to mean greater information gain.

**Figure 2** shows two functions of information gain with respect to different uncertainties at constant external noise. The two information gain functions have an intersection point. The information gain as an index of arousal (in this case, surprise) increases as the prediction error increases. The prediction error and uncertainty have an interaction effect on information gain. The greater the uncertainty, the greater the information gain for zero or small prediction errors. The smaller the uncertainty, the greater the information gain for larger prediction errors. We term this intersection-related phenomenon the *arousal crossover effect*.

## A Functional Model of Emotional Valence

We next investigated how novelty affects the valence dimensions of positivity and negativity. Berlyne (1970) proposed collative variables that consist of stimulus factors, such as novelty, complexity, uncertainty, and conflict. Each collative variable has the quality of arousal potential (i.e., the ability to affect the intensity of arousal). Highly novel stimuli can increase arousal. Berlyne (1967, 1971) assumed that the hedonic qualities of stimuli arise from separate biological incentivization systems. The first system, the *reward system*, generates positive affect whenever arousal potential increases. The second system, the *aversion system*, generates negative affect whenever arousal potential increases. The aversion system has a higher absolute activation threshold than the reward system does. Thus, the joint operation of these two systems creates an inverted U-shaped curve, as shown in **Figure 3**. The valence of a stimulus changes from neutral to positive as the arousal potential increases but shifts from positive to negative after the arousal potential passes

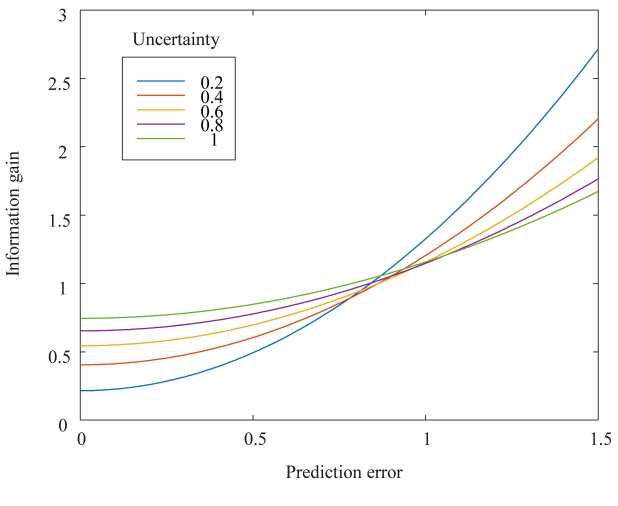


**FIGURE 2 |** Mathematically derived information gain, as a function of prediction errors, for uncertainty levels varying from 0.2 to 1.0. The external noise is set at 0.1.

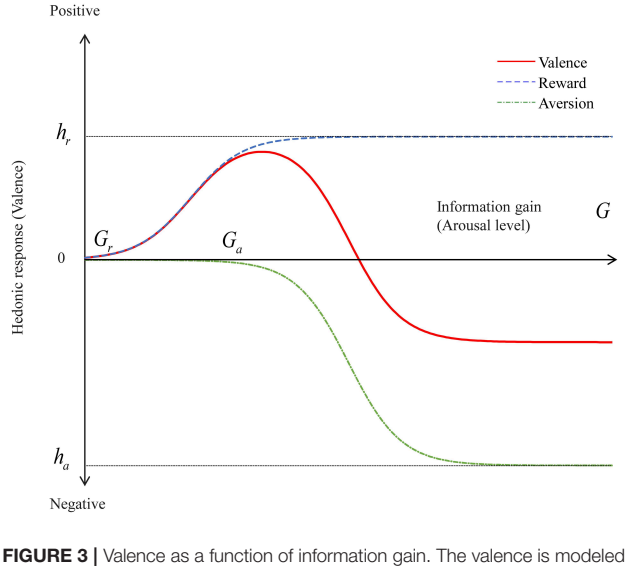


**FIGURE 3 |** Valence as a function of information gain. The valence is modeled as a summation of two sigmoidal functions representing reward and aversion systems.

the peak positive valence. This inverse U shape is reasonable. One may feel safe and experience boredom if stimuli are too familiar (i.e., not novel). Conversely, one may feel uncomfortable if stimuli are extremely unfamiliar and novel. However, in Bayesian models, repeated exposure to the same stimulus decreases both prediction errors and uncertainty. Thus, the iterative information gain for each update decreases. The decreasing information gain and the inverse U-shaped function may explain emotional desensitization, which is the psychological phenomenon of emotional responsiveness to a negative, aversive, or positive stimulus diminishing after repeated exposure to it. The positive hedonic response to a stimulus is diminished by decreasing information gain after repeated exposure to it, and a negative hedonic response to an extremely novel stimulus is shifted to a positive or neutral response by decreasing information gain after repeated exposure.

As noted in Overview, we formalized the arousal level as information gain from an event. If an event does not provide any information, then the valence can be neutral. At the opposite extreme, if an event provides excessive information that is difficult for the brain to process, then the valence can become negative. We can reasonably assume that between these two extremes there lies a "sweet spot" at which an optimum information gain maximizes a positive valence. We formalized valence as a summation of the reward and aversion systems and used sigmoid functions (Saunders, 2012) to model information gain for each system:

$$Valence = Reward + Aversion \tag{15}$$

$$where\ Reward(G) = \frac{h_r}{1 + \exp(-c_r G + G_r)} \tag{16}$$

$$and\ Aversion(G) = \frac{-h_a}{1 + \exp(-c_a G + G_a)}$$

In these formulae, $G_r$ and $G_a$ represent the thresholds of information gain that activate reward and aversion systems, respectively. The variables $h_r$ and $h_a$ are the maxima of positive and negative valence levels, respectively, and $c_r$ and $c_a$ represent the respective gradients. The condition $G_r < G_a$ must always be satisfied because the threshold of the reward system is lower than that of the aversion system. If an extreme information gain occurs, then the condition $h_r < h_a$ must be satisfied to obtain a negative valence. **Figure 3** shows the valence, reward, and aversion functions of formula (15). We can observe that the valence function is an inverse U-shaped curve.

## Model Summary

**Figure 4** shows a schematic of our proposed model. We formalized emotional arousal using information gain from an event, which we represented as the KL divergence from the prior to the posterior. We derived the information gain as a function of three parameters: uncertainty, the prediction error, and external noise, which are represented as the variance of the prior (or entropy), the difference between the prior expectation and the peak of the likelihood function, and the variance of the likelihood function, respectively. We formulated valence (i.e., positivity or negativity) as a summation of reward and aversion systems represented as information gain functions based on Berlyne's theory.

In our model, the information gain is a key parameter to explain the emotional dimensions of arousal and valence. The information gain increases as the prediction error increases. Recent neurological studies have shown that dopaminergic neurons encode the prediction error signal of reward (Schultz et al., 1997; Bayer and Glimcher, 2005). Our model explains a reward system as a function of information gain affected by prediction errors. From a mathematical analysis, we found that uncertainty and prediction errors have interaction effects on information gain. Prediction errors increase information gain. The greater the uncertainty, the more the information gain for zero or small prediction errors. In contrast, the smaller the uncertainty, the more the information gain for large prediction errors. Uncertainty represents the degree of belief in the prior expectation. The familiarity of an event or target and one's knowledge and experience of a target affect uncertainty. For example, if a product is so familiar that everyone knows it well, then uncertainty about the product is small. In contrast, if a product is unfamiliar, then

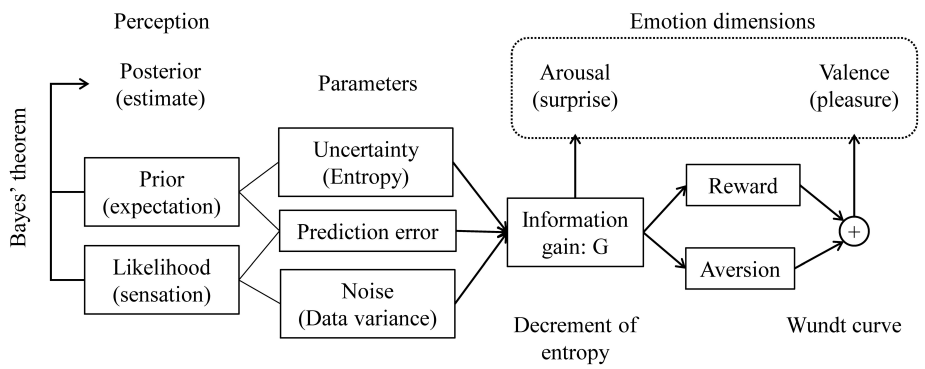


**FIGURE 4 |** Proposed model of the dimensions of novelty-elicited emotions. Emotional arousal is expressed as information gain from the Bayesian prior to the posterior. Valence is a summation of reward and aversion systems, which are functions of information gain.

uncertainty about the product should be considerable. Thus, uncertainty represents prior information before experiencing a target event. Indeed, uncertainty is proportional to the information entropy of the prior, as in formula (9). This model suggests that emotion is influenced by prior information, discrepancies between expectations and reality, and stimulus attributes.

# EFFECTS OF UNCERTAINTY AND PREDICTION ERRORS ON EMOTIONAL AROUSAL RELATED TO PERCUSSION INSTRUMENTS

We investigated the effects of uncertainty and prediction errors on surprise to validate the arousal crossover effect derived from the mathematical model in Interaction Effect of Uncertainty and Prediction Errors on Information Gain. Specifically, we tested the hypothesis that uncertainty increases surprise when prediction errors are small and decreases surprise when prediction errors are large. A set of short videos featuring percussion instruments and accompanying sounds were used as stimuli. In each video, a percussion instrument was presented and then beaten. Different percussive sounds were synthesized. We assumed a transition from a visual prior (i.e., the appearance of an instrument) to an auditory posterior (i.e., the percussive sound). Participants predicted an instrument's sound from its appearance and then listened to a sound. We induced prediction errors by manipulating the congruency between the synthesized percussive sounds and the instrument shown. We assumed that prediction errors were large when the synthesized percussive sounds were incongruent with the instruments shown, and we assumed that the familiarity or unfamiliarity of the instruments shown produced different levels of uncertainty. The appearance of a familiar percussion instrument, such as a hand drum, produces certainty of expectations concerning its sound (i.e., a small uncertainty). The appearance of an unfamiliar percussion instrument, such as the African percussion instrument known as the jawbone, produces uncertain expectations concerning its sound (i.e., a large uncertainty).

We used both questionnaires and ERP recordings to assess participants' levels of surprise in response to the percussive sound in each video. We quantified surprise intensities based on responses to a four-level Likert scale and measurements of ERP P300 amplitudes (Mars et al., 2008).

## Methods

### Participants

Nine right-handed healthy male volunteers (mean age ± standard deviation: 21.7 ± 1.2 years; range: 20–24 years) with normal or corrected-to-normal vision and hearing participated in this study. The study protocol was approved by the Ethics Committee of the Graduate School of Engineering at the University of Tokyo. In accordance with the principles of the Declaration of Helsinki, all participants provided written informed consent prior to their participation in this study. The participants were allowed to interrupt the experiment sessions at their convenience.

**TABLE 1 |** Combinations of percussion instruments and percussive sounds. (Video stimuli are available in the **Supplementary Material**).

| | Instrument | Congruent sound (X) | Incongruent sound (Y) |
|---|---|---|---|
| Familiar (A) | Clave | Clave (AX), (**Video S1**) | Bell (AY), (**Video S3**) |
| | Hand drum | Hand drum (AX), (**Video S2**) | Guiro (AY), (**Video S4**) |
| Unfamiliar (B) | Jawbone | Jawbone (BX), (**Video S5**) | Vibraphone (BY), (**Video S7**) |
| | Slit drum | Slit drum (BX), (**Video S6**) | Snare (BY), (**Video S8**) |

### Stimuli

The stimuli consisted of eight short videos in which a percussion instrument was beaten once and a synthesized percussive sound followed. **Table 1** shows the combinations of instruments shown and the synthesized sounds (Videos are available in **Supplementary Material**). The clave and hand drum were selected as familiar percussion instruments (type A), and the jawbone and slit drum were selected as unfamiliar percussion instruments (type B). To create incongruent conditions, we synthesized percussive sounds that were inconsistent with the instruments shown. Our stimuli included videos with visually familiar instruments and congruent sounds (type AX), videos with visually familiar instruments and incongruent sounds (type AY), videos with visually unfamiliar instruments and congruent sounds (type BX), and videos with visually unfamiliar instruments and incongruent sounds (type BY).

The duration of each video was 2,500 ms. First, a percussion instrument appeared in the center of the screen. The percussion instrument was then beaten once 500 ms into the video while a percussive sound was presented simultaneously. Each video had an 18° horizontal visual angle and a 10° vertical visual angle and was centrally presented against a black background on a 29.8-inch display located 100 cm away from the participant. The participants wore noise-canceling headphones covered by earmuffs while watching the videos.

### Procedure

The participants completed experiments individually in an electromagnetically shielded dark room. After participants received instructions for the procedure, they were asked to start the experiment.

First, we conducted sound-only experiments in which we attempted to ensure uniform surprise levels in response to the percussive sounds used in each video type (i.e., AX, AY, BX, and BY). Achieving this uniformity was necessary so that we could be sure that our observations in later experiments with audiovisual stimuli reflected the effects of visual priors. The eight percussive sounds were presented to the participants via headphones in five random-order sets without any visual stimuli. This phase of the procedure consisted of 40 trials (eight sounds × five presentation sets). The interstimulus interval (ISI) was 1,000–2,000 ms, with an average of 1,500 ms.

Second, we conducted additional sound-only experiments in which we used electroencephalography (EEG) to confirm the uniformity of the surprise levels evoked by the percussive sounds

of each video type. The eight percussive sounds were presented to the participants via headphones in 20 random-order sets without any visual stimuli. This phase of the procedure consisted of 160 trials (eight sounds × 20 presentation sets). The ISI was 1,000–2,000 ms, with an average of 1,500 ms. EEG recordings were obtained for each trial. A short break was inserted after the tenth presentation set.

Third, the participants watched videos of a clave or a hand drum, which we assumed were familiar instruments for our participants, accompanied by congruent percussive sounds. The videos thus belonged to type AX. The participants watched these videos five times to create expectations of certainty and congruity.

Lastly, we conducted the main experiment in which participants watched videos while undergoing EEG recordings and subjectively reporting feelings of surprise. The eight videos described in **Table 1** were presented to the participants in 20 random-order sets. This phase of the procedure consisted of 160 trials (eight videos × 20 presentation sets). The ISI was 1,000–2,000 ms, with an average of 1,500 ms. EEG recordings were obtained for each trial. A short break was inserted after the tenth presentation set. During the first, tenth, and final presentation sets, the participants used a four-level Likert scale to report the intensities of their surprise upon listening to the percussive sounds. The participants used four push buttons under their fingers to provide these reports so that they did not have to avert their eyes from the display.

### EEG Recordings

The EEG data were recorded with a portable digital recorder (Polymate AP1132, TEAC Corporation, Tokyo, Japan) and active electrodes. The data were obtained from three midline electrodes positioned at the Fz, Cz, and Pz points as defined by the international 10–20 system with reference to the nose. The data were recorded at a sampling rate of 500 Hz. The time constant was set at 3 s. All electrode impedances were below 50 k$\Omega$. A digital bandpass filter of 0.1–20 Hz was applied.

### EEG Data Analysis

The ERP waveforms were obtained by averaging data from the period starting 200 ms before the stimulus onset, which we define as the start of the video in video stimulus sessions, and ending 1,500 ms after the stimulus onset. This averaging was done separately for each participant, stimulus type (i.e., AX, AY, BX, and BY), and electrode site for both the sound-only and video stimuli. For each averaged waveform, the 200-ms period preceding the stimulus onset was defined as the baseline. Any epochs containing EEG signals exceeding ± 100 μV were regarded as eye movement–related artifacts and automatically removed. The P300 component was designated as the largest positive peak occurring 250–600 ms after the onset of the percussive sound. The baseline-to-peak P300 amplitudes were measured at the Pz point, which was the dominant electrode site.

### Statistical Analysis

Repeated-measures analysis of variance (ANOVA) was applied to the ERP and Likert scale data. One-way ANOVA of the P300 data from the sound-only sessions was conducted to examine how different percussive sound types affected P300 amplitudes. To identify interaction effects on surprise intensities, we analyzed the P300 amplitude and Likert scale data from the video sessions with two-way ANOVA in terms of congruity and familiarity. Statistical significance was defined as $p < 0.05$ for all statistical tests. We compared the experimental results to the simulation results shown in **Figure 2**.

## Experimental Results

The type of percussive sound did not significantly affect P300 amplitudes in the sound-only sessions ($F = 0.35$, $p = 0.79$).

**Figure 5** shows the grand mean ERP waveforms for the four video types in the main video session. Under the congruent condition, the sounds of unfamiliar percussion instruments (type BX) elicited larger P300 amplitudes than the sounds of familiar percussion instruments (type AX) did. However, under the incongruent condition, the sounds of familiar percussion instruments (type AY) elicited larger P300 amplitudes than the sounds of unfamiliar percussion instruments (type BY) did.

**Figure 6** shows the averaged P300 amplitude for each condition (i.e., all combinations of congruity and familiarity) in the main video session. The interaction effect of congruity and familiarity on P300 amplitudes was significant ($F = 10.99$, $p = 0.01$). The simple main effect of familiarity was significant for both congruent ($F = 4.7$, $p = 0.047$) and incongruent ($F = 11.82$, $p = 0.004$) sounds. When congruent sounds were played, the average P300 amplitude for the unfamiliar instruments was larger than that for the familiar instruments, but when incongruent sounds were played, the average P300 amplitude for the unfamiliar instruments was smaller than that for the familiar instruments. The simple main effect of congruity was significant for the familiar instruments ($F = 6.5$, $p = 0.02$) but not for the unfamiliar instruments ($F = 3.09$, $p = 0.09$). Thus, the average P300 amplitude evoked by incongruent sounds was larger than that evoked by congruent sounds only when familiar instruments were shown.

**Figure 7** shows the average Likert scale surprise rating for each stimulus used in the main video session under different conditions of congruity and familiarity. The interaction effect of congruity and familiarity was significant ($F = 39.06$, $p < 0.001$), as was the simple main effect of congruity ($F = 144.9$, $p < 0.001$). The simple main effect of familiarity was significant for both congruent ($F = 167.14$, $p < 0.001$) and incongruent ($F = 16.72$, $p < 0.001$) sounds. The difference between Likert scale surprise ratings for the familiar and unfamiliar instruments was significant under both congruent and incongruent sound conditions ($p < 0.01$). These results show that subjectively rated surprise under the unfamiliar instrument condition was greater than that under the familiar instrument condition when the sounds were congruent but that subjectively rated surprise under the familiar instrument condition was greater than that under the unfamiliar instrument condition when the sounds were incongruent. The crossover in both **Figures 6**, **7** corresponds to the simulation result in **Figure 2**.

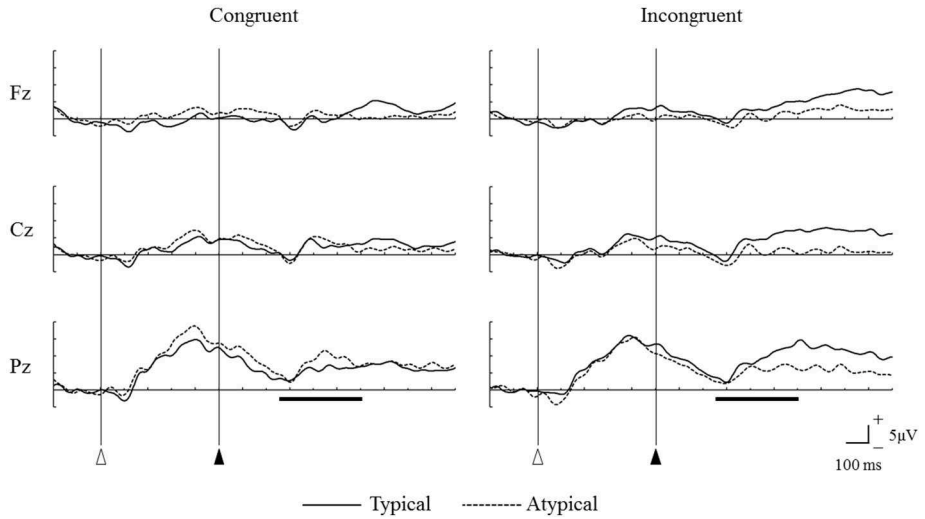


**FIGURE 5 |** Grand mean event-related potential waveforms for the four different video types measured from frontal (Fz), central (Cz), and parietal (Pz) midline regions. Open triangles represent the onsets of the videos, and solid triangle represent the onsets of percussive sounds. The horizontal bars show the time range of 250–600 ms after the onset of the percussive sound.

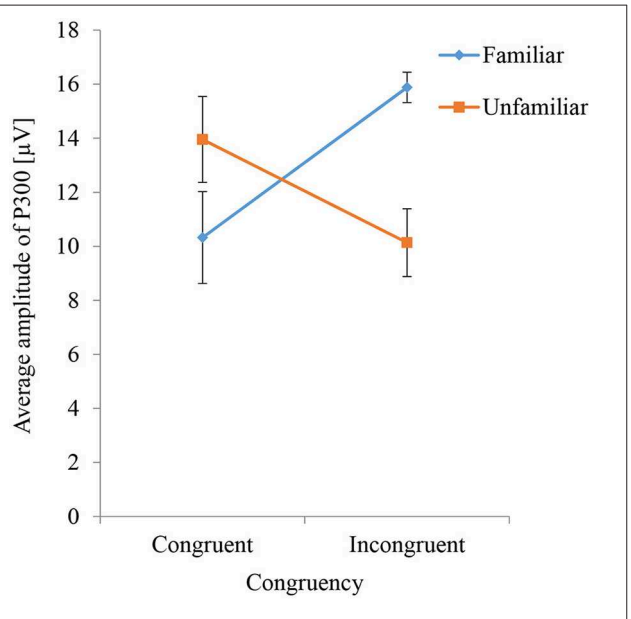


**FIGURE 6 |** P300 amplitudes evoked by percussive sounds that are congruent or incongruent with the instrument shown. The results for familiar and unfamiliar instruments are compared.

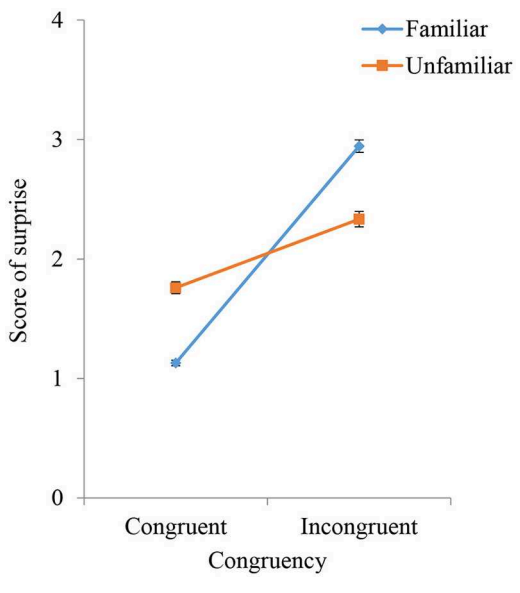


**FIGURE 7 |** Subjectively reported scores for surprise intensities in response to percussive sounds that are congruent or incongruent with the instrument shown. The results for familiar and unfamiliar instruments are compared.

## DISCUSSION

We assumed that information gain from an event, which can be calculated using KL divergence between the Bayesian prior and the posterior, represents the intensity of arousal emotions such as surprise. Prediction errors, which are differences between prior expectations and likelihood function peaks, increase information gain and surprise. We conducted an experiment featuring videos of percussion instruments accompanied by synthesized percussive sounds. We varied uncertainty levels by using familiar and unfamiliar instruments, and we varied prediction error magnitudes by using congruent or incongruent percussive sounds. We used ERP P300 amplitudes and subjective reports to assess the participants' surprise levels in response to the percussive sounds. Compared to congruent sounds, incongruent sounds produced greater subjectively

reported surprise intensities, and this was particularly true when familiar percussion instruments were shown. Similarly, incongruent sounds increased P300 amplitudes when familiar percussion instruments were shown. These results suggest that prediction errors related to visuoauditory incongruities increase surprise, attention, and the amount of information processed in the brain (i.e., the arousal level). Moreover, instrument familiarity, which induces certainty of expectations concerning percussive sounds, provides a greater potential for arousal in the event of visuoauditory incongruity than is possible with unfamiliar instruments, which induce uncertainty of expectations concerning sounds. This result supports our mathematical hypothesis that information gain serves as an index of arousal.

We mathematically derived a hypothesized effect that we termed the *arousal crossover effect*: uncertainty, represented as variance of the prior, increases information gain when prediction errors are zero or small, but uncertainty decreases information gain when prediction errors are large. Both the P300 amplitude data and the subjectively reported surprise intensity data supported this hypothesized effect. When congruent sounds accompanied the instruments shown, videos featuring unfamiliar instruments evoked greater P300 amplitudes and subjectively reported surprise scores than videos featuring familiar instruments did. However, when incongruent sounds accompanied the instruments shown, videos featuring unfamiliar instruments evoked lower P300 amplitudes and subjectively reported surprise scores than videos featuring familiar instruments did.

This concordance between our proposed model's predictions and the experimental results suggests that information gain obtained from a novel event represents the level of emotional arousal. Previous studies have shown that the KL divergence represents surprise that attracts human attention (Itti and Baldi, 2009). We newly formalized the information gain, which is mathematically equivalent to KL divergence, as a function of prediction errors, uncertainty, and noise and showed both mathematically and experimentally that an interaction effect of prediction errors and uncertainty exists. Uncertainty of the prior depends on an individual's knowledge and prior experiences as well as the familiarity of an event. Prior knowledge and experience produce certainty of expectations. This implies that our proposed model may explain individual differences in emotional responses to an identical novel event as resulting from differences in knowledge and prior experience. For example, an expert's expectations should be more certain than those of a novice. Using our model, we can therefore predict that novices are more surprised than experts are when an event differs marginally from prior expectations but that experts are more surprised than novices are when an event greatly differs from prior expectations.

We formalized emotional valence as a function of arousal levels based on Berlyne's theory (Berlyne, 1970). The functional model forms an inverse U-shaped curve that has a positive valence peak at a certain arousal level. Therefore, we can predict that variable uncertainty levels related to an individual's knowledge and experience and the familiarity of an event modulate the effect of prediction errors on valence responses. Although our mathematical model is firmly grounded in Berlyne's theory, further experimental evidence validating the ability of our valence model to predict empirical observations will be more than welcomed. Indeed, the chief limitation of this study is the reliance on mathematical formulations of both arousal and valence and the lack of experimental validation of our formulation of valence. In future studies, we will conduct experiments to test the validity of our valence model.

## AUTHOR CONTRIBUTIONS

HY designed and supervised the study and formalized the mathematical model. HY, OK, and KU designed and conducted the experiment. OK and KU measured and analyzed the EEG data. HY and KU drafted the manuscript. All authors revised the manuscript.

## FUNDING

This study was supported by KAKEN grant number 18H03318 from the Japan Society for the Promotion of Science.

## ACKNOWLEDGMENTS

We thank Prof. Tamotsu Murakami and the members of the Design Engineering Laboratory at the University of Tokyo for supporting this project.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2019.00002/full#supplementary-material

**Video S1 |** Clave (AX).

**Video S2 |** Hand drum (AX).

**Video S3 |** Clave - Bell (AY).

**Video S4 |** Hand drum - Guiro (AY).

**Video S5 |** Jawbone (BX).

**Video S6 |** Slit drum (BX).

**Video S7 |** Jawbone - Vibraphone (BY).

**Video S8 |** Slit drum - Snare (BY).

## REFERENCES

Bayer, H. M., and Glimcher, P. W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47, 129–141. doi: 10.1016/j.neuron.2005.05.020

Berlyne, D. E. (1967). Arousal and reinforcement. *Nebr. Symp. Motiv.* 15, 1–110.

Berlyne, D. E. (1970). Novelty, complexity, and hedonic value. *Percept. Psychophys.* 8, 279–286. doi: 10.3758/bf03212593

Berlyne, D. E. (1971). *Aesthetics and Psychobiology*. New York, NY: Appleton-Century-Crofts.

Ernst, M. O., and Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415, 429–433. doi: 10.1038/415429a

Giacalone, D., Duerlund, M., Bøegh-Petersen, J., Bredie, W. L. P., and Frøst, M. B. (2014). Stimulus collative properties and consumers' flavor preferences. *Appetite* 77, 20–30. doi: 10.1016/j.appet.2014.02.007

Hekkert, P., Snelders, D., and van Wieringen, P. C. W. (2003). Most advanced, yet acceptable: typicality and novelty as joint predictors of aesthetic preference in industrial design. *Br. J. Psychol.* 94, 111–124. doi: 10.1348/000712603762842147

Itti, L., and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Res.* 49, 1295–1306. doi: 10.1016/j.visres.2008.09.007

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference. *Annu. Rev. Psychol.* 55, 271–304. doi: 10.1146/annurev.psych.55.090902.142005

Körding, K. P., and Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature* 427, 244–247. doi: 10.1038/nature02169

Körding, K. P., and Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends Cogn. Sci.* 10, 319–326. doi: 10.1016/j.tics.2006.05.003

Kullback, S., and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* 22, 79–86. doi: 10.1214/aoms/1177729694

Lang, P. J. (1995). The emotion probe: studies of motivation and attention. *Am. Psychol.* 50, 372–385. doi: 10.1037/0003-066X.50.5.372

Loewy, R. (1951). *Never Leave Well Enough Alone: The Personal Record of an Industrial Designer From Lipsticks to Locomotives*. New York, NY: Simon and Schuster.

Ma, W. J., Beck, J. M., Latham, P. E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nat. Neurosci.* 9, 1432–1438. doi: 10.1038/nn1790

Mars, R. B., Debener, S., Gladwin, T. E., Harrison, L. M., Haggard, P., Rothwell, J. C., et al. (2008). Trial-by-trial fluctuations in the event-related electroencephalogram reflect dynamic changes in the degree of surprise. *J. Neurosci.* 28, 12539–12545. doi: 10.1523/jneurosci.2925-08.2008

Mauss, I. B., and Robinson, M. D. (2009). Measures of emotion: a review. *Cogn. Emot.* 23, 209–237. doi: 10.1080/02699930802204677

Russell, J. A. (1980). A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 1161–1178. doi: 10.1037/h0077714

Saunders, R. (2012). Towards autonomous creative systems: a computational approach. *Cognit. Comput.* 4, 216–225. doi: 10.1007/s12559-012-9131-x

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science* 275, 1593–1599. doi: 10.1126/science.275.5306.1593

Shannon, C. E., Weaver, W., Blahut, R. E., and Hajek, B. (1949). *The Mathematical Theory of Communication*. Champaign, IL: University of Illinois Press.

Silvia, P. J. (2005). Emotional responses to art: from collation and arousal to cognition and emotion. *Rev. Gen. Psychol.* 9, 342–357. doi: 10.1037/1089-2680.9.4.342

Stocker, A. A., and Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* 9, 578–585. doi: 10.1038/nn1669

Yanagisawa, H. (2016). A computational model of perceptual expectation effect based on neural coding principles. *J. Sens. Stud.* 31, 430–439. doi: 10.1111/joss.12233

**Conflict of Interest Statement:** 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.

frontiers in Computational Neuroscience

BRIEF RESEARCH REPORT
published: 15 February 2019
doi: 10.3389/fncom.2019.00009

# Enhancing Diagnosis of Autism With Optimized Machine Learning Models and Personal Characteristic Data

*Milan N. Parikh*[1], *Hailong Li*[1] *and Lili He*[1,2]*

[1]Perinatal Institute, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States, [2]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, United States

Autism spectrum disorder (ASD) is a developmental disorder, affecting about 1% of the global population. Currently, the only clinical method for diagnosing ASD are standardized ASD tests which require prolonged diagnostic time and increased medical costs. Our objective was to explore the predictive power of personal characteristic data (PCD) from a large well-characterized dataset to improve upon prior diagnostic models of ASD. We extracted six personal characteristics (age, sex, handedness, and three individual measures of IQ) from 851 subjects in the Autism Brain Imaging Data Exchange (ABIDE) database. ABIDE is an international collaborative project that collected data from a large number of ASD patients and typical non-ASD controls from 17 research and clinical institutes. We employed this publicly available database to test nine supervised machine learning models. We implemented a cross-validation strategy to train and test those machine learning models for classification between typical non-ASD controls and ASD patients. We assessed classification performance using accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC). Of the nine models we tested using six personal characteristics, the neural network model performed the best with a mean AUC (SD) of 0.646 (0.005), followed by k-nearest neighbor with a mean AUC (SD) of 0.641 (0.004). This study established an optimal ASD classification performance with PCD as features. With additional discriminative features (e.g., neuroimaging), machine learning models may ultimately enable automated clinical diagnosis of autism.

**Keywords: autism spectrum disorder, machine learning, diagnosis, biostatistics, support vector machine**

OPEN ACCESS

***Edited by:***
*Dan Chen,*
*Wuhan University, China*

***Reviewed by:***
*Roberto Santana,*
*University of the Basque Country, Spain*
*Jiannan Kang,*
*Beijing Normal University, China*

***Correspondence:**
*Lili He*
*lili.he@cchmc.org*

***Received:*** *09 September 2018*
***Accepted:*** *30 January 2019*
***Published:*** *15 February 2019*

***Citation:***
*Parikh MN, Li H and He L (2019) Enhancing Diagnosis of Autism With Optimized Machine Learning Models and Personal Characteristic Data. Front. Comput. Neurosci. 13:9. doi: 10.3389/fncom.2019.00009*

## INTRODUCTION

Autism spectrum disorder (ASD) is characterized by impaired linguistic, communication, cognitive and social skills (Wetherby and Prutting, 1984). Therapies have been developed to treat the varying degrees of symptoms and improve patient quality of life. However, the diagnosis of ASD remains challenging, especially for marginal cases, resulting in under- and over-diagnosis. To date, behavior-based tests are the standard clinical approach to diagnosing ASD (American Psychiatric Association, 2013). The diagnostic process for ASD is time-consuming and costly (Galliver et al., 2017). An automated ASD diagnostic approach might allow for earlier identification of ASD and could help provide a map of high-risk populations.

**Abbreviations:** ABIDE, Autism Brain Imaging Data Exchange; ASD, Autism Spectrum Disorder; AUC, Area Under the Receiver Operating Characteristic Curve; PCD, Personal Characteristic Data; SVM, Support Vector Machine.

Emerging machine learning approaches are showing great promise for objective evaluation of neuropsychiatric disorders (Nielsen et al., 2013; Bone et al., 2015; Chen et al., 2015; Plitt et al., 2015; Ghiassian et al., 2016; Yahata et al., 2016; Abraham et al., 2017).

Machine learning is a group of statistical techniques that learn with the distribution of data so as to make decisions on new data. It is used to devise complex applications to make accurate classifications/predictions on diverse data (Russell and Norvig, 2010). Autism diagnosis could be formulated as a typical classification problem (i.e., ASD vs. typical control/non-ASD). The constructed model/classifier is then able to evaluate whether a new unknown subject has ASD or not based on input features.

Several studies have employed machine learning to improve ASD diagnosis. Duda et al. (2016) applied machine learning to distinguish autism from attention deficit hyperactivity disorder using a 65-item Social Responsiveness Scale. Bone et al. (2015) trained their models to diagnose autism against healthy controls using the same Social Responsiveness Scale and the Autism Diagnostic Interview-Revised scores. More recently, the Autism Brain Imaging Data Exchange (ABIDE) has gathered data [i.e., personal characteristic data (PCD), structural MRI, functional MRI] from over 1,000 subjects and made it available for the ASD research community (Craddock et al., 2013). This has facilitated the development of machine learning models towards the automated diagnosis of ASD (Ghiassian et al., 2016; Abraham et al., 2017; Heinsfeld et al., 2018; Li et al., 2018). While most studies have focused on brain neuroimaging data, few studies have reported automated machine learning models that solely rely on PCD as input features. As such, the full potential of PCD on ASD classification has yet to be comprehensively evaluated. It is important to note that a true diagnostic classifier of ASD cannot be created due to the retrospective case-control ABIDE study design. In this work, we simply set out to assess the predictive power of PCD for ASD diagnosis and evaluate which machine learning model is most robust for this task. Specifically, we employed and validated nine machine learning models by using PCD, such as age, sex, handedness, and IQ, for ASD classification of individual subjects. Taking advantage of such a large PCD dataset from ABIDE, we systematically evaluated the predictive power of PCD features on ASD classification and compared the performance of those nine machine learning models.

# MATERIALS AND METHODS

## Data

We selected six PCD features of interest—age at testing, sex, handedness, full-scale IQ, verbal IQ, performance IQ—from the ABIDE I Preprocessed Database. Only subjects with information for all 6 features were included ($N$ = 851 of total of 1,112 subjects in ABIDE I database). Of the 851 subjects, 430 were typical non-ASD controls and 421 had a confirmed diagnosis of ASD. To control for site effects, we included site of testing in each of the models. Using a two-sided Student's $t$-test (unequal variance), we identified significant differences between ASD patients and healthy controls in full-scale IQ ($p < 0.001$), verbal IQ ($p < 0.001$), and performance IQ ($p = 0.003$); there was no significant group difference in age ($p = 0.8582$). Sex ($p = 0.017$) and handedness ($p = 0.018$) were also significantly different between groups (chi-squared test; **Table 1**).

**TABLE 1** | Demographic information for our sub-sample of the Autism Brain Imaging Data Exchange (ABIDE) Database.

| Group | ASD ($N$ = 421) | Control ($N$ = 430) | $P$ |
|---|---|---|---|
| Age | 16.8 ± 7.7 | 16.7 ± 6.9 | 0.858 |
| Full-Scale IQ | 105.2 ± 16.8 | 110.9 ± 12.6 | <0.001 |
| Verbal IQ | 104.4 ± 17.8 | 111.3 ± 13.3 | <0.001 |
| Performance IQ | 105.0 ± 17.2 | 108.2 ± 13.3 | 0.003 |
| Sex (%) | | | 0.017 |
| Male | 88 | 82 | |
| Female | 12 | 18 | |
| Handedness (%) | | | 0.018 |
| Left | 13 | 6 | |
| Right | 85 | 92 | |
| Ambidextrous | 2 | 1 | |

*All data are mean ± SD unless otherwise specified.*

A portion of the ABIDE study sites defined handedness as a score based on the Edinburgh Handedness Inventory while others coded it as a category (left, right, or ambidextrous). Thus, we reformatted all handedness data to categorical values. This study included 15 different ABIDE recruitment sites. These were included in the features to control for site of testing.

## Classification Models

In order to comprehensively evaluate the full potential of PCD for ASD classification, we tested a variety of approaches, including k-nearest neighbor (Altman, 1992), linear and nonlinear Support Vector Machine (SVM; Cortes and Vapnik, 1995), decision tree (Breiman et al., 1984), logistic regression (Dobson, 1990), Stacked Sparse Auto-encoder (SSAE)-based neural network (Hinton and Salakhutdinov, 2006), random forest (Breiman, 2001), and majority voting and weighted average ensemble models (Cruz and Wishart, 2006; Zhou, 2012). The models are detailed in the **Supplementary Materials**.

To optimize the performance of each model, we performed a parameters grid search (Cuingnet et al., 2011) for each model (**Supplementary Table S1**; **Supplementary Materials**).

## Model Evaluation

We applied a k-fold cross-validation scheme to train and test the models. The whole dataset was randomly divided into 25 equal sized portions. Of the 25 portions, one portion of data was held out for model testing, and the remaining 24 portions were used for model training. In order to create a validation dataset for model optimization, a 10-fold cross-validation was performed on the training dataset for each model (**Supplementary Materials**; **Supplementary Figure S1**). This process was repeated until each of the 25 portions was evaluated once as the testing data. We evaluated the model based on the concatenated test labels and ground truth labels across 25 iterations. We repeated this k-fold cross-validation 30 times.

The performance of the classification was assessed using four diagnostic metrics: accuracy, sensitivity, specificity and area under the receiver operating characteristic curve (AUC). Accuracy is measured as the percentage of correctly classified subjects within all subjects. Sensitivity is defined as the percentage of correctly classified ASD subjects within all ASD subjects, while specificity is represented by the percentage of correctly classified healthy subjects within all typical non-ASD control subjects. Sensitivity is the ability of the classifier to correctly identify ASD subjects (true positive rate), whereas specificity is the ability of the classifier to correctly identify healthy subjects (true negative rate). AUC reflects the diagnostic ability of a binary classifier system when its discrimination cutoff varies.

## RESULTS

From the models we tested using all six PCD features, we found that the model with the best AUC was the Stacked Sparse Auto-encoder (SSAE)-based neural network ($p < 0.001$) which correctly classified ASD patients with a mean (SD) accuracy of 62.0% (0.9%) and AUC of 0.646 (0.005; **Table 2**). The k-nearest neighbor model displayed an accuracy of 61.8% (0.8%) and the second highest AUC of 0.641 (0.004), but its sensitivity was lower than most models. Compared to this, both linear and non-linear SVM yielded better performance considering overall diagnostic measures.

Using a feature selection method based on the Student's *t*-test, we noted that the most predictive features were full-scale IQ, followed by verbal IQ and performance IQ. By using only these three features, the neural network achieved an AUC (SD) of 0.641 (0.009) which was very comparable to the AUC using all seven features. By removing females ($n = 126$) and only considering male subjects ($n = 725$), the diagnostic performance for neural network was also comparable with an accuracy of 61.1% (1.3%) and AUC of 0.645 (0.014).

## DISCUSSION

This study set out to explore the full potential of PCD as diagnostic features for ASD classification. We developed and compared nine automated machine learning models by using a large PCD dataset from the ABIDE repository. In our evaluation, our neural network model outperformed eight other peer models by achieving the best AUC of 0.646.

PCD have demonstrated strong predictive power for other neurodevelopmental disorders. For example, in the ADHD-200 global competition, PCD features outperformed fMRI features in attention deficit hyperactivity disorder classification (Brown et al., 2012). This inspired us to test the predictive power of PCD for ASD classification. Previous studies using PCD for ASD classification have been limited, and optimal performance for PCD has not been established. In recent studies, PCD were only investigated for the purpose of feature fusion or integration. For instance, Ghiassian et al. (2016) reported an accuracy of 59.6% with non-linear SVM using the same six PCD features and eye stat (eyes open or closed). However, they investigated PCD performance only for model comparison. In addition, their results were based on one classifier whereas we tested multiple classifiers to determine not only the best performance but also the model that consistently yielded the best performance. Finally, when we used the same dataset as Ghiassian et al. (2016) in our neural network model, we obtained a somewhat higher accuracy of 62.3%. Nevertheless, these differences might have also resulted from other factors such as study differences in cross-validation. The more important takeaway is that the six PCD we tested, and particularly the three IQ measures, provide significant predictive power for ASD diagnosis that should be incorporated into future ASD classification studies.

Our results highlight the advantage of neural networks over other commonly employed machine learning models in ASD classification. Traditionally, neural network models have had a significantly higher computational cost than other peer models. With recent rapid advances in deep learning techniques, the current techniques have reduced the optimization process for neural networks to an acceptable training time. As shown in **Supplementary Table S1**, the neural network model has more hyperparameters which provide the model with additional flexibility to learn the PCD distribution for ASD classification. Interestingly, k-nearest neighbor had the second-best AUC among our nine models, but its sensitivity in our experiment was not desirable. Compared to this, both linear and non-linear SVM yielded better performance considering overall diagnostic measures.

In addition, our results compare favorably to recent predictions made using fMRI features from a similar sample

**TABLE 2** | Accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) values for each machine learning model.

| | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC |
|---|---|---|---|---|
| Decision tree | 54.7 ± 1.5 | 53.3 ± 2.0 | 54.9 ± 1.7 | 0.562 ± 0.015 |
| Majority model | 61.9 ± 0.8 | 55.4 ± 1.1 | 69.2 ± 1.3 | 0.568 ± 0.009 |
| Random forest | 57.2 ± 0.8 | 54.4 ± 1.2 | 60.4 ± 1.1 | 0.615 ± 0.007 |
| SVM (linear) | 61.4 ± 0.5 | 57.1 ± 0.6 | 66.7 ± 0.8 | 0.622 ± 0.002 |
| SVM (non-linear) | 61.9 ± 0.4 | 52.3 ± 1.5 | 71.6 ± 1.1 | 0.623 ± 0.005 |
| Confidence model | 61.5 ± 0.9 | 49.1 ± 1.4 | 67.1 ± 1.0 | 0.633 ± 0.008 |
| Logistic regression | 59.1 ± 0.5 | 55.5 ± 0.6 | 62.6 ± 0.8 | 0.635 ± 0.001 |
| k-Nearest neighbor | 61.8 ± 0.6 | 46.6 ± 1.0 | 72.1 ± 0.8 | 0.641 ± 0.004 |
| Neural network | 62.0 ± 0.9 | 53.3 ± 1.3 | 71.2 ± 1.9 | 0.646 ± 0.005 |

*All data are mean ± SD; SVM, Support Vector Machine.*

of the ABIDE database (Abraham et al., 2017). That study achieved a maximum accuracy (SD) of 66.8% (5.4%). Although our model with PCD had a lower accuracy of 62.0%, our standard deviation of 0.9% is substantially lower (i.e., narrower confidence interval) than their model. Additionally, our model only requires six simple PCD features which are low-cost and easy-to-obtain as compared to neuroimaging data. These performance scores compared to fMRI-based classification emphasize the importance of PCD in ASD classification.

The main limitations of our study arise from how the ABIDE data were collected. This international study collected data from 17 unique clinical and research sites. This leads to heterogeneity in the data that might compromise the machine learning models. To mitigate the impact of site bias, we controlled for the site of testing by including it in all the models. However, the heterogeneity of PCD data may require further investigation before such models can be utilized in clinical settings. The small sex difference in ASD vs. controls we observed is likely a function of the high incidence of ASD in males rather than a selection bias for this substudy. Even if this was a biased selection from ABIDE, our secondary analyses in only males from this subpopulation yielded very similar results to our primary analysis that included both sexes, suggesting this difference did not affect performance or bias our results. Another limitation is the size of the dataset. While 851 subjects are considered a large study in this field of clinical research, larger datasets may be needed to yield generalizable machine learning models. Also, our ASD classifiers specifically focused on the classification of ASD and would not be effective in detecting the presence of other developmental disorders. A large prospective study of a more heterogeneous population would be required to confirm the value of PCD and/or other promising features to diagnose ASD.

Future efforts could include combining PCD with neuroimaging data using machine learning models. Along with the addition of fMRI features, the use of other features, such as medical tests or past or family history of disease, might boost the performance of the models to a clinically useful level. The addition of more features may also increase the performance of neural networks and allow for the use of more complex architecture of neural networks. Studies testing new machine learning models show promising results using fMRI features (He et al., 2018; Li et al., 2018). A recent development in machine learning, called transfer learning, mimics the human brain by using large amounts of available information unrelated to the disease of interest (e.g., typical controls) to draw conclusions when presented with a smaller, less accessible amount of information about the disease of interest. Transfer learning has already been shown to improve classification and identify networks in the brains of high-risk premature birth babies (He et al., 2018) and diagnose autism on small subsets of the ABIDE database (Li et al., 2018).

In summary, we developed and compared nine machine learning models for ASD classification by using PCD as input features. We conclude that combining PCD with optimized machine learning models can enhance diagnosis of ASD. When integrated with additional features (e.g., fMRI features), these models have the potential to yield a more objective approach for diagnosing autism.

## DATA AVAILABILITY

The dataset analyzed for this study was the international ABIDE dataset, which can be accessed here: www.preprocessed-connectomes-project.org/abide/index.html. All generated data for this study are included in the manuscript.

## AUTHOR CONTRIBUTIONS

LH and HL conceived the project. MP organized the experiments and analyzed the data with guidance from HL. MP wrote the first draft of the manuscript. All authors contributed to the manuscript revision, read and approved the submitted version.

## FUNDING

This study was supported by the National Institutes of Health (grant numbers R21-HD094085, R01-NS094200, and R01-NS096037).

## ACKNOWLEDGMENTS

We thank the ABIDE project investigators for making their data publicly available. We also thank Nehal Parikh, DO, MS for reviewing an earlier version of this manuscript.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2019.00009/full#supplementary-material

## REFERENCES

Abraham, A., Milham, M. P., Di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., et al. (2017). Deriving reproducible biomarkers from multi-site resting-state data: an autism-based example. *Neuroimage* 147, 736–745. doi: 10.1016/j.neuroimage.2016.10.045

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *Am. Stat.* 46, 175–185. doi: 10.2307/2685209

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders.* Washington, DC: American Psychiatric Publishing.

Bone, D., Goodwin, M. S., Black, M. P., Lee, C. C., Audhkhasi, K., and Narayanan, S. (2015). Applying machine learning to facilitate autism diagnostics: pitfalls and promises. *J. Autism Dev. Disord.* 45, 1121–1136. doi: 10.1007/s10803-014-2268-6

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Boca Raton, FL: Chapman and Hall.

Brown, M., Sidhu, G., Greiner, R., Asgarian, N., Bastani, M., Silverstone, P., et al. (2012). ADHD-200 global competition: diagnosing ADHD using personal

characteristic data can outperform resting state fMRI measurements. *Front. Syst. Neurosci.* 6:69. doi: 10.3389/fnsys.2012.00069

Chen, C. P., Keown, C. L., Jahedi, A., Nair, A., Pflieger, M. E., Bailey, B. A., et al. (2015). Diagnostic classification of intrinsic functional connectivity highlights somatosensory, default, mode, and visual regions in autism. *Neuroimage Clin.* 8, 238–245. doi: 10.1016/j.nicl.2015.04.002

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1023/A:1022627411411

Craddock, C., Benhajali, Y., Chu, C., Chouinard, F., Evans, A., Jakab, A. S., et al. (2013). The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. *Front. Neuroinform.* 7:41. doi: 10.3389/conf.fninf.2013.09.00041

Cruz, J. A., and Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* 2, 59–77. doi: 10.1177/117693510600200030

Cuingnet, R., Gerardi, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M. O., et al. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *Neuroimage* 56, 766–781. doi: 10.1016/j.neuroimage.2010.06.013

Dobson, A. J. (1990). *An Introduction to Generalized Linear Models.* New York, NY: Chapman and Hall.

Duda, M., Ma, R., Haber, N., and Wall, D. P. (2016). Use of machine learning for behavioral distinction of autism and ADHD. *Transl. Psychiatry* 6:e732. doi: 10.1038/tp.2015.221

Galliver, M., Gowling, E., Farr, W., Gain, A., and Male, I. (2017). Cost of assessing a child for possible autism spectrum disorder? An observational study of current practice in child development centres in the UK. *BMJ Paediatr. Open* 1:e000052. doi: 10.1136/bmjpo-2017-000052

Ghiassian, S., Greiner, R., Jin, P., and Brown, M. R. (2016). Using functional or structural magnetic resonance images and personal characteristic data to identify ADHD and autism. *PLoS One* 11:e0166934. doi: 10.1371/journal.pone.0166934

He, L., Li, H., Holland, S. K., Yuan, W., Altaye, M., and Parikh, N. A. (2018). Early prediction of cognitive deficits in very preterm infants using functional connectome data in an artificial neural network framework. *Neuroimage Clin.* 18, 290–297. doi: 10.1016/j.nicl.2018.01.032

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. (2018). Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *Neuroimage Clin.* 17, 16–23. doi: 10.1016/j.nicl.2017.08.017

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Li, H., Parikh, N. A., and He, L. (2018). A novel transfer learning approach to enhance deep neural network classification of brain functional connectomes. *Front. Neurosci.* 12:491. doi: 10.3389/fnins.2018.00491

Nielsen, J. A., Zielinski, B. A., Fletcher, P. T., Alexander, A. L., Lange, N., Bigler, E. D., et al. (2013). Multisite functional connectivity MRI classification of autism: ABIDE results. *Front. Hum. Neurosci.* 7:599. doi: 10.3389/fnhum.2013.00599

Plitt, M., Barnes, K. A., and Martin, A. (2015). Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards. *Neuroimage Clin.* 7, 359–366. doi: 10.1016/j.nicl.2014.12.013

Russell, S. J., and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach.* Essex: Prentice Education Limited.

Wetherby, A. M., and Prutting, C. A. (1984). Profiles of communicative and cognitive-social abilities in autistic children. *J. Speech Hear. Res.* 27, 364–377. doi: 10.1044/jshr.2703.364

Yahata, N., Morimoto, J., Hashimoto, R., Lisi, G., Shibata, K., Kawakubo, Y., et al. (2016). A small number of abnormal brain connections predicts adult autism spectrum disorder. *Nat. Commun.* 7:11254. doi: 10.1038/ncomms11254

Zhou, Z. (2012). *Ensemble Methods: Foundations and Algorithms.* Boca Raton, FL: Chapman and Hall/CRC.

**Conflict of Interest Statement**: 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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 29 March 2019
doi: 10.3389/fncom.2019.00015

# Electroencephalogram-Based Single-Trial Detection of Language Expectation Violations in Listening to Speech

*Hiroki Tanaka*, Hiroki Watanabe, Hayato Maki, Sakti Sakriani and Satoshi Nakamura*

*Division of Information Science, Nara Institute of Science and Technology, Nara, Japan*

We propose an approach for the detection of language expectation violations that occur in communication. We examined semantic and syntactic violations from electroencephalogram (EEG) when participants listened to spoken sentences. Previous studies have shown that such event-related potential (ERP) components as N400 and the late positivity (P600) are evoked in the auditory where semantic and syntactic anomalies occur. We used this knowledge to detect language expectation violation from single-trial EEGs by machine learning techniques. We recorded the brain activity of 18 participants while they listened to sentences that contained semantic and syntactic anomalies and identified the significant main effects of these anomalies in the ERP components. We also found that a multilayer perceptron achieved 59.5% (semantic) and 57.7% (syntactic) accuracies.

**Keywords: electroencephalogram, event-related potentials, N400, P600, single-trial analysis, multilayer perceptron**

OPEN ACCESS

***Edited by:***
*Dezhong Yao,*
*University of Electronic Science and Technology of China, China*

***Reviewed by:***
*Weiyi Ma,*
*University of Arkansas, United States*
*Toshihisa Tanaka,*
*Tokyo University of Agriculture and Technology, Japan*

****Correspondence:***
*Hiroki Tanaka*
*hiroki-tan@is.naist.jp*

***Received:*** *20 December 2018*
***Accepted:*** *01 March 2019*
***Published:*** *29 March 2019*

***Citation:***
*Tanaka H, Watanabe H, Maki H, Sakriani S and Nakamura S (2019) Electroencephalogram-Based Single-Trial Detection of Language Expectation Violations in Listening to Speech. Front. Comput. Neurosci. 13:15. doi: 10.3389/fncom.2019.00015*

## INTRODUCTION

In speech communication, we often face several types of language expectation violations, such as prosodic, semantic, and syntactic errors, especially in conversation through machine output (e.g., human–computer interaction; Koponen, 2010). Questionnaire-based subjective judgments are commonly used to rate such language expectation violations as linguistic discrepancies (Dybkjær et al., 2007). For example, regarding errors in the responses of spoken dialogue systems and machine translation, human examiners in previous research judged each sentence on an error scale from 1 to 5, unlike automatic evaluation metrics, e.g., word error rate (Lippmann, 1997; Och et al., 1999; Papineni et al., 2002). Even though this approach is quick and practical, it suffers from several problems. For instance, such subjective evaluations of participants contain ambiguity and cannot guarantee accurate answers. In this paper, we propose a new objective approach that automatically detects such language expectation violations from physiological signals (Näätänen et al., 2004; Morikawa et al., 2011; Honda et al., 2018) because participants face more obstacles when they are manipulating physiological signals. Although our goal is to develop an online detection tool of the language expectation violations of humans using physiological signals, we simplify the problem by detecting clear language expectation violations as our first step. We assume that this system can also be used for assessing people who exhibit the anomalies of semantic context sensitivity (e.g., autism spectrum, dementia, Olichney et al., 2008; Pijnacker et al., 2010; O'Connor, 2012; Tanaka et al., 2012, 2015, 2017a,b, 2018a; Ujiro et al., 2018).

An electroencephalogram (EEG) is a non-invasive tool that records the electrical activity of the human brain with electrodes placed on the scalp. Regarding real applications using EEGs, in

the context of motor imagery, which is reflected in event-related desynchronization [ERD; (Yeom and Sim, 2008)], the automatic detection of mental states based on convolutional neural networks (CNNs) has been proposed (Tang et al., 2017).

Unlike ERD, an event-related potential (ERP) is a measured time-locked brain response that is a direct result of a specific sensory, cognitive, or motor event. Since ERPs generally have a low signal/noise ratio in individual trials, many consecutive trials (e.g., 30 times) are usually averaged to diminish the random noise. Thus, single-trial detection of ERP components is very challenging due to their low signal/noise ratios (Blankertz et al., 2008; Lotte, 2015; Magee and Givigi, 2015). One public dataset focused on the single-trial detection of P300 components (Hald et al., 2006; Daubigney and Pietquin, 2011), which were elicited with relatively high signal/noise ratios. Most previous works have shown that P300 components can be detected with around 50–70% accuracy (exceeding the chance rate) using several machine learning algorithms (Stewart et al., 2014; Akram et al., 2015; Higashi et al., 2015; Sharma, 2017). Several approaches reached 100% accuracy using four to eight averaged trials in the BCI Competition 2003 (Cashero, 2012). We also need to consider that most works created subject-dependent models (within-subjects) because EEG signals are prone to being subject-dependent, and it remains challenging to generalize to subject-independent models (Terasawa et al., 2017).

Even though P300-based single-trial detection is one successful real application (P300-speller), it failed to detect language expectation violations including semantic and syntactic errors. To achieve single-trial detection of such errors, we focus on other ERP components, e.g., N400 and P600. N400 is a well-known ERP component that is evoked in auditory and visual modalities where semantic anomalies occur (Hagoort and Brown, 2000b). N400 is a phenomenon in which the potential shift in the negative direction increases around the brain's parietal region at around 400 ms from the onset of semantic and syntactic anomalies. Because N400 is strongly influenced by background noise, artifacts, and variations among trials, multiple times must be averaged. One study concluded that N400 is further influenced by a mismatch of the syntactic case information (Frisch and Schlesewsky, 2001). P600 (Narumi, 2014), another well-known ERP component (Hagoort and Brown, 2000a), is evoked in auditory and visual modalities where rule-governed anomalies generally occur. P600 is a language-related ERP that is thought to be elicited by grammatical errors and other syntactic anomalies. Several works have been done in Japanese (Ueno and Kluender, 2003; Mueller et al., 2007). P600 is characterized as a positive-going deflection with an onset around 500 ms after the onset of several types of anomalies. It peaks around 600 ms after the presentation of the stimulus and lasts several 100 ms. P600 is not language-specific, but it can be elicited in non-linguistic (but rule-governed) sequences [e.g., musical chords; (Patel et al., 1998)]. There are few P600 studies on Japanese syntactic violations in auditory modality (e.g., Mueller et al., 2005). To the best of our knowledge, no studies have addressed semantic violations in auditory modality in the Japanese language, which resemble our goal.

Based on our survey, despite the importance of real speech communication, only one study investigated the single-trial detection of semantic anomalies. Geuze et al. (2013) addressed the single-trial detection of semantic priming and the classification of visually presented related and unrelated words with an $L_2$ regularized logistic regression algorithm as a classifier. For more practical applications with such technology, the work-detection keyboard autocorrection of possible semantic and syntactic errors from only EEGs identified the accuracy of the single-trial error detection of around 70% (Putze and Stuerzlinger, 2017). They used linear discriminant analysis as a classifier. Although these two studies detected semantic anomalies in single-trial levels, they did not detect them in spoken sentences.

In this paper, we propose the single-trial detection (from subjects who listened to spoken sentences) of semantic and syntactic anomalies that can be applied to Japanese spoken communication error evaluations. Such linguistic errors might be common across languages. Although we evaluated language expectation violations in Japanese, our approaches may be generalizable to other languages that include semantic (reflecting context expectation) and syntactic (reflecting rule-governed) errors. Understandably, when languages differ, the onset (starting points) of the time-locked ERPs will also be different.

This paper examined the following three research questions:

1. Do semantic violations while listening to spoken Japanese sentences elicit ERPs?
2. How does machine learning contribute to single-trial detection for language expectation violations, including semantic and syntactic errors?
3. Which classification model more proficiently distinguishes semantic and syntactic violations?

We recorded EEG data while Japanese participants listened to sentences that contained semantic and syntactic anomalies and analyzed the ERP effects. We also detected both anomalies from single-trial EEGs with a technique that classified them from multielectrodes and by integrating the time and spectral information with multiple machine learning algorithms.

This paper is an extension of conference proceedings (Tanaka et al., 2018b) in which we reported the overall single-trial detection of semantically incorrect sentences. We added the analysis of syntactic anomalies as well as participant-independent models with more participants.

## METHODS

Our first aim is to confirm whether not only syntactic but also semantic violations in listening to Japanese sentences elicit ERPs. We hypothesized that semantic violations will elicit N400-/P600-related ERP components and syntactic violations will elicit P600-related ERP components. We also attempted to detect such violations from single-trial EEGs. We proposed several machine learning classifiers and confirmed classification above chance levels. In this section, we explain how we performed the EEG experiment and the classification.

## Participants

This study was carried out in accordance with the recommendations of the research ethical committee of the Nara Institute of Science and Technology. The protocol was approved by the research ethical committee of the Nara Institute of Science and Technology. All participants gave written informed consent in accordance with the Declaration of Helsinki.

Nineteen graduate students (16 males and 3 females) between 22 and 41 years of age (mean: 24.2) from the Nara Institute of Science and Technology participated. All were native Japanese speakers with no history of psychiatric problems or hearing disabilities; 18 were right-handed.

## Materials

In this study, we prepared two types of violations to elicit language expectation violations: a selectional restriction (as a semantic condition) and a double-nominative case (as a syntactic condition). Semantic violations very often also elicit biphasic N400 and P600 patterns, particularly when judging linguistic deviancy tasks (Sassenhagen et al., 2014). Note also that the double-nominative case violation that we chose for our syntactic manipulation has elicited N400 effects, including in Japanese (Mueller et al., 2005).

Japanese semantic and syntactic anomalies were manually created by referring to Takazawa et al. (2002) and Mueller et al. (2007). For the semantic condition, we defined error as a selectional restriction between a verb and its arguments. For the syntactic condition, error was defined a double-nominative case of the second phrase. We created an identical number of semantically and syntactically correct and incorrect sentences. We separated these sentences, which means that no two parts of the violated sentences are found in the stimuli.

The following is an example of two matched types of sentences (available on the **Supplementary Material**):

(Semantic)

a. Hanako-ga nikki-o tsuzu-ta
Hanako-NOM a diary-DAT write-PAST
Hanako wrote in her diary .

b. *Hanako-ga beer-o tsuzu-ta
Hanako-NOM a beer-DAT write-PAST
Hanako wrote a beer.

NOM: nominative case marker;
DAT: dative case marker;
PAST: past tense morpheme.

(Syntactic)

c. Gakusei-ga kenchikuka-o tasuke-ta
Student-NOM architect-DAT help-PAST
The student helped the architect.

d. *Gakusei-ga kenchikuka-ga tasuke-ta
Taro-NOM architect-NOM help-PAST

NOM: nominative case marker;
DAT: dative case marker;
PAST: past tense morpheme.

Here, an asterisk indicates semantically (b) and syntactically (d) incorrect sentences. Matched sentences corresponded in the first and third phrases. Due to the speech stimulus, we controlled the phonemes following Hagoort and Brown (2000b) in the third phrase to begin with plosive sounds:/t/,/k/,/d/, and/g/. Since such plosive sounds are in the onset position of the ERPs marked by human annotators, a consistent pattern is required in the spectrogram.

A group composed of the first author (A), the second author (B), and a graduate student who did not join our experiment (C) confirmed and corrected each sentence and reached a consensus about whether a semantic anomaly occurred. We selected the following 200 types of sentence from a total of 360 sentences: 40 semantically correct, 40 semantically incorrect, 40 syntactically correct, 40 syntactically incorrect, and 40 fillers sentences. Fillers were correct sentences that were used as dummies.

We transcribed them into text and recorded speech that was naturally spoken by a professional female narrator whom we instructed to avoid inserting pauses between phrases. The length of the audio files ranged from 1.8 to 3.0 s.

For the semantic case, the syntactic structure of the sentences was matched between the two conditions. We used the same target words in the third phrases. The experiment member A confirmed that the mean frequency of the third phrases was 1.02 in both conditions. Here, a mora is a unit in phonology that determines the syllable weight. The mean number of the moras of the third phrases was 4.25 ($SD = 1.35$). The difference of the two conditions was the second phrases with a mean number of moras of 4.15 ($SD = 0.86$) in the correct condition and 4.63 (SD = 0.93) in the incorrect condition.

For the syntactic case, the difference of the two conditions was the nominative case marker of the second phrases. The mean frequency of the second phrase was 1 in both conditions. The mean number of moras in the second phrases was 4.1 (SD = 0.98) in both conditions.

Moreover, we investigated the predictability of subsequent words (cloze probability) that affect the N400 amplitudes (Borovsky et al., 2010). One hundred crowdsourcing workers were given a list of 40 semantically incorrect sentences from which the final word had been removed. They read the sentences and filled in the blanks at the position of the hidden sentence-final words with the first word that popped into their heads. After that, we manually changed the present tense to the past tense, revised minor typing mistakes, and calculated the cloze probability of the most frequently selected words. The following is the distribution of the cloze probability: mean, 41%, SD, 16%, range, 14–85%. We confirmed that no words appeared as semantically incorrect in our stimuli, which means the cloze probability to the word is zero.

## Synchronization

Since ERPs are the time-locked brain response, we explain details with regard to synchronization between the auditory stimuli and EEG. Experiment members A and C marked the synchronized onset ($t = 0$). For the semantic case, ERP onset is the speech's start position of the third phrases. The onset starts with plosive sounds. The precise beginning position was marked by observing spectrogram of the speech. For the syntactic case, ERP onset

is the speech's start position of the nominative case marker of the second phrases. The onset also begins with plosive sounds (only/g/) and was marked by observing spectrogram of the speech. We used the Wavesurfer (TMH, Speech, Music, and Hearing) in order to visualize spectrogram of the speech.

## Design

The participants entered a soundproof room, sat down, and were instructed to look at the attention point on the monitor and to refrain from blinking and moving as much as possible. The following was the experimental procedure: (1) watch the "+" mark for 1 s on the screen; (2) listen to one randomly presented speech sound for 4 s; and (3) press a key and determine within 2 s whether each speech contains grammatical or semantic errors. We conducted subjective evaluations and prepared practice trials before the EEG recordings. All these steps were completed within 25 min. For speech listening, we used earphones (ER1). This series of experiments was created using presentation software provided by Neurobehavioral Systems (Version 18.0, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com).

The correct answer rates from the behavioral results were 95.8% for semantically correct and incorrect and 96.7% for syntactically correct and incorrect (error rate is <5%).

## Electroencephalogram Recording and Preprocessing

As an EEG cap, we used ActiCAP by Brain Products with 32 ch active electrodes according to all the standard positions of the international 10/20 system (see **Figure 1**). We used a BrainAmp DC from the same company as an amplifier. As a recording filter, we applied a high-pass filter of 0.016 Hz and a low-pass filter of 250 Hz. The sampling rate was 1,000 Hz, the reference electrode was FCz, and the ground electrode was FPz. In order to synchronize the speech signal with EEG, we generated a speech timing signal and recorded it with the EEG amplifier.

For preprocessing the recorded EEGs, we used FieldTrip software (Oostenveld et al., 2011) as follows: (1) Re-referencing was performed on the average of the TP9 and TP10 electrodes. (2) An FIRfilter was applied through a high-pass filter of 0.3 Hz (order: 6192), which is designed for DC suppression (−60 dB at DC) to replace the baseline correction (Maess et al., 2006; Wolff et al., 2008). (3) For each trial condition (excluding fillers), epochs were extracted at −100 to 900 ms of the synchronous onset. Here, the onset is the speech's start positions of the third phrases for the semantic condition and of the nominative case marker of the second phrases for the syntactic condition. (4) First artifact rejection was performed on epochs that exceeded a threshold of −350 and 350 μV in order to remove epochs contaminated with large amplitude of artifacts. This threshold rejection did not consider FP1 and FP2 electrodes where eye-related artifacts mainly contaminated. This large amplitude threshold is to preserve eye-blink artifact, which will be removed by later independent component analysis (ICA). (5) We performed an automatic approach and visual inspection to remove muscle artifacts: automatically identifying artifacts at $Z$ score = 15 by considering amplitude distributions of band-pass-filtered epoch data (110–140 Hz), then rejecting epochs contaminated with muscle artifacts based on visual inspection (Meyer et al., 2017). (6) The recorded EEGs were downsampled to 250 Hz. (7) The logistic infomax ICA algorithm of Bell and Sejnowski (1995) was performed to correct eye-related artifacts, and eye-related components were removed. We identified the components by calculating the correlations to the FP1 and FP2 electrodes and by a visual inspection of the topographies and the waveforms. Four was the maximum number of rejected components because we only intended to remove as few horizontal and vertical ocular artifacts as possible. The rejected components had a mean of 2.1 (SD: 1.2). (8) A second artifact rejection was performed on epochs that exceeded the thresholds of −120 and 120 μV. As a result of the above artifact rejection procedures, one participant was removed because of the large number of rejected epochs (more than 30% of the epochs were rejected). The average rate of rejected trials across participants was 6.2%. We found no effects of the number of rejected trials between the semantically correct and incorrect and the syntactically correct and incorrect by using paired $t$-test {semantic: $[t_{(17)} = 1.32, p = 0.20]$, syntactic: $[t_{(17)} = 0.68, p = 0.51]$}.

## Event-Related Potential Analysis

For further improvement of the signal/noise ratio, we applied another filtering procedure to the ERP data. Since the N400 components are around 6 Hz and the activity in the alpha frequency band tends to contaminate the EEG data, we used a two-pass IIR Butterworth filter of order 8 at 8 Hz to achieve a steeper frequency response than the FIR filter and to preserve the ERP components that also attenuate the alpha activity. Note that this filter was applied for only visualizing and analyzing ERPs, meaning that we did not use these filtered signals to the single-trial analysis.

We computed the grand average of all the participants. Based on a previous studies (Hagoort and Brown, 2000a,b; Mueller et al., 2005; Wolff et al., 2008), we selected the following electrodes in each time window: 100–300, 300–500, and 500–800 ms. These time windows were selected based on the previous study that analyzed syntax- and semantic-related ERP effects (Mueller et al., 2005). To assess the topographic differences in the ERPs, electrodes were summed up in five regions of interest (ROIs)—left anterior: F3, F7, FC1, FC5; right anterior: F4, F8, FC2, FC6; left posterior: CP1, CP5, P3, P7; right posterior: CP2, CP6, P4, P8; and midline: Fz, FCz, Cz, Pz. For the statistical analyses, we calculated the mean amplitudes in the chosen time windows (Wolff et al., 2008).

We used two-way repeated ANOVAs to examine the main effects of the condition and its interaction by ROIs in each time window. We performed a *post hoc* multiple comparison of the interaction between conditions and regions using the Tukey–Kramer method. Finally, we performed cluster-based permutation tests (Maris and Oostenveld, 2007) on the ERPs of the semantic and the syntactic conditions. Regarding the cluster-based permutation tests, for each time step of interest, we marked the electrodes that are members of significant clusters. The significance probability can be calculated by means of the Monte Carlo method. The Monte Carlo significance probability

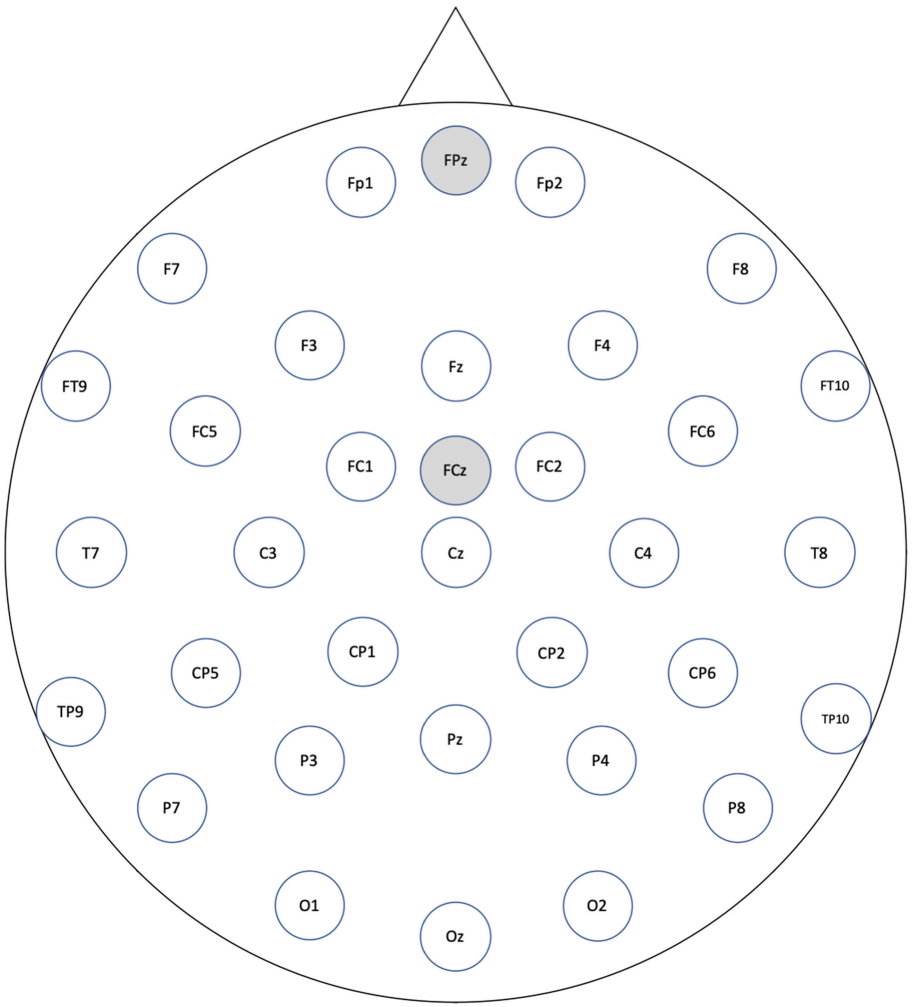


**FIGURE 1 |** All electrode labels: gray electrodes indicate reference and grand position.

is also called a *p*-value. If the *p*-value is smaller than the critical alpha level (5% in this study), then we conclude that the data in the two experimental conditions are significantly different. Overall, we set the significance level to 5%.

## Feature and Classifiers

Based on previous work (Hagoort and Brown, 2000b; Roehm et al., 2004), we extracted the average values of the 100–300, 300–500, and 500–800 ms amplitudes from all of the electrodes (93 time domain features). To avoid overfitting to the training data, we selected specific time domains (possibly important time ranges) rather than using all time sampling points (simplifying the model). We also considered all of the electrodes with frequency domains for the single-trial detection of EEGs (Putze and Stuerzlinger, 2017). The delta band has been associated with N400 and P600 components in language (Correia et al., 2015). Thus, we performed a fast Fourier transform on the waveform between 0 and 900 ms to the onset and calculated the average values of the power spectra of δ (1–3 Hz), θ (4–7 Hz), α (8–12 Hz), and β (13–28 Hz) (124 spectral domain features) by referring to previous work (Hald et al., 2006; Mcmahon et al., 2015). We concatenated time and spectral features (217 dimensions). The feature vectors were normalized to a mean of zero and one standard deviation.

For the classifiers, we used a linear kernel support vector machine (L-SVM), a radial kernel support vector machine (R-SVM), a random forest (RF), and multilayer perceptrons (MLPs). The classifiers were trained on a dataset that combined 13 participants and subsequently tested on five different participants without further training by following Vareka and Mautner (2017). We observed how our detection models performed when they dealt with data from previously unknown participants.

These models were trained using 5-fold cross-validation for hyperparameter tuning on the training set to optimize the accuracies. The hyperparameters included the kernel (linear or radial basis function), $C = \{10^{-5}, 10^{-4}, \ldots, 10^{3}\}$, $\gamma = \{0.00, 0.005, \ldots, 1.00\}$ (in the case of the RBFkernel) for the SVMs, the number of variables tried at each split = {5, 10, 15, 20} for the

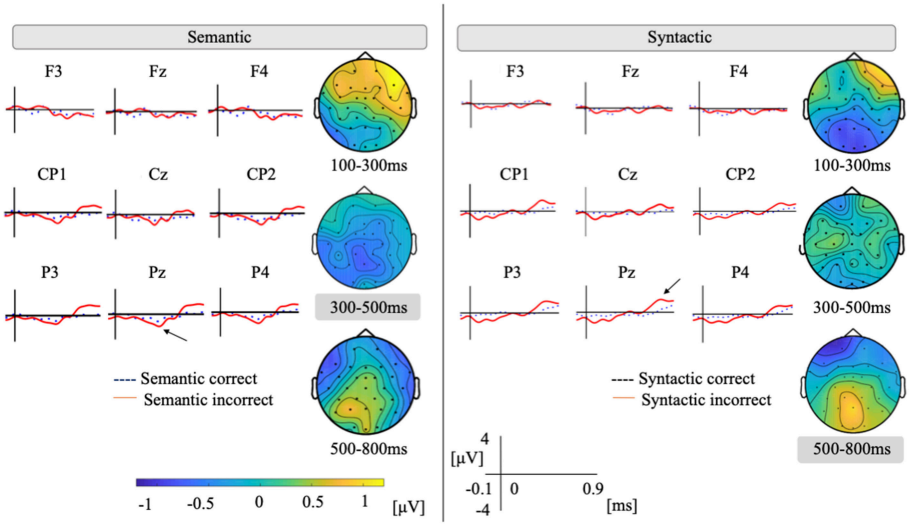


**FIGURE 2 |** Grand average of nine representative electrodes of semantic and syntactic conditions. Vertical axis ranges between −4 and +4. Positivity is plotted up. We plotted −100 to 900 ms to the stimuli onset. Topographical map shows a difference wave between incorrect and correct conditions.

RF, and the number of hidden units {5, 10, 50, 100, 150, 200}, the number of hidden layers {1, 2, 3}, and activation function (logistic, hyperbolic tangent, or rectified linear unit) in the MLP by referring to Vail et al. (2018). After the parameters were found, the models were trained on the whole training dataset and subsequently tested.

By a binomial test, we compared the chance rate (50.4% for the semantic sentences and 50.4% for the syntactic sentences in the test set) and the model that achieved the highest accuracy as well as precision, recall, and F1. We also calculated the correlation between cloze probability and semantic accuracy based on Pearson's correlation coefficient.

# RESULTS

## Event-Related Potential Effects

**Figure 2** plots the ground averages at representative electrodes in the semantic and syntactic conditions. For the semantic condition, a potential shift to the negative around 400 ms can be observed under the semantically incorrect condition over the parietal region, and late positivity (P600) can also be seen.

Based on our assumption, for a time window of 300–500 ms, ANOVAs would show the main effects of the condition [$F_{(1,17)} = 4.69$, $p = 0.04$]. No significant interaction was shown between condition by region [$F_{(4,68)} = 1.18$, $p = 0.32$]. Regarding other time windows, for a mean amplitude of 100–300 ms, we found main effects of condition [$F_{(1,17)} = 4.51$, $p = 0.04$] and also a significant interaction of condition by region [$F_{(4,68)} = 11.5$, $p < 0.001$]. Since there were significant interactions of the condition by region, multiple comparisons were separately calculated for each region. *Post hoc* analysis by the Tukey–Kramer method revealed that the left anterior [difference (incorrect – correct): 0.66, $p = 0.02$, 95% CI = 0.09–1.23] and the right posterior (difference: 0.45, $p = 0.02$, 95% CI = 0.06–0.83) were significantly different between two conditions. For the mean amplitude of 500–800 ms, we found no main effects of condition [$F_{(1,17)} = 0.82$, $p = 0.37$]. However, we did identify a significant interaction of condition by region [$F_{(4,68)} = 5.39$, $p < 0.001$]. *Post hoc* analysis revealed that the left posterior (difference: 0.54, $p = 0.008$, 95% CI = 0.003–1.01) and the right anterior (difference: 0.88, $p < 0.001$, 95% CI = 0.51–1.2) were significantly different between two conditions.

For the syntactic condition, we observed a potential shift to the positive after 500 ms under the syntactically incorrect condition over the parietal region. Based on our assumption, for the time window of 500–800 ms, ANOVAs showed no main effects of condition [$F_{(1,17)} = 1.00$, $p = 0.33$]. ANOVAs showed the interaction of the condition by region [$F_{(4,68)} = 6.03$, $p < 0.001$]. *Post hoc* analysis revealed that the left posterior (difference: 0.51 μV, $p = 0.04$, 95% CI = 0.003–1.01 μV) and the right posterior (difference: 0.45 μV, $p = 0.02$, 95% CI = 0.06–0.83 μV) were significantly different between two conditions. Regarding other time windows, for the mean amplitude of 100–300 ms, we found no main effects of condition [$F_{(1,17)} = 1.28$, $p = 0.27$]. However, we did find a significant interaction of the condition by region [$F_{(4,68)} = 6.86$, $p < 0.001$]. *Post hoc* analysis revealed that the left posterior (difference: −0.65 μV, $p = 0.006$, 95% CI = −1.08 to −0.21 μV) and the right anterior (difference: −0.49 μV, $p = 0.02$, 95% CI = −0.90 to −0.08 μV) were significantly different between two conditions. For the mean amplitude of 300–500 ms, there were no main effects of condition [$F_{(1,17)} = 0.05$, $p = 0.82$] and no interaction of the condition by region [$F_{(4,68)} = 0.05$, $p = 0.79$].

**Figures 3**, **4** show the results of cluster-based permutation tests on ERPs of the semantic and the syntactic conditions.

## Single-Trial Detection

**Table 1** indicates the accuracy of each classifier in the test sets. For the semantic conditions, MLP achieved the highest accuracy of 59.5%. Regarding this accuracy,

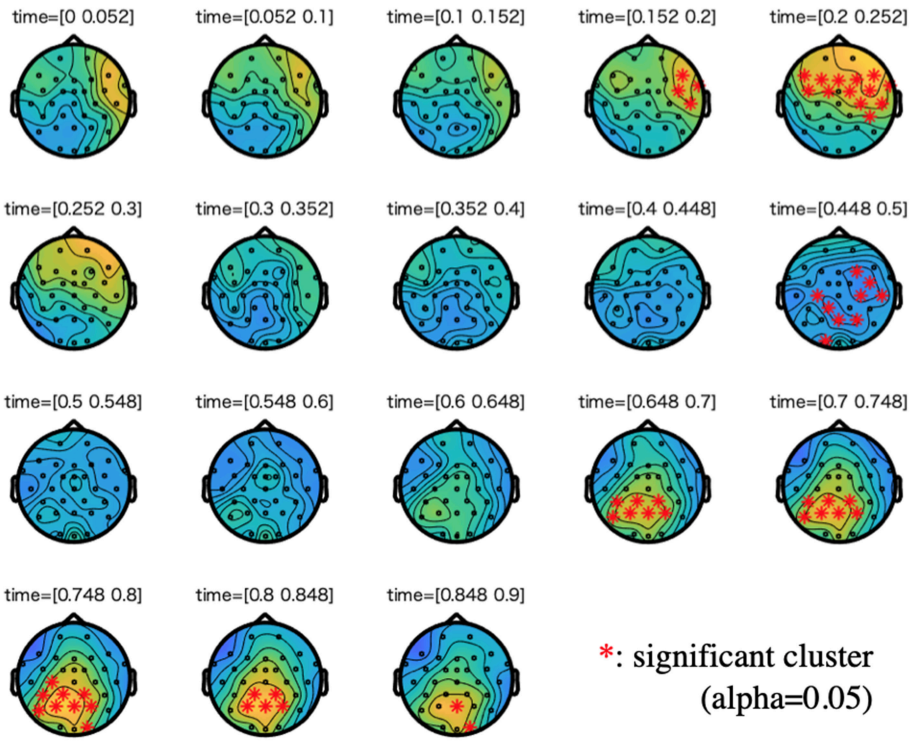


**FIGURE 3 |** Cluster-based permutation tests on the event-related potentials (ERPs) of the semantic condition along with a difference wave between incorrect and correct conditions. We plotted 0–900 ms to the stimuli onset. For each time step of interest (time range: 0.05), we highlighted the electrodes that are members of significant clusters (cluster alpha value: 0.05). A cluster is significant if its $p$-value is less than the critical alpha level.

we confirmed a statistical significance compared to the chance rate ($p < 0.05$): 44.3% precision, 63.1% recall, and 52.1% F1.

We found no significant correlation between the cloze probability or the predicted accuracy in the semantic condition (all classifiers, $r < 0.15$, $p > 0.05$).

For the syntactic conditions, the highest accuracy was also found when using MLP (57.7%), and we confirmed a statistical significance compared to the chance rate ($p < 0.05$): 58.8% precision, 57.9% recall, and 58.4% F1.

## DISCUSSION

The aim of the present study is to observe the time-locked effects of semantic and syntactic anomalies in spoken Japanese sentences and to detect them with single-trial EEGs. We achieved this by focusing on the previous approach: ERPs. We followed two previous studies that elicited the ERP components of N400 and P600 in Japanese: Mueller et al. (2007) and Takazawa et al. (2002). We hypothesized that semantic violations will elicit N400-/P600-related ERP components and syntactic violations will elicit P600-related ERP components. We also attempted to use SVMs, RF, and MLP for single-trial EEGs and confirmed classification that exceeded chance levels. We next summarize our discussion regarding ERP analysis and single-trial detection.

### Event-Related Potential Analysis

For the semantic condition, we used such previously proposed stimuli as selectional restriction (Takazawa et al., 2002). Although the previous study was performed with visual stimuli, our experiment confirmed that ERP components were elicited even in an auditory experimental design.

One of our experiment's drawbacks is that semantically incorrect sentences were limited to the anomalies of the selectional restrictions at the end of sentences. Our 40-filler setting is limited to natural settings, and naturalistic sentence processing is a major analysis challenge. We identified several participants who did not indicate the strong effects of ERPs. We need to control such related factors as social traits and the attention of the participants as well as age (Constantino and Gruber, 2012).

Onset is another critical aspect for analyzing ERPs. We set the ERP onset to the speech's start position of the third phrases for the semantic condition and the speech's start position of the nominative case marker of the second phrases for the syntactic condition. Because this study uses auditory stimuli (speech sequences), we did not know the actual timing when the participants perceived the violations. In the future, we will measure the effects in the onset latency of a representative range of ERPs and implement artificial time shifting (Kiesel et al., 2008; Zoumpoulaki et al., 2013; Sassenhagen et al., 2014).

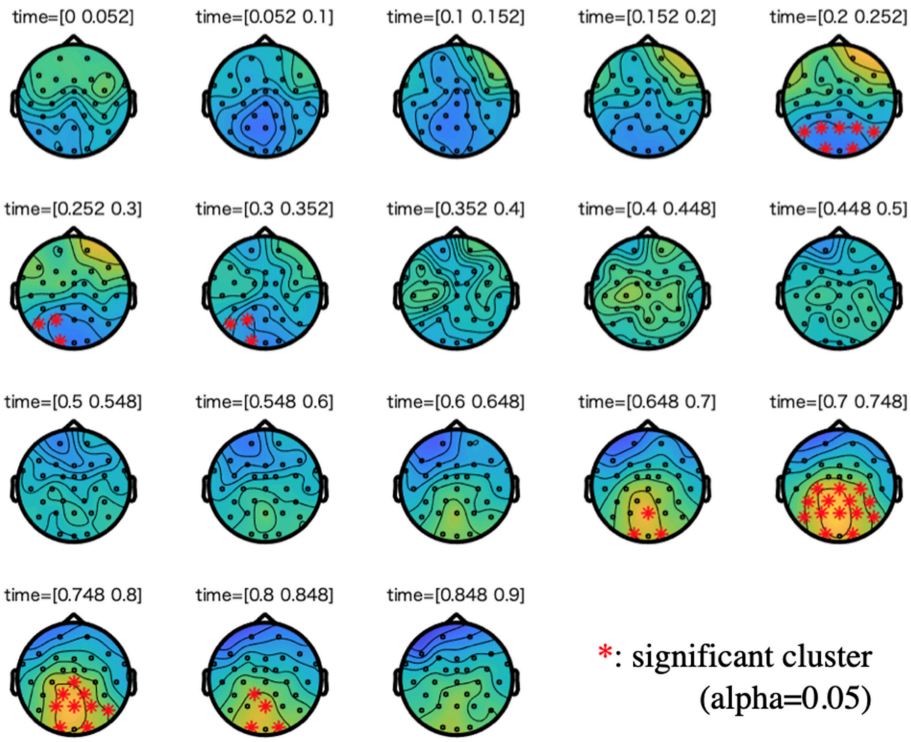


**FIGURE 4 |** Cluster-based permutation tests on the ERPs of the syntactic condition along with a difference wave between incorrect and correct conditions. We plotted 0 to 900 ms to the stimuli onset. For each time step of interest (time range: 0.05), we highlighted the electrodes that are members of significant clusters (cluster alpha value: 0.05). A cluster is significant if its *p*-value is less than the critical alpha level.

**TABLE 1 |** Unweighted accuracies (%) of classifiers.

| Violations | L-SVM | R-SVM | RF | MLP |
|---|---|---|---|---|
| Semantic | 58.2 | 56.0 | 58.2 | **59.5** |
| Syntactic | 54.7 | 54.7 | 55.3 | **57.7** |

*The best model is indicated in bold.*

## Single-Trial Detection

Our classification model achieved 59.5% (semantic) and 57.7% (syntactic) detection accuracies in the incorrect conditions and outperformed the chance rate. MLP outperformed the other classifiers: SVMs and RF. Such accuracies were similar or superior to previous related works (Geuze et al., 2013; Higashi et al., 2015; Putze and Stuerzlinger, 2017). The previous work that detected semantic priming with 12 subjects showed accuracy between 51 and 63%, which is above chance in a cross-subject study (Geuze et al., 2013). Although our evaluation was validated by previously unseen participants, the MLP achieved a similar accuracy.

The N400 amplitude for incongruent words was also modulated by the cloze probability of the expected congruent word for that place. Generally, the best predictor of a word's N400 amplitude in a given sentence is its cloze probability (Kutas and Hillyard, 1984). The N400 amplitude is largest for items with low cloze probability and smallest for items with high cloze probability. Semantic anomaly thus shows the end point on a continuum of expectedness in a particular context (Coulson, 2001). Thus, we hypothesized that detecting low cloze probability items (large N400 amplitude) is easier because of the relatively high signal/noise ratios (Hald et al., 2006; Daubigney and Pietquin, 2011). However, we did not find a relationship between accuracy and cloze probability. This is because we did not control the cloze probability of the semantic incorrect sentences or the semantic correct sentences prior to the experiment (Borovsky et al., 2010).

This study did not consider the effects of the individuality of the frequency band. We fixed the frequency bands rather than individually adapting them based on individual alpha frequencies. This idea needs to be considered due to the high individual variability in this domain (Klimesch, 2012).

To improve classification accuracy, we need to increase the sophistication of the machine learning models, although EEGs have a low signal/noise ratio. We believe that a participant-adaptive technique (e.g., maximum likelihood linear regression; Gales and Woodland, 1996; Pan and Yang, 2010) is one possible future direction. Due to a large amount of P300 data, such as for a BCI competition, we applied several types of machine learning approach to our collected data by transfer learning (Pan et al., 2016).

Another possible direction to improve the classification accuracy is to average several trials (not a single trial) whose

usefulness has already been validated. Several approaches achieved 100% accuracy using only four to eight averaged trials on P300 data (Cashero, 2012). We can apply this approach to detect the language expectation violations toward practical usage.

We will also improve our model using graph regularized tensor factorization (Maki et al., 2018) as well as non-negative matrix factorization, which we previously proposed. Automatic onset detection and the techniques of artificial shifted trials are also needed for completely automated anomaly detection (Kutas and Hillyard, 1980).

## CONCLUSIONS

This study aims to detect semantic and syntactic anomalies from a one-shot EEG, using a machine learning technique. We measured the EEGs of 18 participants while they listened to semantically anomalous sentences and confirmed N400- and P600-related ERP components. When using MLP, we achieved detection accuracies of 59.5% (semantic) and 57.7% (syntactic) with time and spectral domain inputs. From here, the results suggest that machine learning might be able to detect semantic and syntactic anomalies from correct sentences.

## DATA AVAILABILITY

The datasets for this study will not be made publicly available because of the Act on the Protection of Personal Information.

## AUTHOR CONTRIBUTIONS

HT, HW, and HM performed the experiments and data analysis and conceived the methodology and the machine learning algorithms. HT and HW performed EEG preprocessing. SS and SN conceived the entire experiment design and analyzed, and discussed the results. HT wrote this manuscript. All of the authors reviewed the manuscript.

## FUNDING

Part of this work was supported by JSPS KAKENHI Grant Numbers JP17H06101, JP18K11437.

## ACKNOWLEDGMENTS

We thank Rui Hiraoka of the Nara Institute of Science and Technology for creating the stimuli and discussing the study design in our work's early stage. We also thank Yu Odagaki for his helpful suggestions.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2019.00015/full#supplementary-material

## REFERENCES

Akram, F., Han, S. M., and Kim, T.-S. (2015). An efficient word typing P300-BCI system using a modified T9 interface and random forest classifier. *Comput. Biol. Med.* 56, 30–36. doi: 10.1016/j.compbiomed.2014.10.021

Bell, A. J., and Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Comput.* 7, 1129–1159. doi: 10.1162/neco.1995.7.6.1129

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., and Muller, K.-R. (2008). Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Process. Mag.* 25, 41–56. doi: 10.1109/MSP.2008.4408441

Borovsky, A., Kutas, M., and Elman, J. (2010). Learning to use words: event-related potentials index single-shot contextual word learning. *Cognition* 116, 289–296. doi: 10.1016/j.cognition.2010.05.004

Cashero, Z. (2012). *Comparison of EEG Preprocessing Methods to Improve the Performance of the P300 Speller*. Fort Collins, CO: Proquest, Umi Dissertation Publishing.

Constantino, J. N., and Gruber, C. P. (2012). *Social Responsiveness Scale (SRS)*. Torrance, CA: Western Psychological Services.

Correia, J. M., Jansma, B., Hausfeld, L., Kikkert, S., and Bonte, M. (2015). EEG decoding of spoken words in bilingual listeners: from words to language invariant semantic-conceptual representations. *Front. Psychol.* 6:71. doi: 10.3389/fpsyg.2015.00071

Coulson, S. (2001). *Semantic Leaps: Frame-Shifting and Conceptual Blending in Meaning Construction*. New York, NY: Cambridge University Press. doi: 10.1017/CBO9780511551352

Daubigney, L., and Pietquin, O. (2011). "Single-trial P300 detection with Kalman filtering and SVMs," in *ESANN* 2011 (Bruges), 399–404.

Dybkjær, L., Hemsen, H., and Minker, W. (2007). *Evaluation of Text and Speech Systems, Vol. 38*. Dordrecht: Springer Science and Business Media. doi: 10.1007/978-1-4020-5817-2

Frisch, S., and Schlesewsky, M. (2001). The N400 reflects problems of thematic hierarchizing. *Neuroreport* 12, 3391–3394. doi: 10.1097/00001756-200110290-00048

Gales, M. J., and Woodland, P. C. (1996). Mean and variance adaptation within the MLLR framework. *Comput. Speech Lang.* 10, 249–264. doi: 10.1006/csla.1996.0013

Geuze, J., van Gerven, M. A., Farquhar, J., and Desain, P. (2013). Detecting semantic priming at the single-trial level. *PLoS ONE* 8:60377. doi: 10.1371/journal.pone.0060377

Hagoort, P., and Brown, C. M. (2000a). ERP effects of listening to speech compared to reading: the P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation. *Neuropsychologia* 38, 1531–1549. doi: 10.1016/S0028-3932(00)00053-1

Hagoort, P., and Brown, C. M. (2000b). ERP effects of listening to speech: semantic ERP effects. *Neuropsychologia* 38, 1518–1530. doi: 10.1016/S0028-3932(00)00052-X

Hald, L., Bastiaansen, M., and Hagoort, P. (2006). EEG theta and gamma responses to semantic violations in online sentence processing. *Brain Lang.* 96, 90–105. doi: 10.1016/j.bandl.2005.06.007

Higashi, H., Rutkowski, T. M., Tanaka, T., and Tanaka, Y. (2015). "Subspace-constrained multilinear discriminant analysis for ERP-based brain computer interface classification," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (Hong Kong), 934–940. doi: 10.1109/APSIPA.2015.7415409

Honda, M., Tanaka, H., Sakriani, S., and Nakamura, S. (2018). "Detecting suppression of negative emotion by time series change of cerebral blood flow using fNIRS," in *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)* (Las Vegas, NV), 398–401. doi: 10.1109/BHI.2018.8333452

Kiesel, A., Miller, J., Jolicoeur, P., and Brisson, B. (2008). Measurement of ERP latency differences: a comparison of single-participant and

jackknife-based scoring methods. *Psychophysiology* 45, 250–274. doi: 10.1111/j.1469-8986.2007.00618.x

Klimesch, W. (2012). α-band oscillations, attention, and controlled access to stored information. *Trends Cognit. Sci.* 16, 606–617. doi: 10.1016/j.tics.2012.10.007

Koponen, M. (2010). "Assessing machine translation quality with error analysis," in *Electronic Proceeding of the KaTu Symposium on Translation and Interpreting Studies* (Helsinki).

Kutas, M., and Hillyard, S. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science* 207, 203–205. doi: 10.1126/science.7350657

Kutas, M., and Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature* 307, 161–163. doi: 10.1038/307161a0

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Commun.* 22, 1–15. doi: 10.1016/S0167-6393(97)00021-6

Lotte, F. (2015). Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces. *Proc. IEEE* 103, 871–890. doi: 10.1109/JPROC.2015.2404941

Maess, B., Herrmann, C. S., Hahne, A., Nakamura, A., and Friederici, A. D. (2006). Localizing the distributed language network responsible for the N400 measured by MEG during auditory sentence processing. *Brain Res.* 1096, 163–172. doi: 10.1016/j.brainres.2006.04.037

Magee, R., and Givigi, S. (2015). "A genetic algorithm for single-trial P300 detection with a low-cost EEG headset," in *9th Annual IEEE International Systems Conference (SysCon)* (Vancouver, BC), 230–234. doi: 10.1109/SYSCON.2015.7116757

Maki, H., Tanaka, H., Sakti, S., and Nakamura, S. (2018). "Graph regularized tensor factorization for single-trial EEG analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing* (Calgary, AB), 846–850. doi: 10.1109/ICASSP.2018.8461897

Maris, E., and Oostenveld, R. (2007). Non-parametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Mcmahon, T., Zijl, P. C. M. V., and Gilad, A. A. (2015). Gamma- and theta-band synchronization during semantic priming reflect local and long-range lexical-semantic networks. *Brain Lang.* 27, 320–331. doi: 10.1002/nbm.3066.Non-invasive

Meyer, L., Henry, M. J., Gaston, P., Schmuck, N., and Friederici, A. D. (2017). Linguistic bias modulates interpretation of speech *via* neural delta-band oscillations. *Cereb. Cortex* 27, 4293–4302. doi: 10.1093/cercor/bhw228

Morikawa, K., Kozuka, K., and Adachi, S. (2011). "Assessment of speech discrimination based on the event-related potentials to the visual stimuli," in *IEEE International Conference on Communications* (Kyoto), 1–5. doi: 10.1109/icc.2011.5962441

Mueller, J., Hahne, A., Fujii, Y., and Friederici, A. (2005). Native and non-native speakers processing of a miniature version of Japanese as revealed by ERPs. *J. Cognit. Neurosci.* 17, 1229–1244. doi: 10.1162/0898929055002463

Mueller, J. L., Hirotani, M., and Friederici, A. D. (2007). ERP evidence for different strategies in the processing of case markers in native speakers and non-native learners. *BMC Neurosci.* 8:18. doi: 10.1186/1471-2202-8-18

Näätänen, R., Pakarinen, S., Rinne, T., and Takegata, R. (2004). The mismatch negativity (MMN): towards the optimal paradigm. *Clin. Neurophysiol.* 115, 140–144. doi: 10.1016/j.clinph.2003.04.001

Narumi, T. (2014). *An Investigation of the Automaticity in Parsing for Japanese EFL Learners: Examining From Psycholinguistic and Neurophysiological Perspectives*. Ph.D. thesis, Kobe University.

Och, F. J., Tillmann, C., and Ney, H. (1999). "Improved alignment models for statistical machine translation," in *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora* (College Park, MD).

O'Connor, K. (2012). Auditory processing in autism spectrum disorder: a review. *Neurosci. Biobehav. Rev.* 36, 836–854. doi: 10.1016/j.neubiorev.2011.11.008

Olichney, J., Taylor, J., Gatherwright, J., Salmon, D., Bressler, A., Kutas, M., et al. (2008). Patients with MCI and N400 or P600 abnormalities are at very high risk for conversion to dementia. *Neurology* 70, 1763–1770. doi: 10.1212/01.wnl.0000281689.28759.ab

Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput. Intell. Neurosci.* 2011:156869. doi: 10.1155/2011/156869

Pan, S. J., and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Pan, W., Yang, Q., Duan, Y., and Ming, Z. (2016). Transfer learning for semisupervised collaborative recommendation. *ACM Trans. Interact. Intell. Syst.* 6:10. doi: 10.1145/2835497

Papineni, K., Roukos, S., Ward, T., and Jing Zhu, W. (2002). *BLEU: A Method for Automatic Evaluation of Machine Translation.*Philadelphia, PA: Association for Computational Linguistics (ACL), 311–318.

Patel, A. D., Gibson, E., Ratner, J., Besson, M., and Holcomb, P. J. (1998). Processing syntactic relations in language and music: an event-related potential study. *J. Cognit. Neurosci.* 10, 717–733. doi: 10.1162/089892998563121

Pijnacker, J., Geurts, B., Van Lambalgen, M., Buitelaar, J., and Hagoort, P. (2010). Exceptions and anomalies: an ERP study on context sensitivity in autism. *Neuropsychologia* 48, 2940–2951. doi: 10.1016/j.neuropsychologia.2010.06.003

Putze, F., and Stuerzlinger, W. (2017). "Automatic classification of auto-correction errors in predictive text entry based on EEG and context information," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (Glasgow, UK), 137–145. doi: 10.1145/3136755.3136784

Roehm, D., Schlesewsky, M., Bornkessel, I., Frisch, S., and Haider, H. (2004). Fractionating language comprehension *via* frequency characteristics of the human EEG. *Neuroreport* 15, 409–412. doi: 10.1097/00001756-200403010-00005

Sassenhagen, J., Schlesewsky, M., and Bornkessel-Schlesewsky, I. (2014). The P600-as-P3 hypothesis revisited: single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain Lang.* 137, 29–39. doi: 10.1016/j.bandl.2014.07.010

Sharma, N. (2017). *Single-Trial P300 Classification Using PCA With LDA, QDA and Neural Networks. arXiv [preprint] arXiv:1712.01977.*

Stewart, A. X., Nuthmann, A., and Sanguinetti, G. (2014). Single-trial classification of EEG in a visual object task using ICA and machine learning. *J. Neurosci. Methods* 228, 1–14. doi: 10.1016/j.jneumeth.2014.02.014

Takazawa, S., Takahashi, N., Nakagome, K., Kanno, O., Hagiwara, H., Nakajima, H., et al. (2002). Early components of event-related potentials related to semantic and syntactic processes in the Japanese language. *Brain Topogr.* 14, 169–177. doi: 10.1023/A:1014546707256

Tanaka, H., Adachi, H., Ukita, N., Ikeda, M., Kazui, H., Kudo, T., et al. (2017a). Detecting dementia through interactive computer avatars. *IEEE J. Transl. Eng. Health Med.* 5, 1–11. doi: 10.1109/JTEHM.2017.2752152

Tanaka, H., Negoro, H., Iwasaka, H., and Nakamura, S. (2017b). Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. *PLoS ONE* 12:182151. doi: 10.1371/journal.pone.0182151

Tanaka, H., Negoro, H., Iwasaka, H., and Nakamura, S. (2018a). "Listening skills assessment through computer agents," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction (ICMI) ACM* (Boulder, CO), 492–496. doi: 10.1145/3242969.3242970

Tanaka, H., Sakti, S., Neubig, G., Toda, T., Campbell, N., and Nakamura, S. (2012). "Non-verbal cognitive skills and autistic conditions: an analysis and training tool," in *IEEE International Conference on Cognitive Infocommunications (CogInfoCom)* (Kosice), 41–46. doi: 10.1109/CogInfoCom.2012.6422034

Tanaka, H., Sakti, S., Neubig, G., Toda, T., Negoro, H., Iwasaka, H., et al. (2015). "Automated social skills trainer," in *Proceedings of the 20th International Conference on Intelligent User Interfaces. (ACM)* (Atlanta, GA), 17–27. doi: 10.1145/2678025.2701368

Tanaka, H., Watanabe, H., Maki, H., Sakti, S., and Nakamura, S. (2018b). "Single-trial detection of semantic anomalies from EEG during listening to spoken sentences," in *40th IEEE International Engineering in Medicine and Biology Conference* (Honolulu, HI), 977–980. doi: 10.1109/EMBC.2018.8512370

Tang, Z., Li, C., and Sun, S. (2017). Single-trial EEG classification of motor imagery using deep convolutional neural networks. *Optik Int. J. Light Electron Opt.* 130, 11–18. doi: 10.1016/j.ijleo.2016.10.117

Terasawa, N., Tanaka, H., Sakti, S., and Nakamura, S. (2017). "Tracking liking state in brain activity while watching multiple movies," in *Proceedings of the 19th*

*ACM International Conference on Multimodal Interaction. (ACM)* (Glasgow, UK), 321–325. doi: 10.1145/3136755.3136772

Ueno, M., and Kluender, R. (2003). Event-related brain indices of Japanese scrambling. *Brain Lang.* 86, 243–271. doi: 10.1016/S0093-934X(02)00543-6

Ujiro, T., Tanaka, H., Adachi, H., Kazui, H., Ikeda, M., Kudo, T., et al. (2018). "Detection of dementia from responses to atypical questions asked by embodied conversational agents," in *Interspeech* (Hyderabad, India), 1691–1695.

Vail, A. K., Liebson, E., Baker, J. T., and Morency, L.-P. (2018). "Toward objective, multifaceted characterization of psychotic disorders: lexical, structural, and disfluency markers of spoken language," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO), 170–178. doi: 10.1145/3242969.3243020

Vareka, L., and Mautner, P. (2017). Stacked autoencoders for the P300 component detection. *Front. Neurosci.* 11:302. doi: 10.3389/fnins.2017.00302

Wolff, S., Schlesewsky, M., Hirotani, M., and Bornkessel Schlesewsky, I. (2008). The neural mechanisms of word order processing revisited: electrophysiological evidence from Japanese. *Brain Lang.* 107, 133–157. doi: 10.1016/j.bandl.2008.06.003

Yeom, H.-G., and Sim, K.-B. (2008). "ERS and ERD analysis during the imaginary movement of arms," in *IEEE International Conference on Control, Automation and Systems* (Seoul), 2476–2480.

Zoumpoulaki, A., Alsufyani, A., Filetti, M., Brammer, M., and Bowman, H. (2013). ERP latency contrasts using dynamic time warping algorithm. *BMC Neurosci.* 14:434. doi: 10.1186/1471-2202-14-S1-P434

**Conflict of Interest Statement:** 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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 04 April 2019
doi: 10.3389/fncom.2019.00018

# Deep Learning With Asymmetric Connections and Hebbian Updates

*Yali Amit**

*Department of Statistics, University of Chicago, Chicago, IL, United States*

We show that deep networks can be trained using Hebbian updates yielding similar performance to ordinary back-propagation on challenging image datasets. To overcome the unrealistic symmetry in connections between layers, implicit in back-propagation, the feedback weights are separate from the feedforward weights. The feedback weights are also updated with a local rule, the same as the feedforward weights—a weight is updated solely based on the product of activity of the units it connects. With fixed feedback weights as proposed in Lillicrap et al. (2016) performance degrades quickly as the depth of the network increases. If the feedforward and feedback weights are initialized with the same values, as proposed in Zipser and Rumelhart (1990), they remain the same throughout training thus precisely implementing back-propagation. We show that even when the weights are initialized differently and at random, and the algorithm is no longer performing back-propagation, performance is comparable on challenging datasets. We also propose a cost function whose derivative can be represented as a local Hebbian update on the last layer. Convolutional layers are updated with tied weights across space, which is not biologically plausible. We show that similar performance is achieved with untied layers, also known as locally connected layers, corresponding to the connectivity implied by the convolutional layers, but where weights are untied and updated separately. In the linear case we show theoretically that the convergence of the error to zero is accelerated by the update of the feedback weights.

**Keywords: Hebbian learning, asymmetric backpropagation, feedback connections, hinge loss, convolutional networks**

OPEN ACCESS

***Edited by:***
*Wulfram Gerstner,*
*École Polytechnique Fédérale de Lausanne, Switzerland*

***Reviewed by:***
*Benjamin F. Grewe,*
*ETH Zürich, Switzerland*
*Sander Bohte,*
*Centrum Wiskunde & Informatica, Netherlands*

***Correspondence:**
*Yali Amit*
*amit@marx.uchicago.edus*

***Received:*** *22 November 2018*
***Accepted:*** *12 March 2019*
***Published:*** *04 April 2019*

***Citation:***
*Amit Y (2019) Deep Learning With Asymmetric Connections and Hebbian Updates. Front. Comput. Neurosci. 13:18. doi: 10.3389/fncom.2019.00018*

## 1. INTRODUCTION

The success of multi-layer neural networks (deep networks) in a range of prediction tasks as well some observed similarities observed between the properties of the network units and cortical units (Yamins and DiCarlo, 2016), has raised the question of whether they can serve as models for processing in the cortex (Kriegeskorte, 2015; Marblestone et al., 2016). The feedforward architecture of these networks is clearly consistent with models of neural computation: a hierarchy of layers, where the units in each layer compute their activity in terms of the weighted sum of the units of the previous layer. The main challenge with respect to biological plausibility is in the way these networks are trained.

Training of feedforward networks is based on a loss function that compares the output of the top layer of the network to a target. Small random subsets of training data are then used to compute the gradient of the loss with respect to the weights of the network, and these are then updated by moving a small distance in the opposite direction of the gradient. Due to the particular structure

of the function represented by these multi-layer networks the gradient is computed using back-propagation—an algorithmic formulation of the chain rule for differentiation (Rumelhart et al., 1986). In the feedforward step the input is passed *bottom-up* through the layers of the network to produce the output of the top layer and the loss is computed. Back-propagation proceeds top-down through the network. Successively two things occur in each layer: first, the unit activity in the layer is updated in terms of the layer above—feedback, then the weights feeding into this layer are updated. The gradient of each weight is a product of the activity of the units it connects—the feedforward pre-synaptic activity of the input unit in the lower layer and the feedback activity in the post-synaptic unit in the current layer. In that sense the gradient computation has the form of local Hebbian learning. However, a fundamental element of back-propagation is not biologically plausible as explained in Zipser and Rumelhart (1990) and Lillicrap et al. (2016). The feedback activity of a unit is computed as a function of the units in the layer above it in the hierarchy in terms of the *same* weight matrix used to compute the feedforward signal, implying a symmetric synaptic connectivity matrix.

Symmetry of weight connection is an unrealistic assumption. Although reciprocal physical connections between neurons are more common than would be expected at random, these connections are physically separated in entirely different regions of the neuron and can in no way be the same. The solution proposed both in Zipser and Rumelhart (1990) and in Lillicrap et al. (2016) is to create a separate system of feedback connections. The latter model is simpler in that the feedback connections are not updated so that the top-down feedback is always computed with the same weights. The earlier model proposes to update the feedback weights with the same increment as the feedforward weights, which as mentioned above has a Hebbian form. Assuming they are initialized with the same values, they will always have the same value. This guarantees that the back-propagation computation is executed by the network, but in effect reintroduces exact weight symmetry in the back-door, and is unrealistic. In contrast, the computation in Lillicrap et al. (2016) does not replicate back-propagation, as the feedback weights never change, but the price paid is that in deeper networks it performs quite poorly.

The main contribution of this paper is to experiment with the idea proposed in Zipser and Rumelhart (1990), but initialize the feedforward and feedback weights randomly (thus differently). We call this updated random feedback (URFB). We show that even though the feedback weights are never replicates of the feedforward weights, the network performance is comparable to back-propagation, even with deep networks on challenging benchmark datasets such as CIFAR10 and CIFAR100 (Krizhevsky et al., 2013). In contrast, the performance with fixed weights -fixed random feedback (FRFB), as in Lillicrap et al. (2016), degrades with depth. It was noted in Lillicrap et al. (2016) that in shallow networks the feedforward weights gradually align with the fixed feedback weights so that in the long run an approximate back-propagation is being computed, hence the name *feedback alignment*. We show in a number of experiments that this alignment phenomenon is much stronger in URFB even for deep networks. However, we also show that from the very initial iterations of the algorithm, long before the weights have aligned, the evolution of both the training and validation errors is comparable to that of back-propagation.

In our experiments we replace the commonly used unbounded rectified linear unit, with a saturated linearity $\sigma(x) = \min(\max(x, -1), 1)$, which is more biologically plausible, as it is not unbounded, we avoid normalization layers whose gradient is quite complex and not easily amenable to neural computation, and we run all experiments with the simplest stochastic gradient descent that does not require any memory of earlier gradients. We also experiment with randomly zeroing out half of the connections, separately for feedforward and feedback connections. Thus, not only are the feedforward and feedback weights different, but connectivity is asymmetric. In a simplified setting we provide a mathematical argument for why the error decreases faster with updated feedback weights compared to fixed feedback weights.

Another issue arising in considering the biological plausibility of multilayer networks is how the teaching signal is incorporated in learning. The primary loss used for classification problems in the neural network literature is the cross-entropy of the target with respect to the *softmax* of the output layer (see section 3.2). The first step in back-propagation is computing the derivative of this loss with respect to the activities of the top layer. This derivative, which constitutes the feedback signal to the top layer, involves the computation of the softmax—a ratio of sums of exponentials of the activities of all the output units. It is not a local computation and is difficult to model with a neural network. As a second contribution we experiment with an alternative loss, motivated by the original perceptron loss, where the feedback signal is computed locally only in terms of the activity of the top-level unit and the correct target signal. It is based on the one-vs. all method commonly used with support vector machines in the multi-class setting and has been implemented through network models in Amit and Mascaro (2003), La Camera et al. (2004), and Amit and Walker (2012).

Finally, although convolutional layers are consistent with the structure of retinotopic layers in visual cortex, back-propagation through these layers is not biologically plausible. Since the weights of the filters applied across space are assumed identical, the gradient of the unique filter is computed as the sum of the gradients at each location. In the brain the connections corresponding to different spatial locations are physically different and one can't expect them to undergo coordinated updates, see Bartunov et al. (2018). This leads us to the final set of experiments where instead of purely convolutional layers we use a connectivity matrix that has the sparsity structure inherited from the convolution but the values in the matrix are "untied" and undergo independent local updates. Such layers are also called *locally connected* layers and have been used in Bartunov et al. (2018) in experiments with biologically plausible architectures. The memory requirements of such layers are much greater than for convolutional layers, as is the computation, so for these experiments we restrict to simpler architectures.

Overall we observe the same phenomena as with convolutional layers, namely the update of the feedback connections yields performance close to that of regular back-propagation.

The paper is organized as follows. In section 2 we describe related work. In section 3 we describe the structure of a feedforward network, the back-propagation training algorithm and explain how it is modified with separate feedback weights. We describe the loss function and explain why it requires only local Hebbian type updates. In section 4 we report a number of experiments and illustrate some interesting properties of these networks. We show that performance of URFB is lower but close to back-propagation even in very deep networks, on more challenging data sets that actually require a deep network to achieve good results. We show that using locally-connected layers works, although not as well as convolutional networks, and that the resulting filters although not tied apriori show significant similarity across space. We illustrate the phenomenon of weight alignment that is much more pronounced in URFB. In section 5 we describe a simplified mathematical framework to study the properties of these algorithms and show some simulation results that verify that updating the feedback connections yields faster convergence than fixed feedback connections. We conclude with a discussion. Mathematical results and proofs are provided in the **Appendix**.

## 2. RELATED WORK

As indicated in the introduction, the issue of the weight symmetry required for feedback computation in back-propagation, was already raised by Zipser and Rumelhart (1990) and the idea of separating the feedback connections from the feedforward connections was proposed. They then suggested updating each feedforward connection and feedback connection with the same increment. Assuming all weights are initialized at the same value the resulting computation is equivalent to back-propagation. The problem is that this reintroduces the implausible symmetry since the feedback and feedforward weights end up being identical.

In Lillicrap et al. (2016) the simple idea of having fixed random feedback connections was explored and found to work well for shallow networks. However, the performance degrades as the depth of the network increases. It was noted that in shallow networks the feedforward weights gradually align with the fixed feedback weights so that in the long run an approximate back-propagation is being computed, hence the name *feedback alignment*. In Liao et al. (2016) the performance degradation of feedback alignment with depth was addressed by using layer-wise normalization of the outputs. This yielded results with fixed random feedback FRFB that are close to momentum based gradient descent of the back-propagation algorithm for certain network architectures. However, the propagation of the gradient through the normalization layer is complex and it is unclear how to implement it in a network. Furthermore Liao et al. (2016), showed that a simple transfer of information on the sign of the actual back-propagation gradient yields an improvement on using the purely random back-propagation matrix. It is however unclear how such information could be transmitted between different synapses.

In Whittington and Bogacz (2017) a model for training a multilayer network is proposed using a predictive coding framework. However it appears that the model assumes symmetric connections, i.e., the strength of the connection from an error node and a variable in the preceding layer is the same as the reverse connection. A similar issue arises in Roelfsema and Holtmaat (2018), where in the analysis of their algorithm, they assume that in the long run, since the updates are the same, the synaptic values are the same. This is approximately true, in the sense that the correlations between feedforward and feedback weights increase but significant improvement in error rates are observed even early on when the correlations are weak.

Burbank (2015) implements a proposal similar to Zipser and Rumelhart (1990) in the context of an autoencoder and attempts to find STDP rules that can implement the same increment for the feedforward and feedback connections. Again it is assumed that the initial conditions are very similar so that at each step the feedforward and feedback weights are closely aligned.

In a recently archived paper (Pozzi et al., 2018) also goes back to the proposal in Zipser and Rumelhart (1990). However, as in our paper, they experiment with *different* initializations of the feedforward and feedback connections. They introduce a pairing of feedback and feedforward units to model the gating of information from the feedforward pass and the feedback pass. Algorithmically, the only substantial difference to our proposal is in the error signal produced by the output layer, only connections to the output unit representing the correct class are updated.

Here we show that there is a natural way to update all units in the output layer so that subsequent synaptic modifications in the back-propagation are all Hebbian. The correct class unit is activated at the value 1 if the input is below a threshold, and the other classes are activated as $-\mu$ if the input is above a threshold. Thus, corrections occur through top-down feedback in the system when the inputs of any of the output units are not of sufficient magnitude and of the correct sign. We show that this approach works well even in much deeper networks with several convolutional layers and with more challenging data sets. We also present a mathematical analysis of the linearized version of this algorithm and show that the error converges faster when the feedback weights are updated compared to when they are held fixed as in Lillicrap et al. (2016).

Lee et al. (2015) and Bartunov et al. (2018) study target propagation where an error signal is computed in each hidden unit as the difference between the feedforward activity of that unit and a target value propagated from above with feedback connections that are separate from the feedforward connections. The feedback connections between each two consecutive layers are trained to approximate the inverse of the feedforward function between those layers, i.e., the non-linearity applied to the linear transformation of the lower layer. In Bartunov et al. (2018) they analyze the performance of this method on a number of image classification problems and use locally connected layers instead of convolutional layers. In target propagation the losses for both the forward and the backward connections rely on magnitudes of differences between signals requiring a more

complex synaptic modification mechanism than simple products of activities of pre and post-synaptic neurons as proposed in our model.

Such synaptic modification mechanisms are studied in Guerguiev et al. (2017). A biological model for the neuronal units is presented that combines the feedforward and feedback signals within each neuron, and produces an error signal assuming fixed feedback weights as in Lillicrap et al. (2016). The idea is to divide the neuron into two separate compartments one computing feedforward signals and one computing feedback signals, with different phases of learning involving different combinations of these two signals. In addition to computing an error signal internally to the neuron this model avoids the need to compute signed errors, which imply negative as well as positive neuronal activity. However, this is done by assuming the neuron can internally compute the difference in average voltage between two time intervals. In Sacramento et al. (2018) this model is extended to include an inhibitory neuron attached to each hidden unit neuron with plastic synaptic connections to and from the hidden unit. They claim that this eliminates the need to compute the feedback error in separate phases form the feedforward error.

In our model we simply assume that once the feedforward phase is complete the feedback signal *replaces* the feedforward signal at a unit—at the proper timing—to allow for the proper update of the incoming feedforward and outgoing feedback synapses.

## 3. THE UPDATED RANDOM FEEDBACK ALGORITHM

In this section we first describe the structure of a multilayer network, how the back-propagation algorithm works and how we modify it to avoid symmetric connections and maintain simple Hebbian updates to both feedforward and feedback connections. We then describe a loss function, whose derivatives can be computed locally, yielding a Hebbian input dependent update of the weights connecting to the final output layer.

### 3.1. Updated Asymmetric Feedback Connections

A multi-layer network is composed of a sequence of layers $0, \ldots, L$. The data at the input layer is denoted $x_0$. Each layer is composed of $n_l$ units. Let $W_{l,ij}$ be the feedforward weight connecting unit $j$ in layer $l-1$ to unit $i$ in layer $l$. Let $x_l, l = 1, \ldots, L$ be the output of layer $l$, this is computed as

$$x_{l,i} = \sigma(h_{l,i}), \quad h_{l,i} = \sum_{j=1}^{n_{l-1}} W_{l,ij} x_{l-1,j}, i = 1, \ldots, n_l.$$
$$\text{or} \qquad h_l = W_l x_{l-1}, \tag{1}$$

where $\sigma$ is some form of non-linearity and $W_l$ is the $n_l \times n_{l-1}$ matrix of weights connecting layer $l-1$ to layer $l$. We denote $h_{l,i}$ the input of unit $i$ of layer $l$. For classification problems with $C$ classes the top layer $L$, also called the output layer, has $C$ units $x_{L,1}, \ldots, x_{L,C}$. In this last layer no non-linearity is applied, i.e., $x_{L,i} = h_{L,i}$. For given input $x_0$ we can write $x_L = \mathcal{N}(x_0, \mathcal{W})$, where $\mathcal{N}$ represents the function computed through the multiple layers of the network with the set of weights $\mathcal{W}$. The classifier is then defined as:

$$\hat{c}(x_0) = \text{argmax}_i x_{L,i} = \text{argmax}_i \mathcal{N}(x_0, \mathcal{W}).$$

A feedforward network with 3 layers is shown in **Figure 1**.

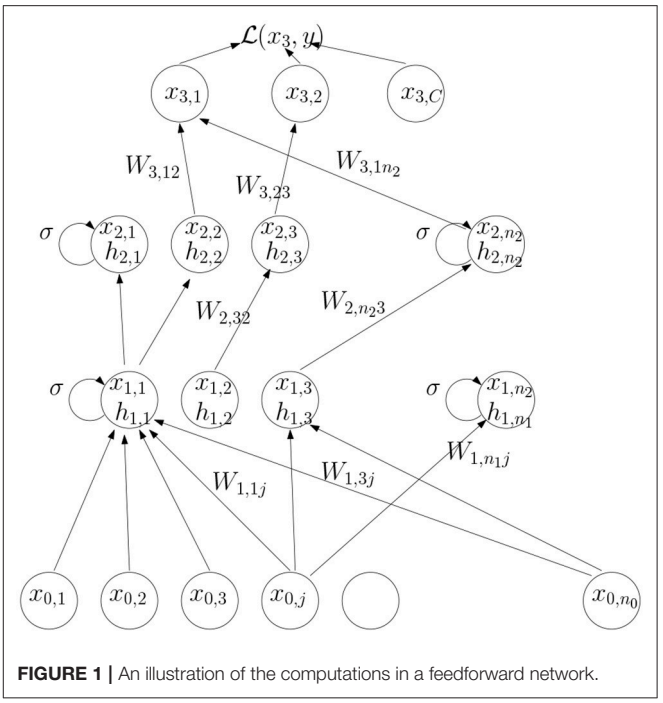


**FIGURE 1 |** An illustration of the computations in a feedforward network.

We define a loss $\mathcal{L}(x_L, y, \mathcal{W})$ comparing the activity of the output layer to a target value, an indicator vector denoting the correct class of the input. At each presentation of a training example the derivative $\partial\mathcal{L}/\partial W_{l,ij}$ of the loss with respect to each weight is computed, and the value of the weight is updated as

$$W_{l,ij} = W_{l,ij} - \eta \partial\mathcal{L}/\partial W_{l,ij},$$

where $\eta$ is a small scalar called the time-step or learning rate. This is done in two phases. In the first phase, the feedforward phase, the input $x_0$ is presented at layer $l = 0$ and passed successively through the layers $l = 1, \ldots, L$ as described in (1). In the second phase the derivatives are computed starting with $W_{L,ij}$ for the top layer and successively moving down the hierarchy. At each layer the following two equalities hold due to the chain rule for differentiation:

$$\frac{\partial\mathcal{L}}{\partial W_{l,ij}} = \frac{\partial\mathcal{L}}{\partial h_{l,i}} \frac{\partial h_{l,i}}{\partial W_{l,ij}} = \frac{\partial\mathcal{L}}{\partial h_{l,i}} x_{l-1,j}$$
$$\frac{\partial\mathcal{L}}{\partial h_{l,i}} = \sigma'(h_{l,i}) \sum_{k=1}^{n_{l+1}} \frac{\partial\mathcal{L}}{\partial h_{l+1,k}} W_{l+1,ki}.$$

If we denote $\delta_{l,i} = \frac{\partial \mathcal{L}}{\partial h_{l,i}}$ we can write this as:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial W_{l,ij}} &= \delta_{l,i} x_{l-1,j} \\ \delta_{l,i} &= \sigma'(h_{l,i}) \sum_{k=1}^{n_{l+1}} \delta_{l+1,k} W_{l+1,ki}. \\ \text{or} \qquad \delta_l &= \sigma'(h_l) W_{l+1}^t \delta_{l+1}, \end{aligned} \tag{2}$$

where $\sigma'(h_l)$ is the diagonal matrix with entries $\sigma'(h_{l,i})$ on the diagonal. So we see that the update to the synaptic weight $W_{l,ij}$ is the product of the *feedback* activity at unit $i$ of layer $l$ denoted by $\delta_{l,i}$, also called the *error signal*, and the input activity from unit $j$ of layer $l-1$. The feedback activity (error signal) of layer $l$ is computed in terms of the feedforward weights connecting unit $i$ in layer $l$ to all the units in layer $l+1$. This is the symmetry problem.

We now separate the feedforward weights from the feedback weights. Let $R_{l+1,ik}$ be the feedback weight connecting unit $k$ of layer $l+1$ to unit $i$ of layer $l$. The second equation in (2) becomes:

$$\delta_{l,i} = \sigma'(h_{l,i}) \sum_{k=1}^{n_{l+1}} \delta_{l+1,k} R_{l+1,ik}.$$

If $R = W^t$ we get the original back-propagation update. We illustrate the general updating scheme computation in **Figure 2**.

In Lillicrap et al. (2016) the values of $R$ are held fixed at some random initial value, which we denote *fixed random feedback* (FRFB). In contrast, in our proposal, since $R_{l+1,ik}$ connects the same units as $W_{l+1,ki}$ it experiences the same pre and post-synaptic activity and so will be updated by the same Hebbian increment - $\delta_{l+1,k} x_{l,i}$. We call this method *updated random feedback* - URFB. If the initial values of $R_{l,ik}$ are the same as $W_{l,ki}$ then equality will hold throughout the update iterations resulting in a symmetric system performing precise back-propagation. This is the proposal in Zipser and Rumelhart (1990). We experiment with different initializations, so that the updates are not performing back-propagation, even in the long run after many iterations the weights are not equal, although their correlation increases. We show that classification rates remain very close to those obtained by back-propagation. In addition, in order to increase the plausibility of the model we also experiment with sparsifying the feedforward and feedback connections by randomly fixing half of each set of weights at 0.

**Remark 1:** It is important to note that the feedback activity $\delta_{l,i}$ replaces the feedforward activity $x_{l,i}$ and needs to be computed before the update of the feedforward weights feeding into unit $i$ and the feedback weights feeding out of that unit, but using the original *feedforward* activity $x_{l-1,i}$ of the units in layer $l-1$. This requires a very rigid sequencing of the algorithm from top to bottom.

**Remark 2:** The feedback signal propagates by computing a linear combination of the feedback signals in the higher layers, but is then multiplied by the term $\sigma'(h_{l,i})$. To simplify as much as possible we have employed a non-linearity $\sigma$ of the form

$$\sigma(h) = \max(-1, \min(1, h)),$$

which is simply a saturated linear function at thresholds $-1$ and 1, and $\sigma'(h) = 1$ if $|h| \leq 1$ and 0 otherwise. Thus, the feedback activity $\delta_{l,i}$ is the linear combination of the feedback activities $\delta_{l+1,k}$ in the layer above unless

$$|h_{l,i}| \geq 1, \text{or } |x_{l,i}| = 1. \tag{3}$$

i.e., bottom-up input $h_{l,i}$ is too high or too low, in which case $\delta_{l,i} = 0$. A local network to compute this thresholding is described in **Appendix 1**. The computation of the top level derivative $\delta_{L,i} = \partial \mathcal{L} / \partial h_{L,i}$ will be discussed in the next section.

## 3.2. Loss Function

The softmax loss commonly used in deep learning defines the probability of each output class as a function of the activities $x_{L,i}$ as follows:

$$\text{softmax}(x_L)_c = p_c = \frac{e^{x_{L,c}}}{\sum_{i=1}^{C} e^{x_{L,i}}}, c = 1, \ldots, C.$$

The loss computes the negative log-likelihood of these probabilities:

$$\mathcal{L}(x_L, y) = -\sum_{i=1}^{C} x_{L,i} y_i + \log \sum_{i=1}^{C} e^{x_{L,i}},$$

where $y_c = 1$ if the class of the input is $c$ and $y_i = 0, i \neq c$. Thus, the initial feedback signal is:

$$\delta_{L,i} = \frac{\partial \mathcal{L}(x_L, y)}{\partial x_{L,i}} = y_i - p_i.$$

This requires the computation of the softmax function $p_i$, which involves the activity of all other units, as well as exponentiations and ratios.

The classification loss function used here is motivated by the hinge loss used in standard linear SVMs. In the simplest case of a two class problem we code the two classes as a scalar $y = \pm 1$ and use only one output unit $x_L$. Classification is based on the sign of $x_L$. The loss is given by

$$\mathcal{L}(x_L, y) = \max(1 - x_L y, 0).$$

The derivative of this loss with respect to $x_L$, is simply

$$\frac{\partial \mathcal{L}}{\partial x_L} = \begin{cases} -y & \text{if } y \cdot x_L \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The idea is that the output $x_L$ should have the same sign as $y$ and be sufficiently large in magnitude. Once the output achieves that, there is no need to change it and the loss is zero.

Writing $x_L = W^t x_{L-1}$, this yields the perceptron learning rule *with margin* (see Shalev-Shwartz et al., 2011):

$$\frac{\partial \mathcal{L}}{\partial W_i} = \begin{cases} -x_{L-1,i} & \text{if } y = 1 \text{ and } W^t x_0 \leq 1 \\ x_{L-1,i} & \text{if } y = -1 \text{ and } W^t x_{L-1} \geq -1 \\ 0 & \text{otherwise} \end{cases},$$

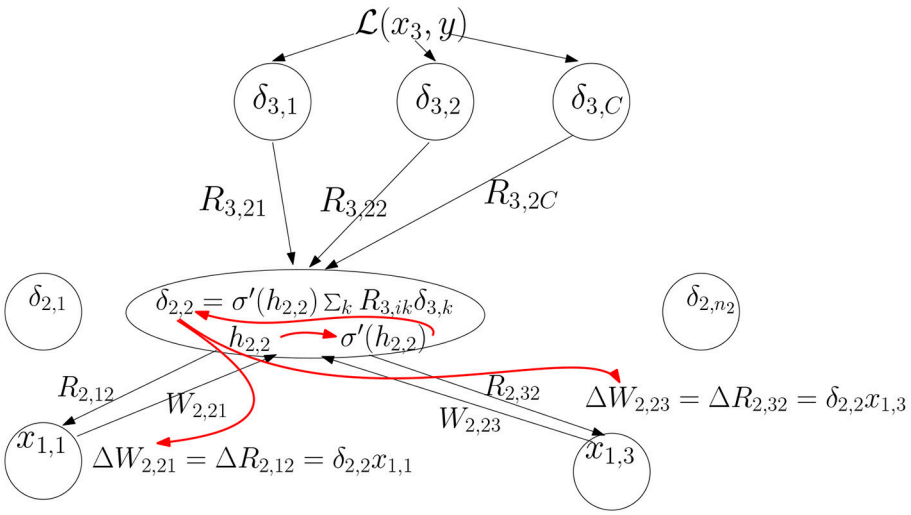


**FIGURE 2 |** The feedback signals $\delta_{3,k}$ from layer 3 are combined linearly and then multiplied by $\sigma'(h_{2,2})$ to produce the feedback signal $\delta_{2,2}$. Then the update to the feedforward weights coming into unit (2, 2) and feedback weights coming out of that unit is computed. The red arrows indicate the order of computation.

If we think of the supervised signal as activating the output unit with $\delta_L = +1$ for one class and $\delta_L = -1$ for the other, unless the input is already of the correct sign and of magnitude greater than 1, then $\delta_L = -\partial\mathcal{L}/\partial x_L$. The update rule can be rewritten as $W_i \leftarrow W_i + \eta\Delta W_i$ where $\Delta W_i = \delta_L \cdot x_{L-1,i}$ if $x_L = W^t x_{L-1}$ satisfies $\delta_L x_L \leq 1$. In other words if the output $x_L$ has the correct sign by more than the margin of **1** then no update occurs, otherwise the weight is updated by the product of the target unit activity and the input unit activity. In that sense the update rule is Hebbian, except for shut down of the update when $x_L$ is "sufficiently correct".

One might ask why not use the unconstrained Hebbian update $\Delta W_i = \eta\delta_L x_{L-1,i}$, which corresponds to a loss that computes the inner product of $y$ and $x$. Unconstrained maximization of the inner product can yield over fitting in the presence of particularly large values of some of the coordinates of $x$ and create an imbalance between the two classes if their input feature distribution is very different. This becomes all the more important with multiple classes, which we discuss next.

For multiple classes we generalize hinge loss as follows. Assume as before $C$ output units $x_{L,1}, \ldots, x_{L,C}$. For an example $x$, of class $c$ define the loss

$$\mathcal{L}(x_L, y) = \max(1 - x_{L,c}, 0) + \mu \sum_{i \neq c} \max(1 + x_{L,i}, 0). \quad (4)$$

where $\mu$ is some balancing factor. The derivative has the form:

$$\frac{\partial\mathcal{L}(x_L, y)}{\partial x_{L,i}} = \begin{cases} -1 & \text{if } i = c \text{ and } x_{L,i} \leq 1 \\ \mu & \text{if } i \neq c \text{ and } x_{L,i} \geq -1 \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

Henceforth we will set $\delta_{L,i} = -\partial\mathcal{L}(x_L, y)/\partial x_{L,i}$. Substituting the feedback signal $\delta_{L,i}$ for the feedforward signal $x_{L,i}$ at the top layer has the following simple form:

$$\delta_{L,i} = \begin{cases} 1 & \text{if } i = c \text{ and } x_{L,i} \leq 1 \\ -\mu & \text{if } i \neq c \text{ and } x_{L,i} \geq -1 \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

and is then applied to compute the feedback to layer $L - 1$ - $\delta_{L-1}$ and the update of the weights $W_L, R_L$. All experiments below use this rule.

Note that $\delta_{L,i}$ is precisely the target signal, *except* when the feedforward signal has the correct value—greater than 1 if $i = c$ (the correct class) and less than $-\mu$ for $i \neq c$ (the wrong class). This error signal only depends on the target value and input to unit $i$, no information is needed regarding the activity of other units. One can ask whether a neuron can produce such an output, which depends both on the exterior teaching signal and on the feedforward activity. In **Appendix 1** we propose a local network that can perform this computation.

This loss produces the well-known one-vs.-all method for multi-class SVM's (see for example Hsu and Lin, 2002), where for each class $c$ a two class SVM is trained for class $c$ against all the rest lumped together as one class. Classification is based on the maximum output of the $C$ classifiers. Each unit $x_{L,c}$ can be viewed as a classifier of class $c$ against all the rest. When an example of class $c$ is presented it updates the weights to obtain a more positive output, when an example of any class other than $c$ is presented it updates the weights to obtain a more negative output. Other global multiclass losses for SVM's can be found in Hsu and Lin (2002). In Amit and Mascaro (2003) and Amit and Walker (2012) a network of binary neurons with discrete synapses was described that implements this learning rule to update connections between discrete neurons in the input and output layers and with positive synapses. Each class was represented by multiple

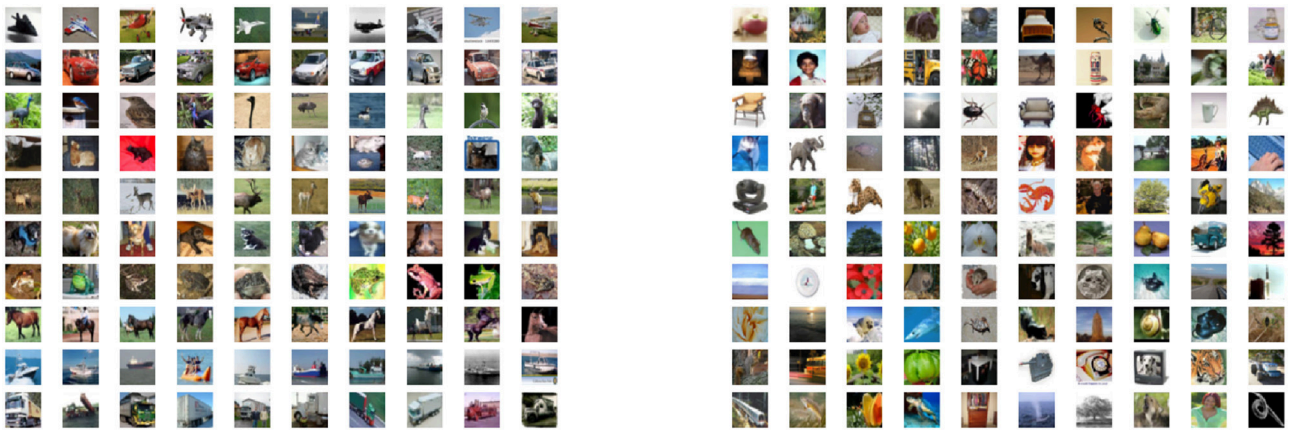

**FIGURE 3 | (Left)** Each row showing 10 images from one of the 10 cifar10 classes. **(Right)** One image from each of the 100 classes in cifar100.

neurons in the output layer. Thus, classification was achieved through recurrent dynamics in the output layer, where the class with most activated units maintained sustained activity, whereas activity in the units corresponding to other classes died out.

## 4. EXPERIMENTS

We report a number of experiments comparing the updated (URFB) to the fixed feedback matrix (FRFB) and comparing the multi-class hinge loss function to the cross-entropy with softmax loss. We restrict ourselves to image data. Since it is quite easy to obtain good results with the widely used MNIST handwritten data set (LeCun et al., 1999) we focus on two more challenging data sets called CIFAR10 and CIFAR100 (Krizhevsky et al., 2013). Each dataset contains 32x32 color images from 10 classes for the first and 100 classes for the second. The classes are broadly defined so that the category bird will contain multiple bird types at many different angles and scales. Some sample images are shown in **Figure 3**. Each data set has 50,000 training images and 10,000 test images.

There are a number of benchmark network architectures that have been developed over the past decade with good results on these datasets, see (Krizhevsky et al., 2012; Simonyan and Zisserman, 2014; Kaiming et al., 2016). These networks are very deep and employ a variety of methods to accelerate convergence, such as adaptive time-steps and batch normalization. These improvements involve steps that are not easily modeled as neural computations. For that reason we restrict our learning method to the simplest form of gradient descent with a fixed time step and no normalization. We do not perform any pre-processing of the input data, nor do we employ any methods of data augmentation to improve classification results. All our weights are initialized based on the method described in Glorot and Bengio (2010). Weights are uniformly drawn between $[-b_l, b_l]$ where $b_l$ is a function of the number of incoming and outgoing connections to a unit in layer $l$.

In the experiments we demonstrate the following:

1. With regular back-propagation (BP) hinge loss performs slightly worse than the softmax loss but results are comparable.
2. For shallow networks URFB performs somewhat better then FRFB but mainly converges faster. It never performs as well as BP but is close.
3. For deeper networks URFB again performs close to BP but FRFB performance degrades significantly.
4. With locally connected—untied—layers replacing convolutional layers results are slightly worse overall but the relationship between the different methods is maintained.
5. In URFB the feedback weights are never the same as the feedforward weights, although the correlation between the two sets of weights increases as a function iteration.
6. Even in initial iterations, when the weights are far from being aligned, training, and validation error rates decrease at similar rates to back propagation.

We first experiment with a shallow network with only two hidden layers, one convolutional and one fully connected.

```
simpnet: Conv 32 5x5; Maxpool 3; Drop .8;
         Full 500; Drop .3; Output
```

The notation `Conv 32 5x5` means that we have 32—5x5 filters, each applied as a convolution to the input images, producing 32 output arrays of dimension 32x32. `Maxpool 3` means that at each pixel the maximal value in the 3x3 window centered at that pixel is substituted for the original value (padding with 0's outside the grid), in each of the 32 output arrays, and then only every second pixel is recorded producing 32 arrays of size 16x16. Drop 0.8 means that for each training batch, a random subset of 80% of the pixels in each array are set to 0 so that no update occurs to the outgoing synaptic weights. This step was introduced in Srivastava et al. (2014) as a way to avoid overfitting to the training set. It is also attractive as a model for biological learning as clearly not all synapses will update at each iteration. `Full 500` means a layer with 500 units, each connected to every unit in the previous layer.

The `Output` layer has $C$ output units one for each class. We use the saturated linearity $\sigma(x) = \min(\max(x, -1), 1)$ and the hinge loss function as given in (4). The update is a simple SGD with a fixed time step of 0.1, and the network is trained for 1,000 epochs with batch-size of 500. We make a point to avoid any adaptive normalization layers as these require a complex gradient that is not amenable to simple neural computations. We avoid the more sophisticated time step adaptations which depend on previous updates and some normalizations, which again do not seem amenable to simple neural computations.

The three parameters we adjusted were the time step and two drop out rates. We experimented with time-steps 0.01, 0.1, and 1.0 for the `simpnet` and found the best behavior on a held out validation set of 5,000 samples was with the value 0.1. We kept this value for all further experiments. We had two dropout layers in each network. One between convolutional layers and one before the output layer. The values were adjusted by running a few tens of iterations and making sure the validation loss was closely tracking the training loss.

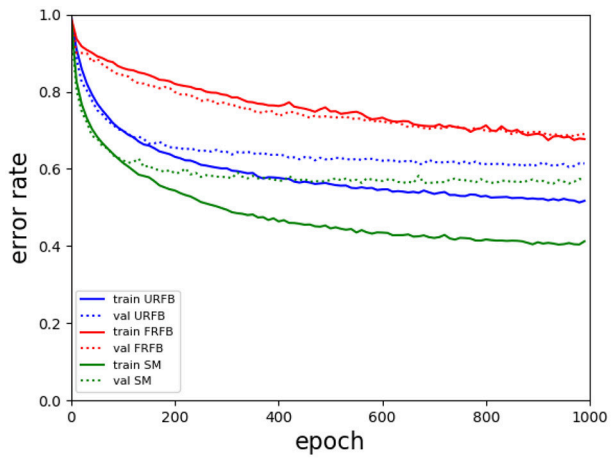


**FIGURE 4 |** Evolution of error rates for `simpnet` as a function of epochs. Solid lines training error, dotted lines validation error. Green–BP, Blue–URFB, Red–FRFB.

We also experiment with pruning the forward and backward connections randomly by 50%. In other words half of these connections are randomly set to 0. The evolution of error rates for the different protocols for simpnet as a function of protocol can be seen in **Figure 4**. Error rates for CIFAR10 and CIFAR100 datasets are shown in **Figure 5**. We note that the use of the multi-class hinge loss leads to only a small loss in accuracy relative to softmax. All experiments with random feedback are performed with the hinge loss. For CIFAR10 the difference between R fixed - FRFB - and R updated - URFB - is small, but becomes more significant when connectivity is reduced to 50% and with the CIFAR100 database.

Note that in the simple network the only layer propagating back an error signal is the fully connected layer. The first layer, which is convolutional, does not need to back-propagate an error.

We experiment with a deep network with multiple convolutional layers, and observe an even larger difference between R fixed and R updated. With the deep network FRFB performs very poorly. The deep architecture is given here.

```
deepnet: Conv 32 5x5; Maxpool 3; Conv 32
         3x3; Conv 32 3x3; Maxpool 3;
         Drop .8;
         Conv 32 3x3; Conv 32 3x3; Maxpool
         3; Drop .3; Full 500; Output
```

Finally we try an even deeper network with residual layers as in Kaiming et al. (2016). This means that after every pair of consecutive convolutional layers at the same resolution we introduce a layer that adds the two previous layers with no trainable parameters. This architecture was found to yield improved results on a variety of datasets.

```
deepernet: conv 16 3x3; conv 16 3x3;
```

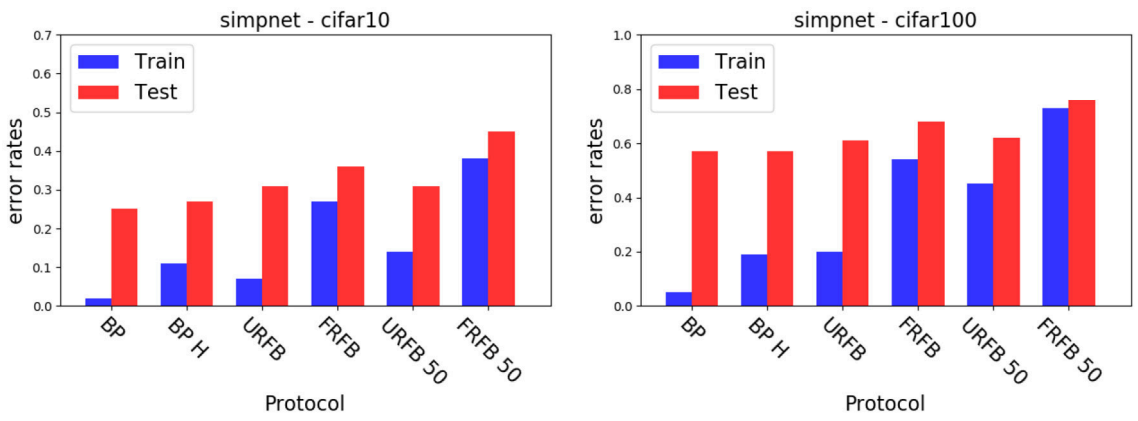


**FIGURE 5 |** Error rates for `simple` network with different update protocols and different losses. **(Left)** CIFAR10, **(Right)** CIFAR100. BP, back-propagation with softmax and cross entropy loss; BP-H, back propagation with hinge loss, all other protocols use the hinge loss as well; URFB, Updated random feedback; FRFB, Fixed random feedback. 50% refers to random connectivity.

```
SUM; conv 32 3x3; conv 32 3x3;
SUM; maxpool 3; drop .5; conv
64 3x3; conv 64 3x3; SUM;
maxpool 3; conv 128 3x3; conv 128
3x3; SUM; maxpool 3; drop .8;
full conn. 500; output
```

We see in **Figure 6** that for the default BP with softmax or hinge loss the error rate decreases from 50% with `deepnet` to 42% with `deepernet`. URFB also shows a decrease in error between `deepnet` and `deepernet` and again FRFB performs very poorly. The evolution of error rates for the different protocols as a function of iteration can be seen in **Figure 7**.

## 4.1. Untying the Convolutional Layers - Locally Connected Layers

We explore "untied" local connectivities determined by the corresponding convolutional layer. These are also called locally connected layers (Bartunov et al., 2018). A convolution corresponds to multiplication by a sparse matrix where the entry values are repeated in each row, but with some displacement. This again is not plausible because it assumes identical weights across a retinotopic layer. Furthermore the back-propagation update of a particular weight in a convolutional layer computes the *sum* of all products $\sum_i \delta_{l,i} x_{l-1,i+k}$, where $i$ represents locations on the grid and $k$ is a fixed displacement. So, it assumes that each one of the identical weights is updated by information summed across the entire layer.

To avoid these issues with biological plausibility we instead assume each of the entries of the sparse matrix is updated separately with the corresponding product $\delta_{l,i} x_{l-1,i+k}$. Only non-zero elements of the sparse matrix, that correspond to connections implied by the convolutional operation are updated. This is implemented using tensorflow sparse tensor operations, and is significantly slower and requires more memory than the ordinary convolutional layers. The error rates are similar to those with the original convolutional layers even with the deeper networks. In **Figure 9** for CIFAR10, we show a comparison of error rates between networks with convolutional layers to networks with corresponding untied layers for the different training protocols. We show comparisons for `simpnet` and `deepnet_s` defined below.

Despite the fact that the weights are updated without being tied across space, the final connectivity matrix retains a strong spatial homogeneity. In other words at each location of the output layer one can restructure the weights to a filter and inspect how similar these filters are across locations. We presume that this is due to the fact that in the data local structures are consistent across space. In **Figure 8** we show a couple of these 5x5 filters across four different locations in the 32x32 grid in the trained `simpnet`. We see that even after 1,000 iterations there is significant similarity in the structure of the filters despite the fact that they were updated independently for each location.

We also experiment with a deeper network:

```
deepnet_s: conv 16 3x3; conv 16 3x3;
```

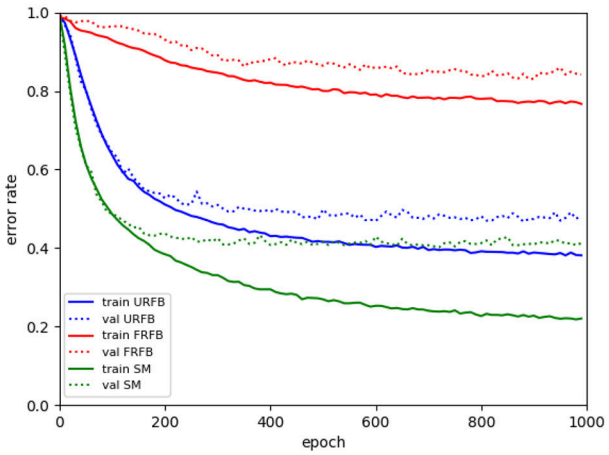


**FIGURE 7 |** Evolution of error rates for `deepernet` as a function of epochs. Solid lines training error, dotted lines validation error. Green–BP, Blue–URFB, Red–FRFB.

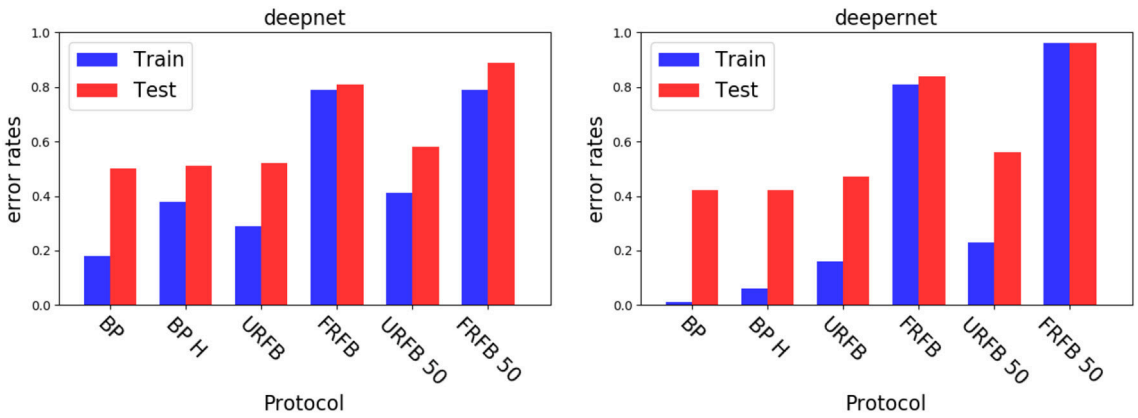


**FIGURE 6 |** Error rates for the `deepnet` **(Left)** and `deepernet` **(Right)**. BP, back-propagation with softmax and cross entropy loss; BP-H, back propagation with hinge loss, all other protocols use the hinge loss as well; URFB, Updated random feedback; FRFB, Fixed random feedback. 50% refers to random connectivity.

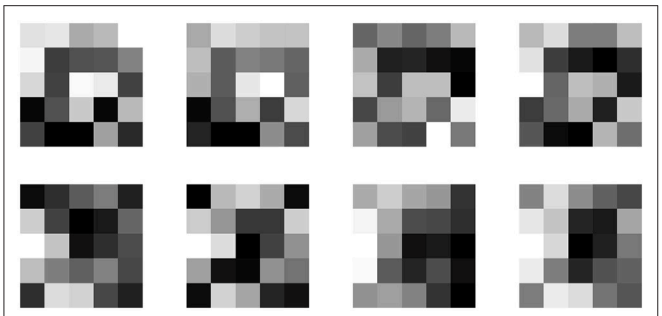

**FIGURE 8 |** Corresponding filters extracted from the sparse connectivity matrix at four different locations on the 32x32 grid. Each row corresponds to a different filter.

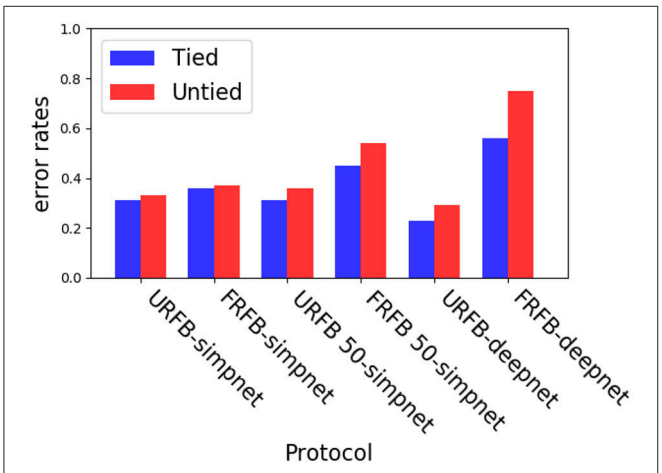


**FIGURE 9 |** Experiments with untying the convolutional layers on `simpnet` and `deepnet_s`. Blue–convolutional layers (tied), Red–untied.

```
SUM;maxpool 3, stride 3; drop .5;
        conv 64 3x3; conv 64 3x3; SUM;
        maxpool 2, stride 2;
        conv 64 2x2; conv 64 2x2; SUM;
        maxpool 2, stride 2; drop .5;
        full conn. 500; output
```

Here we could not run all convolutional layers as untied layers due to memory constraints on our GPUs. Instead we ran the network for 100 epochs with the regular convolutional layers, then we froze the first layer and retrained the remaining layers from scratch using the untied architecture, see **Figure 9**. This would mimic a situation where the first convolutional layer perhaps corresponding to V1 has connections that are predetermined and not subject to synaptic modifications. Once more, we see that the untied layers with URFB reach error rates similar to those of the regular convolutional layers with standard gradient descent. And again, we observe that with a deeper network FRFB performance is much worse.

## 4.2. Weight Alignment

One of the claims in Lillicrap et al. (2016) is that the network gradually aligns the updated feedforward weights to the fixed feedback weights. In **Figure 10** we show the evolution of the correlations between the feedforward weights $W_l$ and $R_l$ for `simpnet`. Recall that the layer with highest index is the output layer and typically reaches high correlations in both URFB and FRFB. We see, however, that the alignment is much stronger for the URFB. Note that when weights are highly correlated the network is effectively implementing back-propagation.

In **Figure 11** we again show the evolution of the correlations between $W_l, R_l$ for the seven updated layers of the deeper network `deepnet_s`. Note that for some but not all layers the final correlations are very close to one. However, the training loss and error rates change very rapidly in the initial iterations when the correlations are very low. Interestingly the correlation levels are not a monotone function of layer depth.

# 5. MATHEMATICAL ANALYSIS OF UPDATED RANDOM FEEDBACK

The mathematical analysis closely follows the methods developed in Saxe et al. (2013) and thus focuses on linear networks, i.e., $\sigma(x) = x$ and a simple quadratic loss. We start with a simple two layer network.

Let the input $x \in \mathbb{R}^{n_0}$, and the output $y = W_2 W_1 x \in \mathbb{R}^{n_2}$ with weights $W_1 \in \mathbb{R}^{n_1 \times n_0}$, $W_2 \in \mathbb{R}^{n_2 \times n_1}$. If $X$ is the $n_0 \times N$ matrix of input data and $Y$ the $n_2 \times N$ of output data the goal is to minimize

$$C(W_1, W_2) = |Y - W_2 W_1 X|^2.$$

We write $T = YX^t \in \mathbb{R}^{n_2 \times n_0}$, and assume that $XX^t = I$, namely the input coordinates are uncorrelated. The gradient of $L$ with respect to $W_1$ and $W_2$ yields the following gradient descent ODE's, which corresponds to regular back-propagation:

$$\begin{aligned}\dot{W}_2 &= (T - W_2 W_1) W_1^t \\ \dot{W}_1 &= W_2^t (T - W_2 W_1),\end{aligned}$$

with some initial condition $W_1(0), W_2(0)$. If we implement the FRFB or URFB described above we get the following three equations:

$$\begin{aligned}\dot{W}_2 &= (T - W_2 W_1) W_1^t \\ \dot{W}_1 &= R_2 (T - W_2 W_1) \\ \dot{R}_2 &= \epsilon W_1 (T - W_2 W_1)^t,\end{aligned} \tag{7}$$

where $R_2 \in \mathbb{R}^{n_1 \times n_2}$ and $\epsilon$ is a parameter. Setting $\epsilon = 0$ corresponds to FRFB, as there is no modification of the matrix $R$. The URFB corresponds to $\epsilon = 1$. Our goal is to show that the larger $\epsilon$ the faster the convergence of the error to 0.

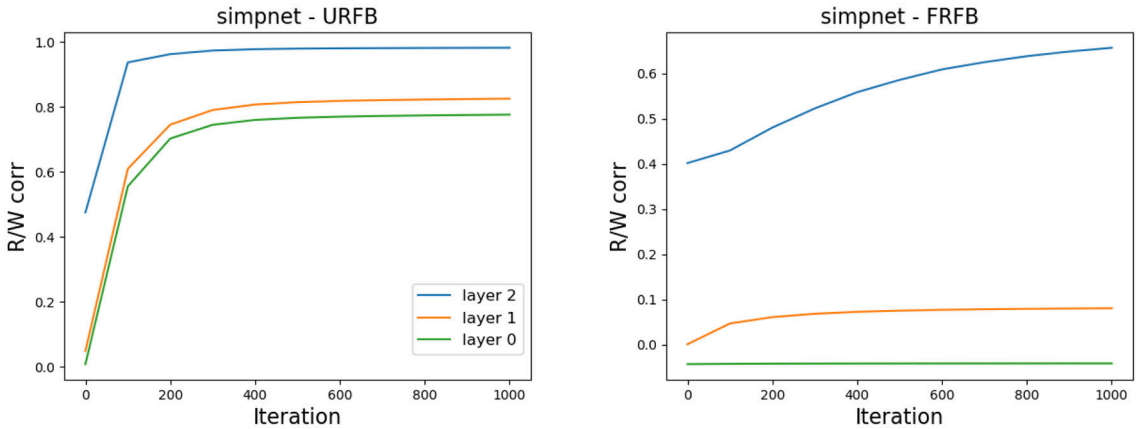


**FIGURE 10 |** Correlation between $W_l$ and $R_l$ for the three layers in `simpnet`. **(Left)** URFB, **(Right)** FRFB.

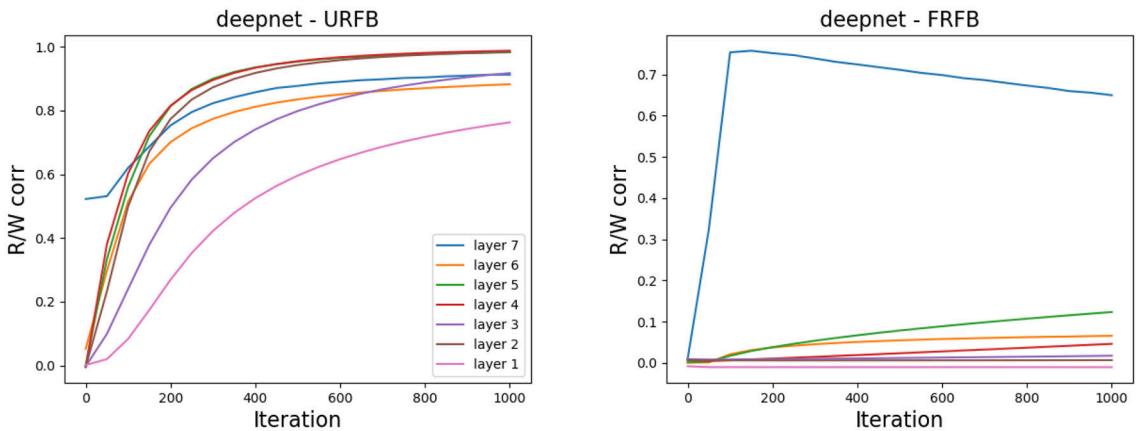


**FIGURE 11 |** Correlation between $W_l$ and $R_l$ for the seven updated layers in `deepnet_s`. **(Left)** URFB, **(Right)** FRFB.

To simplify the analysis of (7) we assume $W_1(0) = W_2(0) = 0$ and $R_2(0)$ is random. Then $R_2 = R_2(0) + \epsilon W_2^t$ and the system reduces to

$$\begin{aligned}\dot{W}_2 &= (T - W_2 W_1) W_1^t \\ \dot{W}_1 &= (R_2(0) + \epsilon W_2^t)(T - W_2 W_1).\end{aligned} \tag{8}$$

For deeper networks, and again assuming the $W_l$ matrices are initialized at 0, we have the following equations for URFB:

$$\begin{aligned}\dot{W}_k &= E W_1^t \cdots W_{k-1}^t \\ &\vdots \\ \dot{W}_i &= (R_{i+1}(0) + \epsilon W_{i+1}^t) \cdots (R_k(0) + \epsilon W_k^t) E W_1^t \cdots W_{i-1}^t \\ &\vdots \\ \dot{W}_1 &= (R_2(0) + \epsilon W_2^t) \cdots (R_k(0) + \epsilon W_k^t) E,\end{aligned} \tag{9}$$

where $E = T - W_k \cdots W_1$, $T \in \mathbb{R}^{n_k \times n_0}$ and $W_i \in \mathbb{R}^{n_i \times n_{i-1}}, i = 1, \ldots, k$. Again our goal is to show that as $\epsilon$ increases from 0 to 1, the error given by $e = tr(E^t E)$ converges faster to 0.

The precise statements of the results and the proofs can be found in **Appendix 2**. Here we show through a simulation that convergence is indeed faster as $\epsilon$ increases from $\epsilon = 0$ (FRFB) to $\epsilon = 1$ (URFB).

## 5.1. Simulation

We simulated the following setting. An input layer of dimension 40, two intermediate layers of dimension 100 and an output layer of dimension 10. We assume $X = I_{40}$ so that $T = W_1^* W_2^* W_3^*$ with $W_1^* \in R^{40\times100}, W_2^* \in R^{100\times100}, W_3^* \in R^{100\times10}$. We choose the $W_i^*$ to have random independent normal entries with sd = 0.2. We then initialize the three matrices randomly as $W_i(0), i = 1, 2, 3$ to run regular back propagation. For comparison we initialize $W_i(0) = 0$ and initialize $R_i(0)$ randomly. We run the differential equations with $\epsilon = 0, 0.25, 0.5, 1.$, where $\epsilon = 0$ corresponds to *FRFB* and $\epsilon = 1$ to URFB. We run 1,000 iterations until all 5 algorithms have negligible

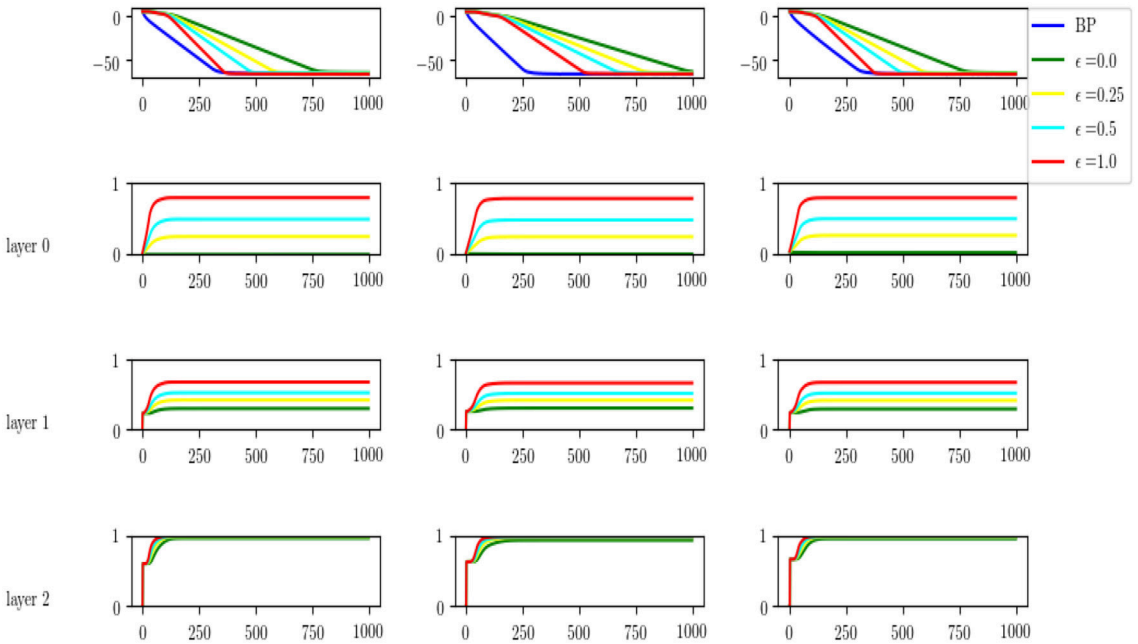


**FIGURE 12 |** Top: comparison of log-error rates as a function of iteration for original BP and for four different values of $\epsilon = 0, 0.25, 0.5, 1$. Results for three runs of the experiment. Last three rows, for each level of the network we show the evolution of the correlation between the $W$ and $R^t$ weights, for each of the values of $\epsilon$.

error. We see the results in **Figure 12**. In the first row, for 3 different runs we show the log-error as a function of iteration, and clearly convergence rate increases with $\epsilon$. In the three rows below that we show the evolution of the correlation of $W_l$ and $R_l^t$ with the same color code. We see that for FRFB (green) the correlation of the weights feeding into the last layer increases to 1 but for the deeper layers that does not hold. Moreover, as $\epsilon$ increases to 1 the correlations approach higher values at each layer. The top layer always converges to a correlation very close to 1, lower layers do not reach correlation 1., and interestingly the correlation reached in the input layer is slightly higher than that of the middle layer. Similar non-monotonicity of the correlation was observed in the experiments in **Figure 11**.

## 6. DISCUSSION

The original idea proposed in Zipser and Rumelhart (1990) of having separate feedback weights undergoing the same Hebbian updates as the feedforward weights yields the original back-propagation algorithm if the feedforward and feedback weights are initialized with the same values. We have shown that even when these weights are initialized differently the performance of the algorithm is comparable to that of back-propagation and significantly outperforms fixed feedback weights as proposed in Lillicrap et al. (2016). The improvement over fixed feedback weights increases with the depth of the network and is demonstrated on challenging benchmarks such as CIFAR10 and CIFAR100. We have also shown that in the long run the feedforward and feedback weights increase their alignment but the performance of the algorithm is comparable to back-propagation even at the initial iterations. We have introduced a cost function whose derivatives lead to local Hebbian updates and provided a proposal for how the associated error signal in the output layer could be implemented in a network. We have shown theoretically, in the linear setting, that adding the update to the feedback weights accelerates the convergence of the error to zero.

These contributions notwithstanding, there are still many aspects of this learning algorithm that are far from biologically plausible. First, although we have removed the need for symmetric connections, we have maintained a symmetric update rule, in that the update of a feedback and feedforward connection connecting two units is the same. To use the formulation in Gerstner et al. (2013) a typical Hebbian update has the form $\Delta W = f(x_{pre})g(x_{post})$, where $f, g$ are typically *not* the same function, however in our setting both $f$ and $g$ are linear which yields a symmetric Hebbian update. In Burbank (2015) it is shown that a mirrored version of STDP could produce this type of symmetric update. Whether this is actually biologically realistic is still an open question.

Another important issue is the timing of the feedforward and feedback weight updates that needs to be very tightly controlled. The update of the feedforward and feedback connections between layer $l$ and $l+1$ requires the feedback signal to layer $l+1$ to have replaced the feedforward signal in all its units, while the feedforward signal is maintained in layer $l$. This issue is discussed in detail in Guerguiev et al. (2017). They propose a neural model with several compartments. One that receives bottom-up or feedforward input and one that receives top-down feedback

input. In a transient phase corresponding to the feedforward processing of the network the top-down input contribution to the neural voltage at the soma is suppressed. Then in a second phase this voltage is allowed in and combined with the feedforward voltage contribution to enable synaptic modifications. In our proposal, instead of combining the two voltages, the top-down voltage would replace the bottom up voltage. Still, in a multilayer network, this would need to be timed in such a way that the previous layer is still responding only to the feedforward input.

An important component of the model proposed in Roelfsema and Holtmaat (2018) are the synaptic tags that maintain the information on the firing of the pre and post-synaptic neurons allowing for a later synaptic modification based on some reinforcement signal. This may offer a mechanism for controlling the timing of the updates. An alternative direction of research would be to investigate the possibility of desynchronizing the updates, i.e., making the learning process more stochastic. If images of similar classes are shown in sequence it could be that it is not so important when the update occurs, as long as the statistics of the error signal and the feedforward signal are the same.

We have defined the network with neurons that have negative and positive values, and synapses with negative and positive weights. Handling negative weights can be achieved with properly adjusted inhibitory inputs. Handling the negative neural activity is more challenging and it would be of interest to explore an architecture that employs only positive neural activity. Finally we mention the issue of the training protocol. We assume randomly ordered presentation of data from all the classes, many hundreds of times. A more natural protocol would be to learn classes one at a time, perhaps occasionally refreshing the memory of previously learned ones. Because our loss function is local and updates to each class label are independent, one could potentially experiment with alternative protocols and see if they are able to yield similar error rates.

## DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: a https://www.cs.toronto.edu/kriz/cifar.html.

## CODE

Code for URFB can be found in https://github.com/yaliamit/URFB.git.

## AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

## ACKNOWLEDGMENTS

This work was supported in part by NIMH award no. R01 MH11555. I'd like to thank Nicolas Brunel, Ulises Pereira, and the referees for helpful comments.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2019.00018/full#supplementary-material

## REFERENCES

Amit, Y., and Mascaro, M. (2003). An integrated network for invariant visual detection and recognition. *Vision Res.* 43, 2073–2088. doi: 10.1016/S0042-6989(03)00306-7

Amit, Y., and Walker, J. (2012). Recurrent network of perceptrons with three state synapses achieves competitive classification on real inputs. *Front. Comput. Neurosci.* 6:39. doi: 10.3389/fncom.2012.00039

Bartunov, S., Santoro A., Richards B. A., Marris L., Hinton G. E., and Lillicrap, T. P. (2018) "Assessing the scalability of biologically-motivated deep learning algorithms and architectures," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018* (Montréal), 9390–9400.

Burbank, K. S. (2015). Mirrored stdp implements autoencoder learning in a network of spiking neurons. *PLoS Comput. Biol.* 11:e1004566. doi: 10.1371/journal.pcbi.1004566

Fusi, S. (2003) Spike-driven synaptic plasticity for learning correlated patterns of mean firing rates. *Rev. Neurosci.* 14:73–84. doi: 10.1515/REVNEURO.2003.14.1-2.73

Gerstner W., Lehmann, M., Liakoni, V., Corneil, D., and Brea, J. (2013). Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules. *Front. Neural Circuits* 12:53. doi: 10.3389/fncir.2018.00053

Glorot, X., and Bengio, Y. (2010) "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, eds Y. W. Teh and M. Titterington (Sardinia: PMLR), 249–256.

Guerguiev, J., Lillicrap, T. P., and Richards, B. A. (2017). Towards deep learning with segregated dendrites. *elife* 6:e22901. doi: 10.7554/eLife.22901

Hsu, C.-W., and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Trans. Neural Netw.* 13, 415–425. doi: 10.1109/72.991427

Kaiming, H., Xiangyu, Z., Shaoqing, R., and Jian,S. (2016). "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 770–778.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* 1, 417–446. doi: 10.1146/annurev-vision-082114-035447

Krizhevsky A., Nair, V., and Hinton,G. (2013). Available online at: https://www.cs.toronto.edu/kriz/cifar.html

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, eds F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc.), 1097–1105.

La Camera, G.,Rauch, A., Luscher, H. R., Senn, W., and Fusi, S. (2004). Minimal models of adapted neuronal response to *in vivo*-like input currents. *Neural Comput.* 16:2101–2124. doi: 10.1162/0899766041732468

LeCun, Y., Cortes, C., and Burges, C. J. C. (1999). Available online at: http://yann.lecun.com/exdb/mnist/

Lee, D.-H.,Zhang, S., Fishcer, A., and Bengio, Y. (2015). "Difference target propagation," in *Machine Learning and Knowledge Discovery in Databases* (Porto: Springer International), 498–515.

Liao, Q., Leibo, J. Z., and Poggio, T. (2016). "How important is weight symmetry in backpropagation?," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16 (Phoenix, AZ: AAAI Press), 1837–1844.

Lillicrap, T. P., Cownden, D., Tweed, D. B., and Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nat. Commun.* 7:13276. doi: 10.1038/ncomms13276

Marblestone, A. H., Wayne, G., and Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Front. Comput. Neurosci.* 10:94. doi: 10.3389/fncom.2016.00094

Pozzi, I., Bohté, S., and Roelfsema,P. R. (2018). A biologically plausible learning rule for deep learning in the brain. *CoRR*, abs/1811.01768.

Roelfsema, P. R., and Holtmaat, A. (2018). Control of synaptic plasticity in deep cortical networks. *Nat. Rev. Neurosci.* 19:166–180. doi: 10.1038/nrn.2018.6

Rumelhart, E. D., Hinton, G. E., and Williams,R. J. (1986). Learning representation by back-propagating errors. *Nature* 323, 533–536. doi: 10.1038/323533a0

Sacramento, J., Costa, R., Bengio, Y., and Senn, W. (2018). "Dendritic cortical microcircuits approximate the backpropagation algorithm," in *Advances in Neural Information Processing Systems 31*, eds S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Montreal, QC: Curran Associates, Inc.), 8735–8746.

Saxe, A. M., McClelland, J. L., and Ganguli, S. (2013). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *CoRR*, abs/1312.6120.

Shalev-Shwartz, S., Singer, Y., Srebro, N., and Cotter, A. (2011). Pegasos: primal estimated sub-gradient solver for svm. *Math. Program.* 127, 3–30. doi: 10.1007/s10107-010-0420-4

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.(2014). Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* 15, 1929–1958.

Whittington, J. C. R., and Bogacz, R. (2017). An approximation of the error backpropagation algorithm in a predictive coding network with local hebbian synaptic plasticity. *Neural Comput.* 29, 1229–1262. doi: 10.1162/NECO_a_00949

Yamins,D. L .K., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19, 356–365.

Zipser, D., and Rumelhart, D. (1990). "The neurobiological significance of the new learning models," in *Computational Neuroscience*, ed E. L. Schwarz (Cambridge, MA: MIT Press).

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

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 12 April 2019
doi: 10.3389/fncom.2019.00021

# End-to-End Deep Image Reconstruction From Human Brain Activity

*Guohua Shen*[1†], *Kshitij Dwivedi*[1†], *Kei Majima*[2], *Tomoyasu Horikawa*[1] *and Yukiyasu Kamitani*[1,2*]

[1] Computational Neuroscience Laboratories, Advanced Telecommunications Research Institute International, Kyoto, Japan, [2] Graduate School of Informatics, Kyoto University, Kyoto, Japan

Deep neural networks (DNNs) have recently been applied successfully to brain decoding and image reconstruction from functional magnetic resonance imaging (fMRI) activity. However, direct training of a DNN with fMRI data is often avoided because the size of available data is thought to be insufficient for training a complex network with numerous parameters. Instead, a pre-trained DNN usually serves as a proxy for hierarchical visual representations, and fMRI data are used to decode individual DNN features of a stimulus image using a simple linear model, which are then passed to a reconstruction module. Here, we directly trained a DNN model with fMRI data and the corresponding stimulus images to build an end-to-end reconstruction model. We accomplished this by training a generative adversarial network with an additional loss term that was defined in high-level feature space (feature loss) using up to 6,000 training data samples (natural images and fMRI responses). The above model was tested on independent datasets and directly reconstructed image using an fMRI pattern as the input. Reconstructions obtained from our proposed method resembled the test stimuli (natural and artificial images) and reconstruction accuracy increased as a function of training-data size. Ablation analyses indicated that the feature loss that we employed played a critical role in achieving accurate reconstruction. Our results show that the end-to-end model can learn a direct mapping between brain activity and perception.

**Keywords: brain decoding, visual image reconstruction, functional magnetic resonance imaging, deep neural networks, generative adversarial networks**

OPEN ACCESS

***Edited by:***
*Taro Toyoizumi,*
*RIKEN Brain Science Institute (BSI), Japan*

***Reviewed by:***
*Takahiro Ezaki,*
*PRESTO, Japan Science and Technology Agency, Japan*
*Topi Tanskanen,*
*RIKEN, Japan*

***Correspondence:**
*Yukiyasu Kamitani*
*kamitani@i.kyoto-u.ac.jp*

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

***Received:*** *23 October 2018*
***Accepted:*** *26 March 2019*
***Published:*** *12 April 2019*

***Citation:***
*Shen G, Dwivedi K, Majima K, Horikawa T and Kamitani Y (2019) End-to-End Deep Image Reconstruction From Human Brain Activity. Front. Comput. Neurosci. 13:21. doi: 10.3389/fncom.2019.00021*

## INTRODUCTION

Advances in the deep learning have opened new directions to decode and visualize the information present in the human brain. In the past few years, deep neural networks (DNNs) have been successfully used to reconstruct visual content from brain activity measured by functional magnetic resonance imaging (fMRI) (Güçlütürk et al., 2017; Han et al., 2017; Seeliger et al., 2018; Shen et al., 2019).

The reconstruction studies avoid training a DNN model directly on the fMRI data because of limited dataset size in fMRI studies. To solve the limited dataset size issue, the feature representation from a DNN pre-trained on a large scale image dataset is usually used as a proxy for the neural representations of the human visual system. Hence, these decoded-feature-based methods involve two independent steps, (1) decoding DNN features from fMRI activity and (2) reconstruction using the decoded DNN features.

Different from fMRI studies, DNNs in computer vision for image generation are usually trained in the end-to-end manner with large datasets. For instance, Mansimov et al. (2015) trained a caption-to-image model on Microsoft COCO dataset that consists of 82,783 images, each annotated with at least 5 captions. Dosovitskiy and Brox (2016a) trained a DNN model on ImageNet training dataset (over 1.2 million images) to reconstruct images from corresponding DNN features. Due to availability of large-scale image datasets, the above image-generation studies can train DNNs using an end-to-end approach to directly generate images from a modality correlated with the images. In contrast, the largest fMRI dataset used for reconstruction in Shen et al. (2019) consisted of only 6,000 training samples. Thus, training a DNN to reconstruct images directly from fMRI data is often avoided and considered infeasible because of the smaller datasets.

Learning a direct mapping between brain activity and perception of the outside world or subjective experiences would be advantageous over the previous decoded-feature-based methods due to the following reason. Decoding features from fMRI is constrained by the pre-trained DNN features which were optimized in a prior without brain data that may not be optimal for decoding them from brain activity. Therefore, information loss occurs in the decoding process which affects the quality of reconstruction. A direct mapping can help in reducing the information loss mentioned above.

In this study, we sought to evaluate the potential of the end-to-end approach for directly mapping fMRI activity to stimulus space given a limited training dataset. In the end-to-end approach, the input to the DNN is the fMRI activity and the output of the DNN is the reconstruction of the perceived stimulus. If reconstruction using the end-to-end approach is successful, we can avoid the feature-decoding step (step 1 above) that leads to information loss.

For designing an end-to-end DNN model to reconstruct images from fMRI data, we considered the models that transform image representations such as DNN features to original image as the potential candidates. The motivation behind this is that the fMRI activity is the neural representation of the perceived image and thus can be considered as an image representation. Also, in previous studies (Agrawal et al., 2014; Khaligh-Razavi and Kriegeskorte, 2014; Güçlü and van Gerven, 2015a,b; Cichy et al., 2016; Horikawa and Kamitani, 2017) fMRI activity has already been shown to be correlated to DNN features. Therefore, for this study, we adopted the model from Dosovitskiy and Brox (2016b) to reconstruct the image from fMRI activity.

For the end-to-end image reconstruction model used in this study, the model needs to be optimized with suitable choice of loss functions relevant to our problem. Dosovitskiy and Brox (2016a) first proposed a DNN-based method for reconstructing original images from their corresponding features by optimization within image space. Loss in image space usually results in poor reconstruction because it generates an average of all possible reconstructions having the same distance in image space, and hence the reconstructed images are blurred. The feature loss in high dimensional space, also called perceptual loss, constrains the reconstruction to be perceptually similar to the original image. Adversarial loss (Goodfellow et al., 2014) constrains the distribution of the reconstructed images to be close to the distribution of natural images. In a subsequent study, Dosovitskiy and Brox (2016b) have also showed that reconstruction from features is improved by introducing feature and adversarial loss terms. Thus, we adopted this latter approach for reconstructing perceived stimuli directly from the fMRI activity. Specifically, we modified their model to take input directly from the fMRI activity and trained the model from scratch with the dataset from Shen et al. (2019).

Here, we present a novel approach to visualize perceptual content from human brain activity by an end-to-end deep image reconstruction model which can directly map fMRI activity in the visual cortex to stimuli observed during perception. Our end-to-end deep image reconstruction model was accomplished by directly training a deep generative adversarial network with a perceptual loss term with fMRI data and the corresponding stimulus images. We demonstrated that the reconstructions obtained from our proposed method resembled the original stimulus images. We further explored the generalizability of our reconstruction model (trained solely with natural images and fMRI responses) to artificial images. To understand the effect of training-dataset size on reconstruction quality, we compared reconstruction results across gradually increasing dataset sizes (from 120 to 6,000 samples). Finally, to investigate the effects of different loss functions used in the reconstruction, we performed an ablation study that objectively and subjectively compared reconstruction results as loss functions were removed one at a time.

## MATERIALS AND METHODS

In this section, we briefly describe the methods we used for our experiments and the details of the dataset. For more details regarding image reconstruction, please refer to Dosovitskiy and Brox (2016b), and for details regarding the dataset, please refer to Shen et al. (2019).

### Problem Statement

Let $\mathbf{x} \in \mathbf{R}^{w \times h \times 3}$ be the stimulus image displayed in the perception experiment, where $w$ and $h$ are width and height of the stimulus image and 3 denotes the number of channels (RGB) of the color image. Let $\mathbf{v} \in \mathbb{R}^D$ be the corresponding preprocessed fMRI vector for the brain activity recorded during the perception of the image, with $D$ being the number of voxels in the visual cortex (VC). We are interested in obtaining a reconstruction of the stimulus from fMRI vector $\mathbf{v}$.

To solve this problem, we use a DNN $\mathbf{G_\theta}$ with parameters $\boldsymbol{\theta}$, which performs non-linear operations on $\mathbf{v}$ to obtain a plausible reconstruction $\mathbf{G_\theta}(\mathbf{v})$ of the stimulus image.

### Image Reconstruction Model

To reconstruct stimulus images from fMRI data, we modified the DNN model proposed by Dosovitskiy and Brox (2016b).

For each fMRI data vector $\mathbf{v}$ corresponding to a stimulus image $\mathbf{x}$, the model was trained to generate a plausible reconstructed image $\mathbf{G_\theta}(\mathbf{v})$. In the training step, the network

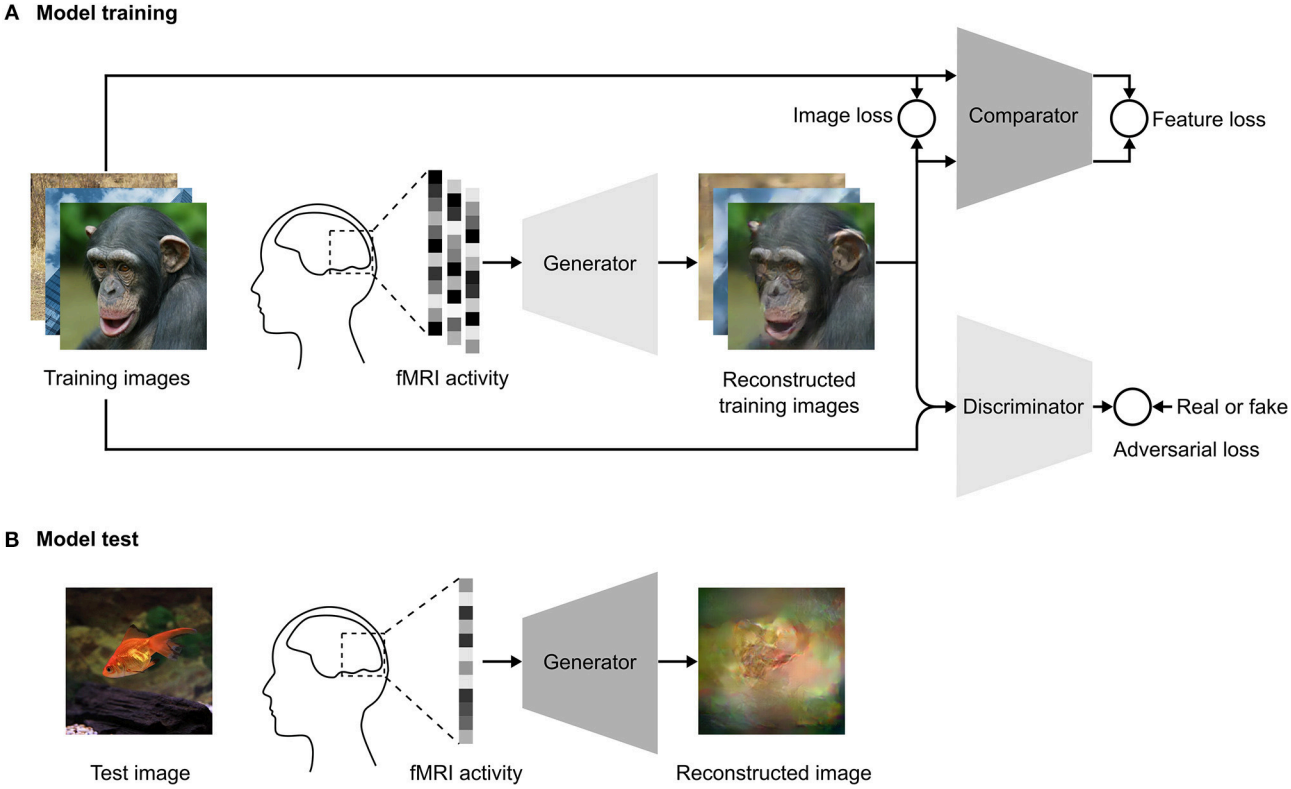


**FIGURE 1 |** Schematics of our reconstruction approach. **(A)** Model training. We use an adversarial training strategy adopted from Dosovitskiy and Brox (2016b), which consists of three DNNs: a generator, a comparator, and a discriminator. The training images are presented to a human subject while brain activity is measured by fMRI. The fMRI activity is used as input to the generator. The generator is trained to reconstruct the images from the fMRI activity to be as similar as possible to the presented training images in both pixel and feature space. The adversarial loss constrains the generator so that reconstructed images fool the discriminator into classifying them as the true training images. The discriminator is trained to distinguish between the reconstructed image and the true training image. The comparator is a pre-trained DNN that was trained to recognize objects in natural images. Both the reconstructed and true training images are used as input to the comparator, which compares the image similarity in feature space. **(B)** Model test. In the test phase, the images are reconstructed by providing the fMRI activity associated with the test image as the input to the generator.

architecture (**Figure 1A**) consisted of three convolutional neural networks: a generator $\mathbf{G_\theta}$ that transformed the fMRI vector $\mathbf{v}$ to $\mathbf{G_\theta(v)}$, a discriminator $\mathbf{D_\Phi}$ that discriminated the reconstructed image $\mathbf{G_\theta(v)}$ from the natural image $\mathbf{x}$, and a comparator $\mathbf{C}$ that compared the reconstructed image $\mathbf{G_\theta(v)}$ with the original stimulus image $\mathbf{x}$ in feature space. During test time, we only need the generator (**Figure 1B**) to reconstruct images from fMRI data.

The input to the generator was the fMRI vector $\mathbf{v}$ from the VC and the output was the reconstructed image $\mathbf{G_\theta(v)}$. The generator consisted of three fully connected layers followed by six upconvolution layers that generated the final reconstruction image $\mathbf{G_\theta(v)}$. The comparator network $\mathbf{C}$ was Caffenet (Krizhevsky et al., 2012), which was trained on the ImageNet dataset for the image classification task. The Caffenet model is a replication of the Alexnet model with the order of pooling and normalization layers switched and without relighting data-augmentation during training. The network consisted of five convolutional and three fully connected layers. We used the last convolutional layer of the comparator to compare the reconstructed image with the original image in feature space. The parameters of the comparator were not updated during training of the reconstruction model.

The discriminator $\mathbf{D_\Phi}$ consisted of five convolutional layers followed by an average pooling layer and two fully connected layers. The output layer of the discriminator was a 2-way softmax and the network was trained to discriminate the reconstructed image from the original image. The generator was trained concurrently to optimize the adversarial loss function, which fooled the discriminator into classifying the reconstructed image as the real stimulus image. The adversarial loss forces the generator to generate more realistic images that are closer to the image distribution of the training data.

The generator was modified to take its input from fMRI data instead of DNN features. The discriminator in Dosovitskiy and Brox (2016b) was provided two inputs, the image and corresponding feature from the comparator, however, we modified the discriminator to receive only the image as the input. The architecture of the comparator network was the same as in Dosovitskiy and Brox (2016b).

Let $\mathbf{X}_i$ denote the $i$ th stimulus image in the dataset, $\mathbf{V}_i$ denote the corresponding fMRI data, and $\mathbf{G}_{\boldsymbol{\theta}}(\mathbf{V}_i)$ denote the corresponding reconstructed output image of the generator. The generator $\mathbf{G}_{\boldsymbol{\theta}}$ had parameters $\boldsymbol{\theta}$, which were updated to minimize the weighted sum of three loss terms for a mini-batch that used stochastic gradient descent: loss in image space $L_{\text{img}}$, feature loss $L_{\text{feat}}$, and adversarial loss $L_{\text{adv}}$:

$$L(\boldsymbol{\theta},\boldsymbol{\Phi}) = \lambda_{\text{img}} L_{\text{img}}(\boldsymbol{\theta}) + \lambda_{\text{feat}} L_{\text{feat}}(\boldsymbol{\theta}) + \lambda_{\text{adv}} L_{\text{adv}}(\boldsymbol{\theta},\boldsymbol{\Phi})$$

where

$$L_{\text{img}}(\boldsymbol{\theta}) = \sum_i \left|\left| \mathbf{G}_{\boldsymbol{\theta}}(\mathbf{V}_i) - \mathbf{X}_i \right|\right|_2^2$$

$$L_{\text{feat}}(\boldsymbol{\theta}) = \sum_i \left|\left| \mathbf{C}\left( \mathbf{G}_{\boldsymbol{\theta}}(\mathbf{V}_i) \right) - \mathbf{C}\left(\mathbf{X}_i\right) \right|\right|_2^2$$

$$L_{\text{adv}}(\boldsymbol{\theta},\boldsymbol{\Phi}) = -\sum_i \log \mathbf{D}_{\boldsymbol{\Phi}}\left( \mathbf{G}_{\boldsymbol{\theta}}(\mathbf{V}_i) \right)$$

and $\lambda_{\text{img}}$, $\lambda_{\text{feat}}$, and $\lambda_{\text{adv}}$ denote the weights of the loss in image space $L_{\text{img}}$, feature loss $L_{\text{feat}}$, and adversarial loss $L_{\text{adv}}$, respectively.

The discriminator was trained concurrently with the generator to minimize $L_{\text{discr}}$:

$$L_{\text{discr}}\left(\boldsymbol{\Phi}\right) = -\sum_i \log\left(\mathbf{D}_{\boldsymbol{\Phi}}\left(\mathbf{X}_i\right)\right) + \log\left(1 - \mathbf{D}_{\boldsymbol{\Phi}}\left(\mathbf{G}_{\boldsymbol{\theta}}\left(\mathbf{V}_i\right)\right)\right).$$

The parameters of the comparator $\mathbf{C}$ were fixed throughout the training because it was only used for the comparison in feature space, and thus did not require any update.

We trained the system using the Caffe framework (Jia et al., 2014) and modified the implementation of the model provided by Dosovitskiy and Brox (2016b). The weights of the generator and discriminator were initialized using MSRA (He et al., 2015) initialization. The comparator weights were initialized by Caffenet weights trained on ImageNet classification. We used Adam solver (Kingma and Ba, 2015) with momentum $\beta_1 = 0.9$, $\beta_2 = 0.999$ and an initial learning rate 0.0002 for optimization. We used a batch size of 64 and trained for 500,000 mini-batch iterations in all experiments. Following this training procedure similar to Dosovitskiy and Brox (2016b), we temporarily stopped updating the discriminator if the ratio of $L_{\text{discr}}$ to $L_{\text{adv}}$ was below 0.1. This was done to prevent the discriminator from overfitting. The weights of the individual loss functions $\lambda_{\text{img}}$, $\lambda_{\text{feat}}$, and $\lambda_{\text{adv}}$ were set to $\lambda_{\text{img}} = 2 \times 10^6$, $\lambda_{\text{feat}} = 0.01$, and $\lambda_{\text{adv}} = 100$.

We applied image jittering during the training for data augmentation and to take into account subject's eye movement during the image presentation experiment. Generally, eye movement was approximately one degree of visual angle for a typical subject. The viewing angle for the stimulus images was 12°. All training images were resized to 248 × 248 pixels before training. During training, we randomly cropped a 227 × 227 pixel window from each training image to use as the target image for each iteration. This ensured that the largest jittering size was approximately one degree viewing angle.

To analyze the size of the dataset, we trained the reconstruction model with a variable number of training samples for 1,000 epochs with a batch size of 60. The rest of the hyperparameters were the same as in the previous analysis.

## Dataset From Shen et al. (2019)

We used an fMRI dataset from our previous reconstruction study (Shen et al., 2019). This dataset was used to reconstruct stimulus images from the visual features of a deep convolutional neural network that was decoded from the brain. The dataset analyzed for this study can be found in the OpenNeuro (https://openneuro.org/datasets/ds001506) repository.

The dataset comprises fMRI data from three human subjects. For each subject, the stimulus images in the dataset are categorized into four types: training and test natural images, artificial shapes, and alphabetical letters. The natural images used for the experiment were selected from 200 representative object categories in the ImageNet dataset (2011, fall release) (Deng et al., 2009). The training dataset of natural images were 1,200 images that were taken from 150 object categories and the test dataset of natural images were 50 images from the remaining 50 object categories. Thus, the categories used in the training and test datasets did not overlap. The artificial shapes were 40 images obtained by combining 8 colors and 5 shapes. The artificial shapes stimuli set was controlled by shape and color, but figure-ground separation and brightness were consistent across all the stimuli. The alphabetical letters were 10 black letters from the English alphabet. The alphabetical letters stimuli set had consistent color, brightness and figure ground separation. The only variable in this stimuli set was the shape of the alphabet.

The image presentation experiments comprised four distinct types of sessions that corresponded to the four categories of stimulus images described above. In one training-session set (natural images), 1,200 images were each presented once. This set of training session was repeated five times. In each test-session (natural image, artificial shape, and alphabetical letters), 50, 40, and 10 images were presented 24, 20, and 12 times each, respectively. The presentation order of the images was randomized across runs.

The fMRI data obtained during the image presentation experiment were preprocessed for motion correction followed by co-registration to the within-session high-resolution anatomical images of the same slices and subsequently to T1-weighted anatomical images. The coregistered data were then re-interpolated as 2 × 2 × 2 mm voxels.

The fMRI data samples were created by first regressing out nuisance parameters from each voxel's amplitude for each run, including a linear trend and temporal components proportional to six motion parameters. These were calculated by the SPM (http://www.fil.ion.ucl.ac.uk/spm) motion correction procedure. After that, voxel amplitudes were normalized relative to the mean amplitude of the initial 24 s rest period of each run, and were despiked to reduce extreme values (beyond ± 3 SD for each run). The voxel amplitudes were then averaged within each 8 s (training sessions) or 12 s (test sessions) stimulus block (four or six volumes), after shifting the data by 4 s (two volumes) to compensate for hemodynamic delays.

The voxels used for reconstruction were selected from the VC, which consisted of lower-order visual areas (V1, V2, V3, and V4) as well as higher-order visual areas (the lateral occipital complex, fusiform face area, and parahippocampal place area). The lower-order regions were identified using retinotopy experiments and the higher-order areas were identified using functional localizer experiments (Shen et al., 2019).

The fMRI data from the training image dataset were further normalized to have zero mean and unit standard deviation for each voxel. The mean and standard deviation of the training fMRI data were then used to normalize the test fMRI data.

We performed trial-averaging for the test fMRI data while we considered each trial as an individual sample for the training fMRI data. Therefore, to compensate for the statistical difference between training and test fMRI data, we rescaled the test fMRI data by a factor of $\sqrt{n}$ where $n$ is number of trials averaged, before we use the test fMRI data as the input to the generator.

We train reconstruction models with the training natural images and their corresponding fMRI data for each individual subject, and test reconstruction models with the test fMRI dataset of the corresponding subject. For training in the dataset size-analysis, we initially selected a fixed number of training images and their corresponding fMRI data from five trials. As we increased the size of the dataset, we added more training images and fMRI data. Specifically, we gradually increased the size of the training dataset from 120 ($5 \times 24$) to 6,000 ($5 \times 1{,}200$) training samples.

## Evaluation

We evaluated the quality of reconstruction using both objective and subjective assessment methods. For both methods, we performed a pairwise similarity comparison, following previous studies (Cowen et al., 2014; Lee and Kuhl, 2016; Seeliger et al., 2018; Shen et al., 2019), in which one reconstructed image was compared with two candidate images: the original stimulus image from which the reconstruction was derived and a "lure" image, which was a different test image. The lure image was randomly selected from the test dataset of the same type as the original stimulus image. For each reconstructed image, the pairwise similarity comparison was conducted for all possible combinations of candidate images: the original stimulus image and every other stimulus image of the same type in the test dataset. For example, to evaluate the reconstruction quality for one of the 50 test natural images, the lure image is randomly selected from the remaining 49 test natural images. Then, for each reconstructed natural image, the pairwise similarity comparison is conducted for all 49 pairs of candidate images.

For the subjective assessment, we conducted a behavioral experiment similar to Shen et al. (2019). In this experiment, a group of 13 raters (6 females and 7 males, aged between 19 and 48 years) were presented with a reconstructed image and two candidate images and were asked to select the candidate image that appeared more similar to the reconstructed image. The trials for different test images were presented in a randomized order for each rater to prevent them from memorizing the correspondence between reconstructed and the true images.

For the objective assessment, we conducted pairwise similarity comparison analysis based on two metrics separately: Pearson correlation coefficient and structural similarity index (SSIM) (Wang et al., 2004). We computed the two metrics between the reconstructed image and each of the two candidate images. For the pairwise similarity comparison, we selected the candidate image with the higher Pearson correlation coefficient or higher SSIM, respectively.

For computing pixel-wise Pearson correlation coefficients, we first reshaped an image (a 3D array with dimensions of height, width, and RGB color channels) into a 1-dimensional vector. During this reshaping, the pixels of different color channels are concatenated in a vector. Then we calculated the Pearson correlation coefficient between the reshaped reconstructed and candidate images.

Since Pearson correlation coefficient considers each pixel as an independent variable, we also used SSIM to take into account the similarity of local structures of the spatially close pixels between two given images. We computed SSIM between the reconstructed and candidate images in the original 2D form for each of the RGB color channels, and then average the SSIM across the RGB color channels.

For both assessments, we calculated the percentage of trials in which the original stimulus image was selected, and used this value as the reconstruction accuracy of each reconstructed image. Trials for each reconstructed image were conducted by pairing the original stimulus image with every other stimulus image of the same type. For the study of dataset size, we reduce the trials for each reconstructed image by randomly selected 500 trials (10 trials for each test image) from all the possible trials, while the selected trials are fixed for all the conditions (here the modes trained with different number of samples) to be compared. For each type of test images (natural images, artificial shapes and alphabetical letters), we used the mean reconstruction accuracy as the quality measure, which was obtained by averaging across all the samples after pooling the three subjects.

We compliment the evaluation using pairwise similarity comparison with modified RV coefficient (Smilde et al., 2009). We compute the modified RV coefficient between two matrices: matrix of the reconstructed images and matrix of the true images. The rows of both these matrices correspond to test samples and columns correspond to individual pixels. With this setting, the modified RV coefficient evaluates the correlation between similarity relation within the true images and within the reconstructed images. We compared the results with a baseline of modified RV coefficient computed with randomly shuffled ordered of reconstructed images and correctly ordered true images to see whether the reconstructions preserve the similarity relation among the true images.

We conducted another behavioral experiment to study the effect of different loss terms in the proposed approach. Another group of 5 raters (2 females and 3 males, aged between 25 and 37 years) were presented with one original stimulus image and two reconstructed images that were generated from different combinations of loss terms. The raters were asked to judge which

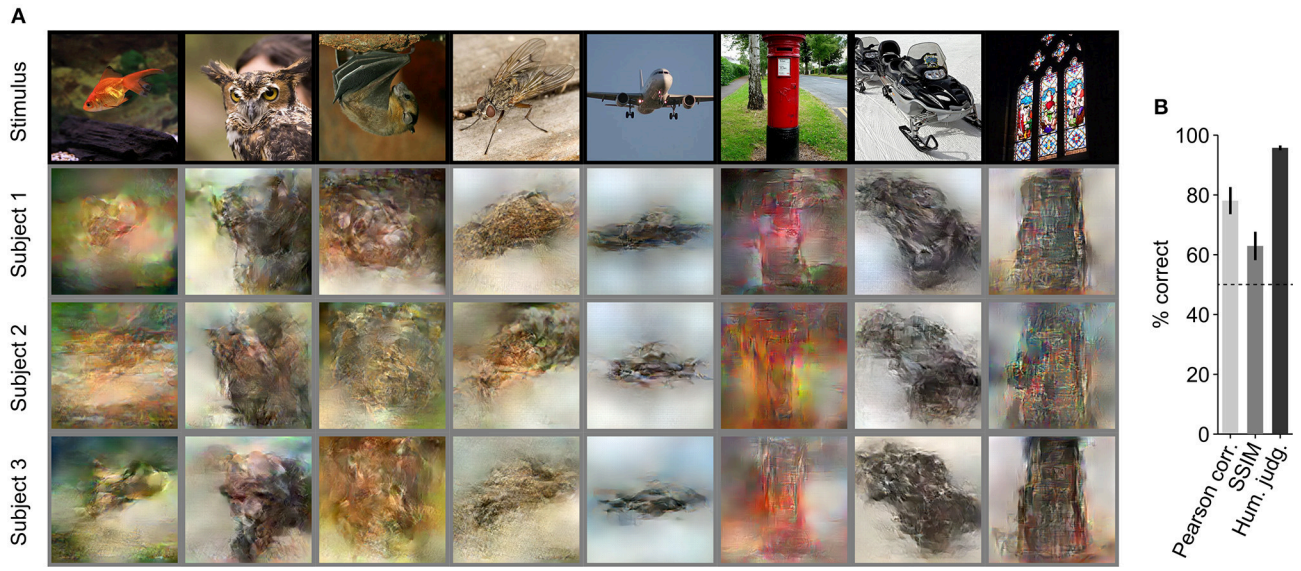


**FIGURE 2 |** Reconstruction of natural images. **(A)** Stimulus and reconstructed natural images. The stimulus images (in black frames) are shown in the top row. Three corresponding reconstructed images (in gray frames) from each of the three subjects are shown underneath. **(B)** Reconstruction accuracy for natural images in terms of the accuracy of pairwise similarity comparison based on Pearson correlation, structural similarity index (SSIM) and human judgment (error bars, 95% confidence interval (CI) across samples; three subjects pooled, the number of samples ($N$) = 150; chance level, 50%).

of the reconstructions more resembled the original stimulus image. This pairwise comparison was conducted for 6 pairs of loss term combinations for each stimulus image in the test dataset. We used the winning percentage as the quantitative measure for comparing reconstructions that were generated using different combinations of loss terms. The winning percentage was the percentage of trials in which the reconstruction from one combination was judged better than that of the other. For computing the winning percentage from objective metrics, the reconstructions with higher similarity (Pearson correlation coefficients or SSIM) were selected. For more details regarding the design of the behavioral experiments, please refer to Shen et al. (2019).

# RESULTS

## Image Reconstruction

We trained the reconstruction model on the Shen et al. (2019) training-session samples of fMRI visual perception data. In the training session, each stimulus image had been presented to each subject five times. Here, we treated each stimulus presentation as a separate training sample for the reconstruction model. Therefore, the training dataset we used consisted of 6,000 ($5 \times 1{,}200$) samples.

We evaluated reconstruction quality using three test datasets: natural images, artificial shapes and alphabetical letters. For generating reconstructions, fMRI samples corresponding to the same image (24 samples for the natural image session, 20 for the artificial shapes session, and 12 for the alphabetical letters session) were averaged across trials to increase the signal to noise ratio. The averaged fMRI samples were used as input to the trained generator (**Figure 1B**). **Figure 2A** shows example images from the natural image test dataset and their corresponding reconstructions from three different subjects. The reconstructions from all three subjects closely resembled shape of the object in the natural image stimuli. The color, however, was not preserved in some of the reconstructions. The reconstruction results from our model show that despite utilizing a small dataset, training a model from scratch and reconstructing visually similar images from fMRI data was possible with high accuracy (**Figure 2B**) The mean reconstruction accuracy (three subjects pooled, $N = 150$) is 78.1% by Pearson correlation (78.9, 75.3, and 79.9% for Subject 1, 2, and 3), 62.9% by SSIM (63.0, 61.9, and 63.8% for Subject 1, 2, and 3), and 95.7% by human judgment (95.6, 95.1, and 96.4% for Subject 1, 2, and 3). Additionally, we calculated modified RV coefficient, which evaluates the correlation between the similarity relation within the true images and the reconstructed images to see whether the reconstructions preserve the similarity relation within the true images. The higher modified RV coefficients (0.34, 0.32, and 0.32 for Subject 1, 2, and 3) for natural image test dataset as compared to the baseline calculated by random permutation ($p < 0.0001$ for all three subjects, permutation test) demonstrate that reconstructed images from our approach preserve the similarity relation within the true images.

Further, we evaluated the generalizability of our reconstruction model (trained solely with natural images and fMRI responses) using artificial images as similarly performed by Shen et al. (2019) (**Figure 3A**). Using the proposed approach, artificial shapes were reconstructed with high accuracy (**Figure 3B**. 69.3% by Pearson correlation, 56.9% by SSIM, and 92.7% by human judgment) and alphabetical letters

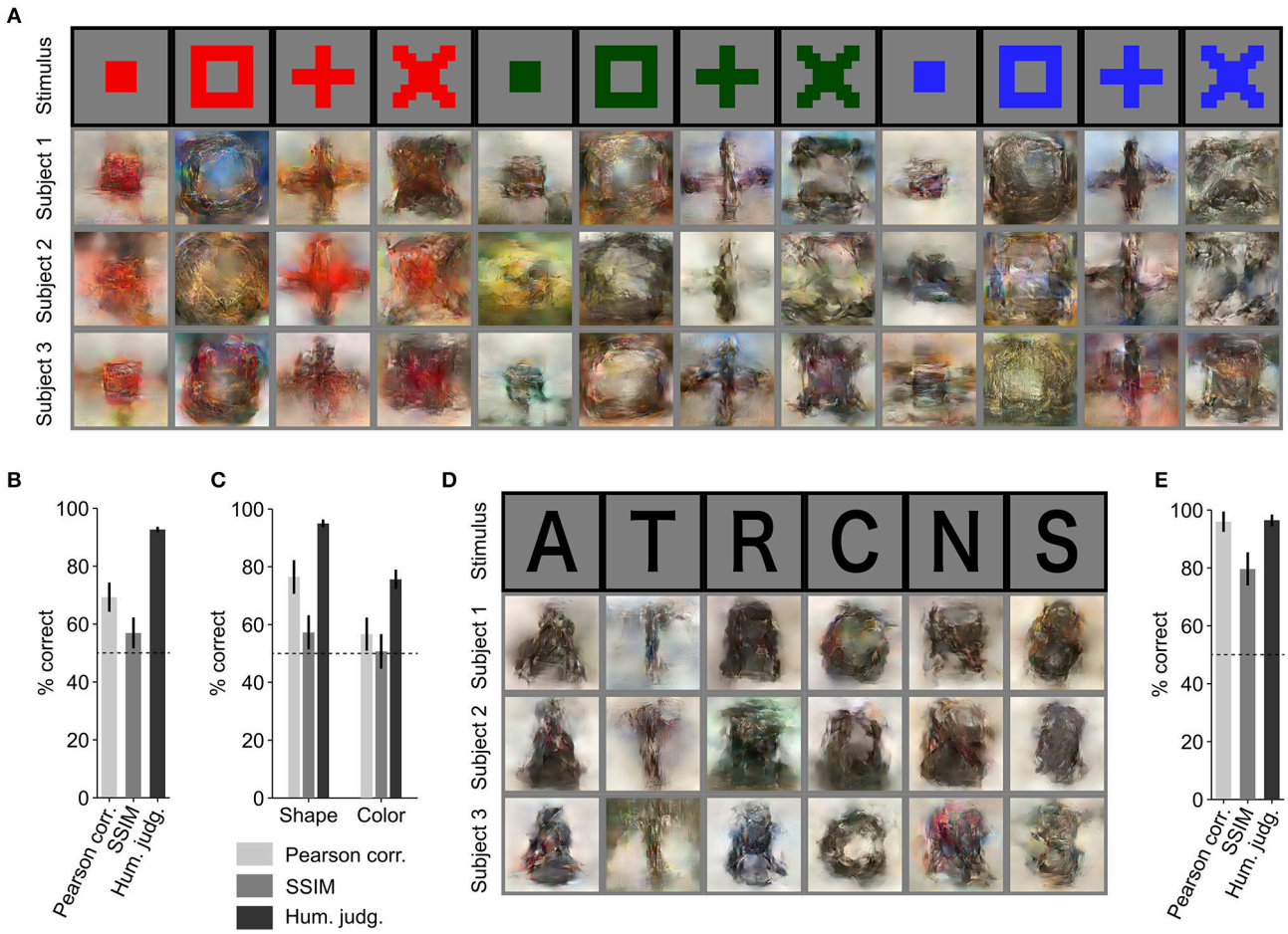


**FIGURE 3 |** Reconstruction of artificial images. **(A)** Reconstruction of artificial shapes. The stimulus images (in black frames) are shown in the top row. Three corresponding reconstructed images (in gray frames) from each of the three subjects are shown underneath. **(B)** Reconstruction accuracy for artificial shapes. **(C)** Reconstruction accuracy for both shape and color. **(D)** Reconstruction of alphabetical letters. **(E)** Reconstruction accuracy for alphabetical letters. For **(B,C,E)**, reconstruction accuracy is assessed in terms of the accuracy of pairwise similarity comparison based on Pearson correlation, structural similarity index (SSIM) and human judgment (error bars, 95% CI across samples; three subjects pooled, $N = 120$ for artificial shapes, $N = 30$ for alphabetical letters; chance level, 50%).

were also reconstructed with high accuracy (**Figures 3D,E**; 95.9% by Pearson correlation, 79.6% by SSIM, and 96.4% by human judgment), even though the model was trained on natural images.

From the results for artificial shape reconstruction, we observed that the shape of the stimulus was well preserved in the reconstructions. However, the color was preserved only for the red-colored shapes. To evaluate reconstruction quality in terms of shape and color, we compared reconstructed images of the same colors and shapes, respectively. The quantitative results are shown in **Figure 3C** (shape: 76.5% by Pearson correlation, 57.3% by SSIM, and 95.0% by human judgment; color: 56.7% by Pearson correlation, 50.7% by SSIM, and 75.6% by human judgment) and confirm that the reconstructed images were more similar in shape to the original images than in color.

While the main purpose of this study is to evaluate the potential of the end-to-end method in learning direct mapping from fMRI data to visual images, we compared the reconstruction accuracy of the proposed method with that of Shen et al. (2019) to analyze the difference between the two methods. We observed that our new method achieved almost same performance as Shen et al. (2019) on the Pearson correlation metric (natural images: ours 78.1 vs. 76.1%; two-sided signed-rank test, no significantly difference, $N = 150$), whereas our new method did not outperform Shen et al. (2019) on the subjective judgment (natural images: ours 95.7 vs. 99.1%; two-sided signed-rank test, $P < 0.006$, $N = 150$). Shen et al. (2019) used a natural image prior that helps their reconstructions look more natural, which could explain why that method outperforms our new method in terms of human judgment. We tried to introduce a natural image prior through use of a discriminator, but the reconstructions did not appear as natural as those from Shen et al. (2019).

## Effect of Dataset Size

The results of the previous analyses show that it is possible to reconstruct images from human brain activity by training

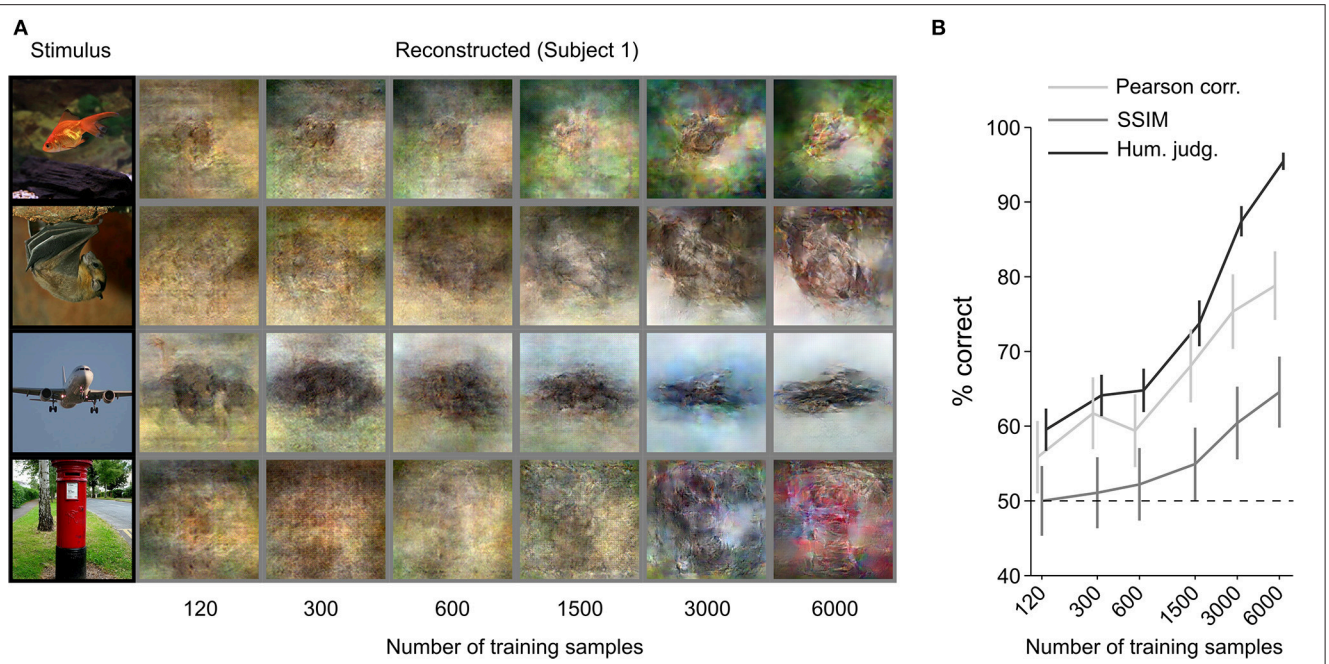


**FIGURE 4 |** Effect of training-dataset size. **(A)** Reconstruction from brain activity (Subject 1) using models trained with different dataset sizes. The stimulus images (in black frames) are shown in the first column. The corresponding reconstructed images (in gray frames) are shown to the right of each stimulus image (from left to right, the number of training samples increases). **(B)** Reconstruction accuracy in terms of the accuracy of pairwise similarity comparison based on Pearson correlation, structural similarity index (SSIM) and human judgment (error bars, 95% CI across samples; three subjects pooled, $N = 150$, chance level, 50%). The horizontal axis is scaled using a base 2 logarithm.

an end-to-end model from scratch with only 6,000 training samples. Next, we sought to investigate the effect of dataset size on reconstruction quality. We checked how many samples are enough to achieve recognizable reconstruction and assessed the possibility of improving reconstruction quality using more training samples.

We increased the training dataset from 120 to 6,000 (120, 300, 600, 1,500, 3,000, and 6,000) samples. **Figure 4** shows a qualitative comparison of reconstructions (**Figure 4A**) and the quantitative objective and human judgment scores (**Figure 4B**). Through visual inspection of the reconstruction results in **Figure 4A**, we can see that reconstruction quality improved with the number of training samples. Objective and human judgment scores quantitatively confirm this trend. The results showed that the increasing trend in the reconstruction quality is not saturated for our reconstruction model, which suggests that although we can obtain highly accurate reconstructions with only 6,000 training samples, better reconstruction quality might be achieved if larger datasets are available.

## Effect of Loss Functions: Ablation Study

We performed an ablation study to understand the effects of the different loss functions used in training the reconstruction model. We removed one loss function at a time and compared the reconstructions with those obtained using all three loss functions. Visual inspection showed that the best resemblance to the original images was obtained using all three loss terms (**Figure 5A**). To quantitatively compare the reconstruction quality of different models in the ablation study, the winning percentage of the pairwise similarity comparisons based on either objective or human judgment was used. The difference in winning percentage between the model optimized with all three loss terms and the model optimized with one loss term removed indicates the importance of the corresponding loss term. From **Figure 5B**, we can observe that the model trained with all three loss terms showed the highest winning percentage followed by the model where the loss in the image space is removed. The results demonstrate that the model trained with all three loss terms was preferred by the human raters.

Removing the loss in image space resulted in a moderate drop for both objective and subjective assessments (Pearson correlation 7.3% decrease, SSIM 13.8% decrease, and human judgment 18.5% decrease), but the difference in human judgement was not as pronounced as it was for the other two loss functions. Removing feature loss produced the highest drop in winning percentage for human judgment (36.9% decrease) and a moderate drop in Pearson correlation (5.6% decrease) and SSIM (11.1% decrease). This demonstrates the importance of optimization in high dimensional feature space, as it not only enhances the spatial details, but also makes the reconstruction more perceptually similar to its corresponding original stimulus image. Although removing adversarial dramatically reduced human judgement scores (30.0% decrease) and SSIM (41.8% decrease), it surprisingly showed improvement in Pearson correlation (10.9% increase). This suggests that optimizing adversarial loss forces the reconstruction to appear

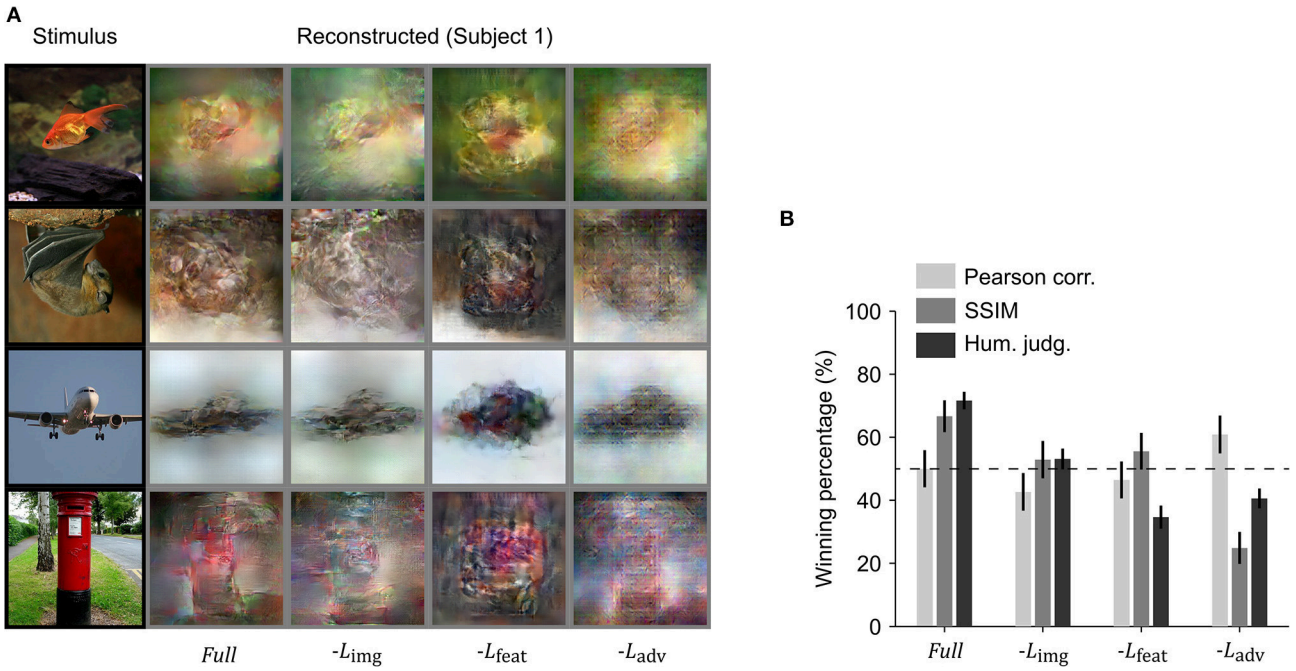


**FIGURE 5 |** Ablation of loss terms. **(A)** Reconstruction from brain activity (Subject 1) using reconstruction models with some loss components removed. The stimulus images (in black frames) are shown in the first column. The corresponding reconstructed images (in gray frames) obtained with different models are shown to the right of each stimulus image (from right to left, the model is: full reconstruction model (*Full*), with image loss removed ( $-L_{\text{img}}$ ), with feature loss removed ($-L_{\text{feat}}$), and with adversarial loss removed ($-L_{\text{adv}}$). **(B)** Reconstruction accuracy in terms of winning percentage of pairwise similarity comparison based on Pearson correlation, structural similarity index (SSIM) and human judgment (error bars, 95% CI across samples; three subjects pooled, $N = 150$, chance level, 50%). The winning percentage is the percentage of pairwise similarity comparison trials in which the reconstruction from one model was judged better than that of the other.

closer to a natural image distribution and preserve structural similarity but has a negative impact on preservation of the spatial details.

## DISCUSSION

Here, we have demonstrated that end-to-end training of a DNN model can directly map fMRI activity in the visual cortex to stimuli observed during perception, and thus reconstruct perceived images from fMRI data. The reconstructions of natural images were highly similar to the perceived stimuli in shape, and in some cases in color (**Figure 2**). Although trained only on natural images, the model generated accurate reconstructions of artificial shapes and alphabetical letters (**Figure 3**), thus showing generalizability that is similar to Shen et al. (2019). We also demonstrated that reconstruction quality improved as the number of training samples increased (**Figure 4**), and thus we may be able to further improve reconstruction accuracy with even more training samples.

We performed an ablation study by removing one loss function at a time to understand the importance of each loss term used for training the proposed model (**Figure 5**). The results showed that the model trained with all three loss terms achieved the best performance in terms of human judgement while the model trained without the adversarial loss showed the best performance in terms of Pearson correlation. The removal of loss in image space resulted in moderate changes in winning percentage calculated from behavioral experiments and both objective measures (Pearson correlation and SSIM). The removal of feature loss resulted in a drop in all the three types of winning percentage, although the drop in human ratings was more pronounced. Although removal of adversarial loss showed significant increase in winning percentage based on Pearson correlation, winning percentage based on human ratings and SSIM dropped significantly. This suggests that the addition of adversarial loss in the optimization process constrains the reconstructed images so that their distribution is closer to that of the training images (natural images). The increase in Pearson correlation winning percentage, however, suggests that adversarial loss has negative impact on preserving the spatial details of the reconstructed image. The results suggest that both the perceptual and adversarial losses are critical for our end-to-end deep image reconstruction model to achieve perceptually similar reconstructions.

Earlier studies on decoding stimuli in pixel space either searched for a match in the exemplar set (Naselaris et al., 2009; Nishimoto et al., 2011) or tried to reconstruct the stimulus (Miyawaki et al., 2008; Wen et al., 2016; Güçlütürk et al., 2017; Han et al., 2017; Seeliger et al., 2018; Shen et al., 2019). In the exemplar matching methods, visualization is limited to the samples in the exemplar set and hence these methods cannot be generalized to stimuli that are not included in the exemplar set. In contrast, reconstruction methods are more robust in generalizing to a new stimulus domain (**Figure 3**).

DNN-based reconstruction methods have typically avoided directly training a DNN model for reconstruction (Güçlütürk et al., 2017; Han et al., 2017; Seeliger et al., 2018; Shen et al., 2019). Instead, they have used decoded features as a proxy for hierarchical visual representations encoded in the fMRI activity that was used as the input to a reconstruction module. This method is effective since the decoded features can easily be plugged into known image reconstruction/generation methods. It is also thought to be efficient given the lack of large-scale diverse fMRI datasets (which contrasts with the large computer-vision datasets used for end-to-end training of vision tasks). The lack of large fMRI datasets makes learning a direct mapping from brain activity to stimulus space difficult without overfitting to the training dataset. Thus, developing a way to learn this direct mapping from limited numbers of training samples was the main motivation for this work.

A potential advantage of direct mapping is that it avoids information loss that occurs in the feature-decoding step. Even though the decoded features are correlated with the original image features, in Horikawa and Kamitani (2017) the maximum correlation coefficient on average was $< 0.5$. Thus, we argue that information in the decoded features is not all the visual information that can be decoded from the brain. Therefore, if enough training samples are available, direct mapping may help in preventing this information loss.

Our proposed method can easily be extended to other modalities such as text, sounds and video. This can be achieved by a suitable choice of generator, discriminator, and comparator modules for the corresponding modality. Further, our approach can be extended for reconstruction of multimodal data where a single generator module with multiple heads can generate reconstructions of multiple modalities simultaneously. Therefore, we believe an end-to-end approach has a wide potential for transforming the internal representations of the brain to meaningful visual and auditory contents for brain-machine interfaces.

## ETHICS STATEMENT

This study was carried out in accordance with the recommendations of the Ethics Committee of Advanced Telecommunications Research Institute International (ATR). The protocol was approved by the Ethics Committee of ATR. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

## AUTHOR CONTRIBUTIONS

YK directed the study. GS, KD, and KM developed the reconstruction methods. TH performed the experiments. GS performed the analyses. KD and YK wrote the paper.

## FUNDING

This research was supported by grants from the New Energy and Industrial Technology Development Organization (NEDO), JSPS KAKENHI Grant number JP15H05710, JP15H05920, JP26870935, and JP17K12771, and the ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan).

## ACKNOWLEDGMENTS

The authors thank Mohamed Abdelhack for valuable comments on the manuscript and Mitsuaki Tsukamoto for the help in computational environmental setting. We also thank Adam Phillips, Ph.D., from Edanz Group (www.edanzediting.com/ac) for editing a draft of this manuscript. This study was conducted using the MRI scanner and related facilities of Kokoro Research Center, Kyoto University. An earlier version of this manuscript was released as a preprint at BioRxiv (Shen et al., 2018).

## REFERENCES

Agrawal, P., Stansbury, D., Malik, J., and Gallant, J. L. (2014). *Pixels to Voxels: Modeling Visual Representation in the Human Brain. arXiv [Preprint]* arXiv:1407.5104. Available online at: https://arxiv.org/abs/1407.5104 (accessed October 23, 2017).

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Sci. Rep.* 6:27755. doi: 10.1038/srep27755

Cowen, A. S., Chun, M. M., and Kuhl, B. A. (2014). Neural portraits of perception: reconstructing face images from evoked brain activity. *NeuroImage* 94, 12–22. doi: 10.1016/j.neuroimage.2014.03.018

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. (2009). "ImageNet: a large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL), 248–255. doi: 10.1109/CVPR.2009.5206848

Dosovitskiy, A., and Brox, T. (2016a). "Inverting visual representations with convolutional networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 4829–4837.

Dosovitskiy, A., and Brox, T. (2016b). "Generating images with perceptual similarity metrics based on deep networks," in *Advances in Neural Information Processing Systems 29 (NIPS 2016)* (Barcelona), 658–666. Available online at: http://arxiv.org/abs/1602.02644 (accessed October 23, 2017).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems. (NIPS)*, Vol 27 (Montreal, QC), 2672–2680. Available online at: https://arxiv.org/abs/1406.2661 (accessed October 23, 2017).

Güçlü, U., and van Gerven, M. A. (2015a). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35, 10005–10014. doi: 10.1523/JNEUROSCI.5023-14.2015

Güçlü, U., and van Gerven, M. A. (2015b). Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *Neuroimage* 145(Pt B), 329–336. doi: 10.1016/j.neuroimage.2015.12.036

Güçlütürk, Y., Güçlü, U., Seeliger, K., Bosch, S., van Lier, R., and van Gerven, M. (2017). "Reconstructing perceived faces from brain activations with deep adversarial neural decoding," in *Advances in Neural Information Processing Systems 30 (NIPS)* (Long Beach, CA), 4249–4260.

Han, K., Wen, H., Shi, J., Lu, K.-H., Zhang, Y., and Liu, Z. (2017). Variational autoencoder: an unsupervised model for modeling and decoding fMRI activity in visual cortex. *bioRxiv [Preprint]*. doi: 10.1101/214247

He, K., Zhang, X., Ren, S., and Sun, J. (2015). "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *in Proceedings of the IEEE International Conference on Computer Vision* (Santiago), 1026–1034.

Horikawa, T., and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.* 8:15037. doi: 10.1038/ncomms15037

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). "Caffe: convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia MM '14*. (New York, NY: ACM), 675–678.

Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* 10, 1–29. doi: 10.1371/journal.pcbi.1003915

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* (San Diego, CA). Available online at: https://arxiv.org/abs/1412.6980 (accessed May 31, 2017).

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS),* Vol. 25 (Tahoe, CA), 1097–1105.

Lee, H., and Kuhl, B. A. (2016). Reconstructing perceived and retrieved faces from activity patterns in lateral parietal cortex. *J. Neurosci.* 36, 6069–6082. doi: 10.1523/JNEUROSCI.4286-15.2016

Mansimov, E., Parisotto, E., Ba, J. L., and Salakhutdinov, R. (2015). Generating images from captions with attention. *arXiv:*1511.02793

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M. A., Morito, Y., Tanabe, H. C., et al. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* 60, 915–929. doi: 10.1016/j.neuron.2008.11.004

Naselaris, T., Prenger, R. J., Kay, K. N., Oliver, M., and Gallant, J. L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron* 63, 902–915. doi: 10.1016/j.neuron.2009.09.006

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Curr. Biol.* 21, 1641–1646. doi: 10.1016/j.cub.2011.08.031

Seeliger, K., Güçlü, U., Ambrogioni, L., Güçlütürk, Y., and van Gerven, M. A. J. (2018). Generative adversarial networks for reconstructing natural images from brain activity. *Neuroimage* 181, 775–785. doi: 10.1016/j.neuroimage.2018.07.043

Shen, G., Dwivedi, K., Majima, K., Horikawa, T., and Kamitani, Y. (2018). End-to-end deep image reconstruction from human brain activity. *BioRxiv [Preprint]*. doi: 10.1101/272518

Shen, G., Horikawa, T., Majima, K., and Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLoS. Comput. Biol.* 15:e1006633. doi: 10.1371/journal.pcbi.1006633

Smilde, A. K., Kiers, H. A. L., Bijlsma, S., Rubingh, C. M., and van Erk, M. J. (2009). Matrix correlations for high-dimensional data: the modified RV-coefficient. *Bioinformatics* 25, 401–405. doi: 10.1093/bioinformatics/btn634

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process* 13, 600–612. doi: 10.1109/TIP.2003.819861

Wen, H., Shi, J., Zhang, Y., Lu, K. H., and Liu, Z. (2016). Neural encoding and decoding with deep learning for dynamic natural vision. *arXiv:*1608.03425

**Conflict of Interest Statement:** 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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 30 April 2019
doi: 10.3389/fncom.2019.00027

# Depth and the Uncertainty of Statistical Knowledge on Musical Creativity Fluctuate Over a Composer's Lifetime

*Tatsuya Daikoku**

*Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany*

Brain models music as a hierarchy of dynamical systems that encode probability distributions and complexity (i.e., entropy and uncertainty). Through musical experience over lifetime, a human is intrinsically motivated in optimizing the internalized probabilistic model for efficient information processing and the uncertainty resolution, which has been regarded as rewords. Human's behavior, however, appears to be not necessarily directing to efficiency but sometimes act inefficiently in order to explore a maximum rewards of uncertainty resolution. Previous studies suggest that the drive for novelty seeking behavior (high uncertain phenomenon) reflects human's curiosity, and that the curiosity rewards encourage humans to create and learn new regularities. That is to say, although brain generally minimizes uncertainty of music structure, we sometimes derive pleasure from music with uncertain structure due to curiosity for novelty seeking behavior by which we anticipate the resolution of uncertainty. Few studies, however, investigated how curiosity for uncertain and novelty seeking behavior modulates musical creativity. The present study investigated how the probabilistic model and the uncertainty in music fluctuate over a composer's lifetime (all of the 32 piano sonatas by Ludwig van Beethoven). In the late periods of the composer's lifetime, the transitional probabilities (TPs) of sequential patterns that ubiquitously appear in all of his music (familiar phrase) were decreased, whereas the uncertainties of the whole structure were increased. Furthermore, these findings were prominent in higher-, rather than lower-, order models of TP distribution. This may suggest that the higher-order probabilistic model is susceptible to experience and psychological phenomena over the composer's lifetime. The present study first suggested the fluctuation of uncertainty of musical structure over a composer's lifetime. It is suggested that human's curiosity for uncertain and novelty seeking behavior may modulate optimization and creativity in human's brain.

**Keywords: statistical learning, entropy, mutual information, information theory, Markovian, Bayesian, order, n-gram**

OPEN ACCESS

***Edited by:***
*MatjaŽ Perc,*
*University of Maribor, Slovenia*

***Reviewed by:***
*Haroldo Valentin Ribeiro,*
*Universidade Estadual de Maringá, Brazil*
*Rodrigo Laje,*
*Universidad Nacional de Quilmes (UNQ), Argentina*

***Correspondence:**
*Tatsuya Daikoku*
*daikoku@cbs.mpg.de*

***Received:*** *06 December 2018*
***Accepted:*** *11 April 2019*
***Published:*** *30 April 2019*

***Citation:***
*Daikoku T (2019) Depth and the Uncertainty of Statistical Knowledge on Musical Creativity Fluctuate Over a Composer's Lifetime. Front. Comput. Neurosci. 13:27. doi: 10.3389/fncom.2019.00027*

# INTRODUCTION

## Statistical Learning and Uncertainty in the Brain

The brain models external phenomena as a hierarchy of dynamical systems that encode probability distributions and complexity (i.e., entropy and uncertainty) over states in the world. Based on the internalized hidden model, it can predict a future state and optimize behavior and action to resolve the uncertainty (Friston, 2010). Within predictive-coding framework, this behavior mandates the suppression of prediction errors (prediction of content) and uncertainty (prediction of the context or precision of predictability and uncertainty) through updating internal model that generates predictions and the belief (Kanai et al., 2015). It has been considered that aesthetic appreciation of music can be modulated by these brain function: Through musical experience over lifetime, a human is intrinsically motivated in optimizing the internalized probabilistic model for efficient information processing and the uncertainty resolution, which has been regarded as rewords. For example, previous studies demonstrated that the precise prediction (Przysinda et al., 2017) and uncertainty perception (Hansen and Pearce, 2014) in music is stronger in proficient musicians than non-musicians.

This generative model could cover statistical learning (SL) theory of brain (Saffran et al., 1996; Cleeremans et al., 1998; Perruchet and Pacton, 2006). The SL is an implicit process by which the brain automatically calculates the statistical distribution of sequential phenomena based on Bayesian inference (Daikoku et al., 2012, 2014, 2016, 2017a, 2018; Yumoto and Daikoku, 2016; Daikoku and Yumoto, 2017), grasps the uncertainty (Hasson, 2017), predicts a future state based on the internal statistical model, and optimize action for achieving a given goal (Monroy et al., 2017a,b). By SL, generation of culture (Feher et al., 2016), individuality of creativity (Daikoku, 2018b) can be originated. Although brain tries to realize valuable behaviors at the lowest possible informational cost and uncertainty, it also seeks slightly suboptimal solution if the solution can be afforded at a significantly low uncertainty (Tishby and Polani, 2011). In other words, human's behavior appears to be not necessarily directing to efficiency but sometimes act unefficiently to explore a maximum rewards of uncertainty resolution. Previous studies suggest that the drive for novelty seeking behavior (high uncertain phenomenon) reflects human's curiosity and that the curiosity rewards encourage humans to create and learn new regularities (Kagan, 1972; Wittmann et al., 2008; Krebs et al., 2009; Schwartenbeck et al., 2013). Furthermore, a certain degree of uncertainty generates excitement and pleasure (Shen et al., 2015) because we explore a maximum curiosity rewards. Although brain generally minimizes prediction errors and uncertainty (Friston, 2010), we sometimes derive pleasure from prediction errors under conditions such as enjoying music listening due to curiosity and motivation for novelty seeking behavior by which we anticipate the resolution of uncertainty. Some literatures propose the hypothesis that the recurrent resolution of uncertainty activates reward networks that underwrite pleasure induced by listening to music (Koelsch, 2014; Salimpoor et al., 2015). It has been suggested that creativity can be explained as by-products of such intrinsic curiosity rewards (Schmidhuber, 2006). That is, human seems to look for some forms of optimality between uncertain and certain situations through action by which we are expected a maximum curiosity rewards, and hence our action gives rise to increasing as well as decreasing uncertainty. Recent studies imply that the curiosity rewards encourage humans to create and learn new regularities (Schmidhuber, 2006), and the fluctuations in uncertainty of predictions could contribute to aesthetic appreciation of art and music (Koelsch, 2014). Thus, it is hypothesized that human's intrinsic curiosity and motivation may modulate optimization and efficiency of prediction and action involved in SL. Recent computational studies on music suggest that, from early to late periods in the composer's lifetime, the transitional probabilities (TPs) of familiar phrase that ubiquitously appears in all of his music were gradually decreased (Daikoku, 2018d). This suggests that the statistical knowledge (Daikoku, 2018a) may be susceptible to long-term experience that modulates brain's probabilistic model (Hansen and Pearce, 2014). A neurophysiological study also suggested that sequences with higher entropy were learned based on higher-order TP, whereas those with lower entropy were learned based on lower-order TP (Daikoku et al., 2017b; Daikoku and Yumoto, 2019). Another study suggested that certain regions or networks perform specific computations of entropy (i.e., uncertainty), which are different from TP (i.e., prediction) of each content (Hasson, 2017). Thus, interaction between prediction and uncertainty in perceptive systems is an important topic to understand whole process of brain SL in both computational and neurophysiological areas (Daikoku, 2018c; Yumoto and Daikoku, 2018). Nevertheless, to our knowledge, few study examined relationships between SL, uncertainty and musical creativity and how curiosity for uncertain and novelty seeking behavior modulates musical creativity. The present study investigated how the probabilistic model and the uncertainty in music fluctuate over a composer's lifetime (all of the 32 piano sonatas by Ludwig van Beethoven).

## Computational Model

The computational model and simulation have been used to understand SL systems (e.g., Pearce and Wiggins, 2012; Rohrmeier and Rebuschat, 2012; Daikoku, 2018a; Wiggins, 2018; Daikoku and Yumoto, 2019). Although experimental approaches are necessary for understanding the real-world brain's function, the modeling approaches partially outperform experimental results under conditions that are impossible to replicate in an experimental approach (e.g., long-term statistical variation over the decades within a person and across cultures) and serves an important dual role in providing a quantitative account of observed empirical effects and in generating novel predictions to guide empirical research (e.g., Elman, 1990; Thiessen et al., 2013; Carreiras et al., 2014). Computational modeling can also express the relevant neural networks and neural hardware of sensory cortices (Turk-browne et al., 2009; Roux and Uhlhaas, 2014). For example, simple recurrent

network (SRN), which is classified as a neural network and was firstly devised by Elmer Elman (1990), learns sequential co-occurrence statistics by error-driven learning in which the gap between the prediction of a next input and the actual input drives changes to the weights on its internal connections. The SRN (Rogers and McClelland, 2004) and a modified SRN (Altmann, 1999; Dienes et al., 1999) implement a similarity space in which words referring to similar objects or actions were located more closely to one another than to words referring to dissimilar objects or actions. The neural network and deep learning such as Long-Short Term Memory (LSTM) (Hochreiter and Urgen Schmidhuber, 1997), on the other hand, is not intended to be a model of the relationship between human episodic and semantic memory although they proceed in this direction. Corpus-based approaches such as hyperspace analog to language (HAL) (Lund and Burgess, 1996), bound encoding of the aggregate language environment (BEAGLE) (Jones and Mewhort, 2007), Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997) are based on abstraction of episodic memory of input information and encoding in a multidimensional semantic space as semantic memory. Their models could also generate semantic similarity spaces in the similar way. For instance, when a verb of "drink" occurs, the models predict subsequent words that can be drunken. PARSER (Perruchet and Vinter, 1998), Competitive Chunker (Servan-Schreiber and Anderson, 1990), Information Dynamics of Music (IDyOM) (Pearce and Wiggins, 2012), Information Dynamics of Thinking (IDyOT) (Wiggins, 2018), and other *Markovian* models including the n-gram and nth-order Markov models (Daikoku, 2018b), can implement chunking hypotheses that learning is based on extracting, storing, and combining small chunks. Particularly, information-theoretical models including *Markovian* processes have been applied to neurophysiological studies of SL in human brain as well as computational simulation (Pearce et al., 2010a; Pearce and Wiggins, 2012; Daikoku et al., 2014, 2016, 2017a, 2018; Yumoto and Daikoku, 2016, 2018; Daikoku and Yumoto, 2017; Daikoku, 2018c). These neurophysiological experiments showed consistent evidence: neural activities for stimuli with high information content (i.e., low probability) are larger than those with low information content (i.e., high probability). Furthermore, these SL effects were larger when humans are exposed stimulus sequence with less information entropy (uncertainty), compared with when they are exposed stimulus sequence with high information entropy (Daikoku et al., 2017c). The mutual information of information theory, which has been assumed as the reduction of uncertainty afforded by observations (see section Mutual Information of nth-order SL model), is also correlated with neuronal activity in limbic cortex (Harrison et al., 2006). This neural phenomenon is in agreement with a *Bayesian* hypothesis in theoretical neurobiology that the brain encodes probabilities (beliefs) about the causes of sensory data, and that these beliefs are updated in response to new sensory evidence based on Bayesian inference (Kersten et al., 2004; Knill and Pouget, 2004; Doya et al., 2007; Friston, 2010; O'Reilly et al., 2012; Parr and Friston, 2018; Parr et al., 2018). Formulations of self-organization (Karl Friston, 2013; Kirchhoff et al., 2018) and brain connectivity (Parr and Friston, 2018) are also expressed using an information-theoretical concept called Markov blankets (Pearl, 1988). The blanket of a state is the only knowledge necessary to predict the behavior of that state and the adjacent state. If we know everything within a blanket, knowledge about things outside the blanket becomes uninformative about things inside the blanket. For example, Parr and Friston (2018) hypothesized that a neuronal population reflecting a given variable only need receive connections from those populations representing its blanket and explained this notion from perception, planning, attention, and movement. The Markov blanket may also represent in part chunk formation although it's not sufficient. Markov decision process (MDP) (Schwartenbeck et al., 2013; Karl Friston et al., 2014, 2015; Pezzulo et al., 2015), which has often been used for reinforcement learning in AI and robotics, extends the simple perceptive process by adding active process (controlling predictability by choice, called "*policy*") and "*rewards*" (giving motivation). The IDyOM is also an extension of Markov model to precisely modeling SL of musical sequence combining several concomitant information such as pitch, duration, onset, scale degree, and so on. The SL based on IDyOM could also be reflected in neurophysiological responses within the predictive-coding framework (Pearce et al., 2010b). The IDyOT also takes advantage of information theory to represent domain-general SL mechanisms that cover both language and music (Wiggins, 2018). Particularly, this model implements semantic and episodic memory systems, and captures hierarchical SL process from lower- to higher-level using boundary entropy: spectrum of auditory sequence is chunked into phonemes, then morphemes, then words (Wiggins, in press). In summary, information-theoretical models including Markovian processes can capture a variety of neurophysiological phenomena on SL such as prediction, uncertainty, a part of chunk formation, and policy of action, across domains, and modality.

A previous study reported that SL effects based on TPs occurs action as well as perception (Monroy et al., 2017c). This suggests that SL also contributes production of sequences. In other words, from psychological perspective, TP distribution sampled from music based on Markov models may also refer to the characteristics of a composer's statistical knowledge: a high-probability transition in music may be one that a composer is more likely to predict and choose based on the latest n states, compared to a low-probability transition. Thus, the Markov model is used in the interdisciplinary realms of neuroscience, behavioral science, engineering, and informatics. A computational study using nth-order Markov or n-gram models suggested that time-course variations of statistics in music reflect time-course variations of a composer's statistical knowledge (Daikoku, 2018d). Neurophysiological studies also suggested that time-course variations of statistics of auditory sequence modulate SL effects (Daikoku et al., 2017c) and that the SL effects of sequences with higher entropy were lower than those with lower entropy, even when TP itself is same between these two sequences (Daikoku et al., 2017c). These studies suggest that time-course variations of TPs and entropy may partially be able to predict the SL model in human's brain.

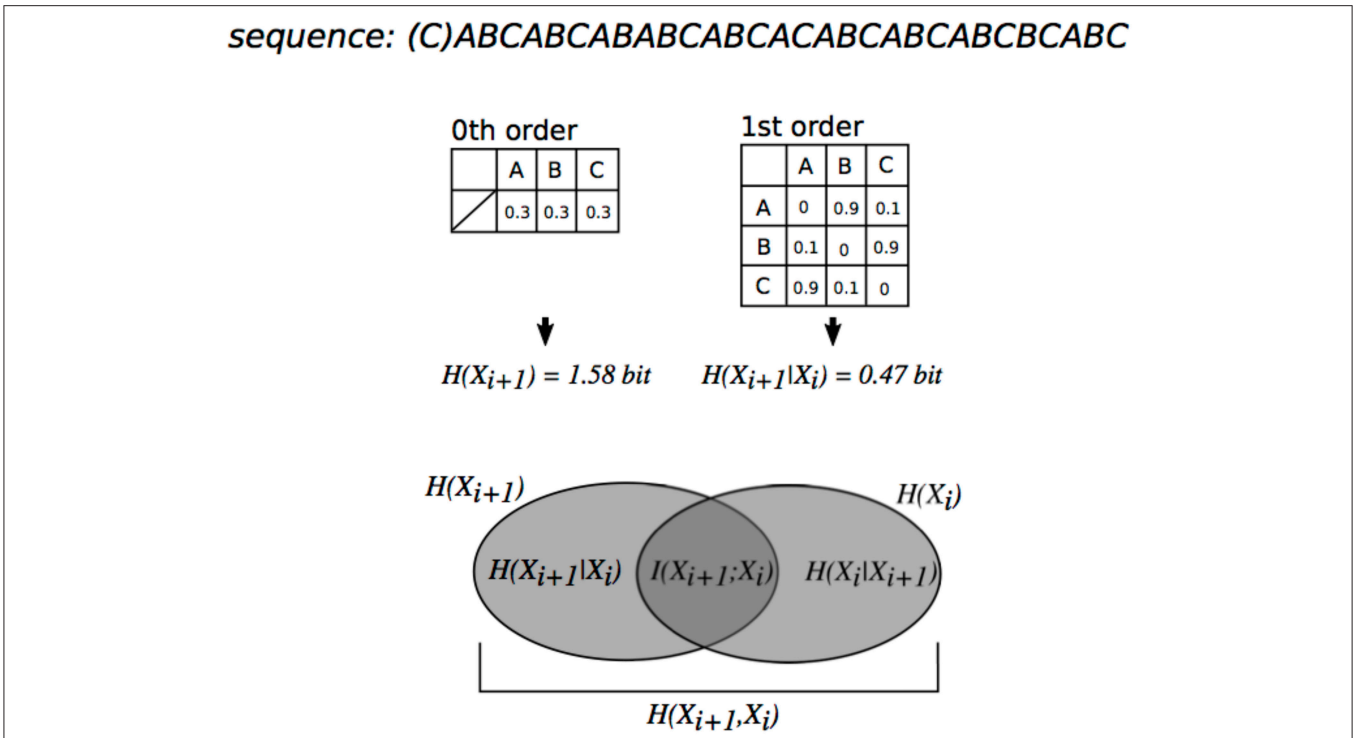


**FIGURE 1 |** Relationship between order of transitional probabilities, conditional entropy, and mutual information illustrated using a Venn diagram. The degree of dependence on $X_i$ for $X_{i+1}$ is measured by mutual information [mutual information ($I(X;Y)$] = entropy ($H(X_{i+1})$)–conditional entropy ($H(X_{i+1}|X_i)$)]. The mutual information of sequences in this figure is more than 0. Thus, each event $X_{i+1}$ in the sequence is dependent on a preceding event $X_i$.

## Mathematical Interpretation of Brain's Statistical Learning Based on Information Theory

### Nth-Order Transitional Probability

According to SL theory, the brain automatically computes both lower- and higher-order TPs in sequences (Furl et al., 2011; Yumoto and Daikoku, 2016, 2018, grasps uncertainty/entropy in the whole sequences Hasson, 2017, and predicts a future state based on the internalized statistical model Friston, 2010. The TP is a conditional probability of an event B given that the latest event A has occurred, written as P(B|A). The nth-order TP distributions sampled from sequential information such as music and language can be expressed by nth-order Markov models. The nth-order Markov model is based on the conditional probability of an event $e_{n+1}$, given the preceding $n$ events based on Bayes' theorem ($P(e_{n+1}|e_n)$). From psychological perspective, the conditional probability ($P(e_{n+1}|e_n)$) can be interpreted as positing that the brain predicts a subsequent event $e_{n+1}$ based on the preceding events $e_n$ in a sequence. In other words, learners expect the event with the highest TP based on the latest n states, whereas they are likely to be surprised by an event with lower TP. Furthermore, TPs are often translated as information contents (IC) ($-log_2 1/P(e_{n+1}|e_n)$) of information theory (Pearce and Wiggins, 2006). The lower IC (i.e., higher TPs) means higher predictabilities and smaller surprising, whereas the higher IC (i.e., lower TPs) means lower predictabilities and larger surprising. In the end, a tone with lower IC may be one that a composer is more likely to predict and choose as a next tone, compared to tones with higher IC. IC can be used in computational studies of music to discuss psychological phenomena involved in prediction and SL.

### Entropy and Uncertainty

Entropy as well as TP of each event is used to understand predictability of a sequence (Pearce, 2005). Entropy (e.g., see **Figure 1**) is calculated from probability distribution, interpreted as uncertainty (Friston, 2010), and used to evaluate neurophysiological effects of uncertainty in SL (Harrison et al., 2006) and curiosity (Loewenstein, 1994). A previous study reported that neural systems of uncertainty perception were partially independent of those of prediction of each content (Hasson, 2017). Some articles, however, suggest that uncertainty modulates predictability of each content in SL (Daikoku et al., 2017c). Furthermore, uncertainty of auditory and visual statistics is coded by modality-general, as well as modality-specific, neural systems (Strange et al., 2005; Nastase et al., 2014). This suggests that the neural basis that codes uncertainty as well as prediction, is a domain-general system. Thus, there seems to be neural and psychological interactions of perceptions between prediction and uncertainty.

### Mutual Information of nth-order SL Model

Mutual information (MI) and pointwise Mutual information (PMI) are a measure of the mutual dependence between the two variables. The PMI refers to each event in sequence (local

dependence), whereas MI refers to the average of all events in the sequence (global dependence). In the framework of SL based on TPs ($P(e_{n+1}|e_n)$), MI explains how an event $e_{n+1}$ is dependent on the preceding event $e_n$. Thus, MI is a key to understanding order of SL. For instance, conventional oddball sequence, which consists of a frequent stimulus with high probability of appearance and a deviant stimulus with low probability of appearance, has weak dependence between two adjacent events ($e_n$, $e_{n+1}$) and shows low MI, because event $e_{n+1}$ appears independently of preceding events $e_n$. In contrast, SL sequence based on TPs, but not probabilities of appearance, has strong dependence between two adjacent events and shows larger MI. For example, typical SL paradigm that consists of concatenation of pseudo-words with three stimuli has large MI until 2nd-order Markov or tri-gram models [i.e., $P(C|AB)$)], whereas it has low MI from 3rd-oder Markov or four-gram models [i.e., $P(D|ABC)$)]. Thus, MI is sometimes used to evaluate hierarchical SL in both neurophysiological and computational studies (Harrison et al., 2006; Pearce et al., 2010b).

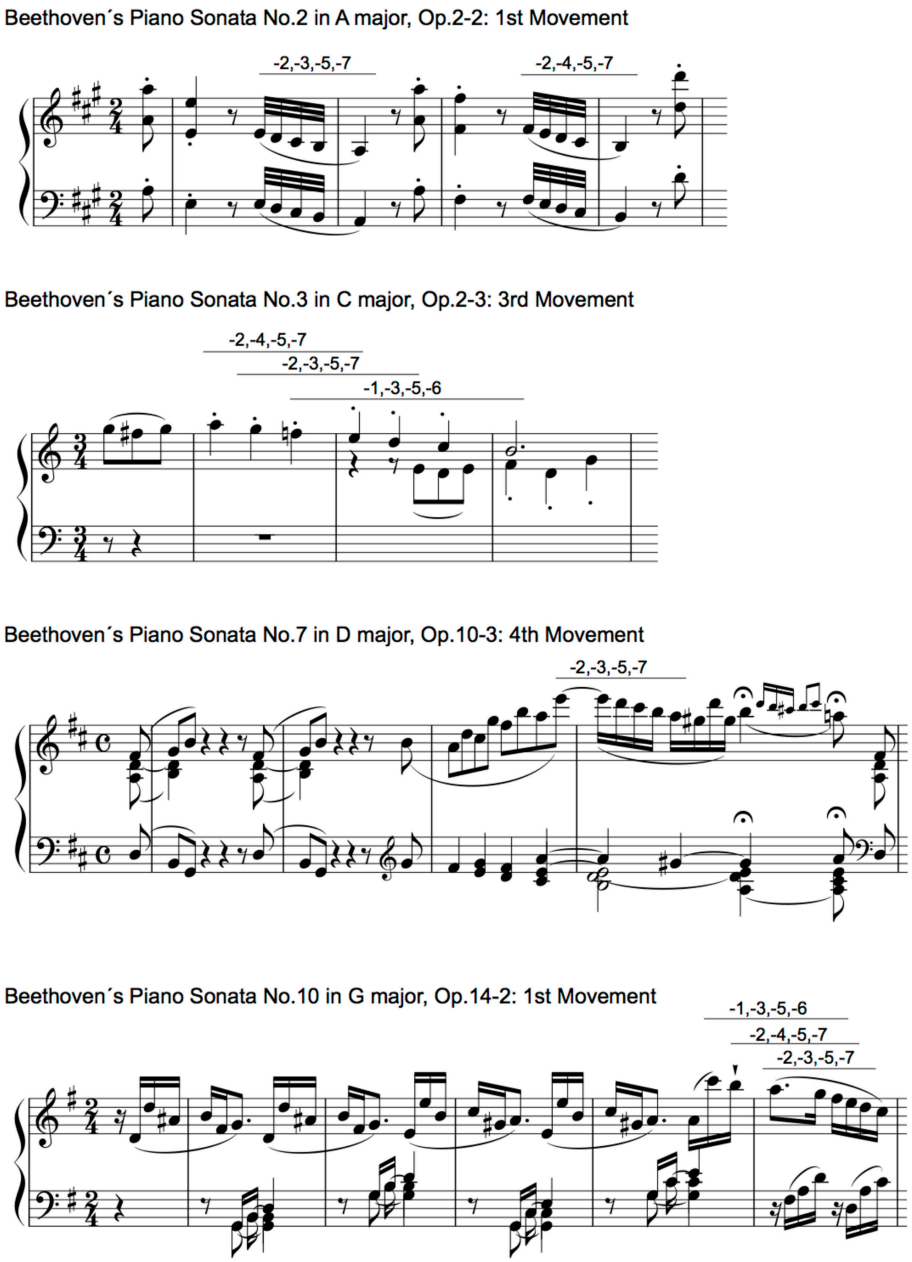


**FIGURE 2 |** Representative sequences of [0,−2,−3,−5,−7], [0,−2,−4,−5,−7], and [0,−1,−3,−5,−6] in Beethoven's piano sonatas in the early period.

In this section, the three types of information-theoretical evaluation of SL models (i.e., IC, entropy, and MI) were explained from psychological aspects. In sum, (1) IC reflects surprising/predictability. A tone with lower IC (i.e., higher TPs) may be one that a composer is more likely to predict and choose as a next tone, compared to tones with higher IC. (2) Entropy reflects uncertainty of whole sequences. (3) MI reflects hierarchy of statistics and is interpreted as dependence of preceding sequential events in SL. Using them, the present study investigated how prediction, uncertainty, and the depth of implicit knowledge in music vary over a composer's lifetime (all of the 32 piano sonatas by Ludwig van Beethoven).

## Ludwig Van Beethoven's Piano Sonata

The German composer and pianist Ludwig van Beethoven (1770–1827) remains one of the most famous and influential of all composers. It is believed that his music strongly expresses the psychological variations and visions of his life (Sullivan, 1927; Boucourechliev, 1963). Beethoven's compositional career is often divided into the early (around 1802), middle (around 1802–1814), and late periods (from about 1814) (Dahlhaus, 1991; Adorno-Wiesengrund, 1993). It is generally thought that his works in the early period were strongly influenced by his predecessors in classicism, such as Wolfgang Amadeus Mozart (1756–1792) and Franz Joseph Haydn (1732–1809), whereas his works in the late period show his personal character and experience (Sullivan, 1927) and accompanying intellectual depth and personal expression (Dahlhaus, 1991; Adorno-Wiesengrund, 1993). Thus, his psychological variations on thinking and experience may form the statistical characteristics of his music that may reflect a composer's statistical knowledge (Johnson et al., 1985). It is believed that he always explored new directions of musical composition and gradually expanded his scope of music over his lifetime (Dahlhaus, 1991; Adorno-Wiesengrund, 1993). Using Beethoven's piano sonatas over his lifetime, the present study examined time-course variations of three types of statistics in music: TPs (ICs) of sequential patterns that appear in all 32 sonatas, entropy of whole TP distribution, and the MI. It was hypothesized that, because of his exploration of new directions in musical composition over his lifetime, TP of phrases that frequently appear in the early period (i.e., sequences with high TP) might decrease in the late period (i.e., decreasing TP), whereas entropy might increase in the late period. It would be very interesting if the psychological variations in which Beethoven explored new directions and gradually expanded his scope of music over his lifetime were reflected in the SL models of his music.

# METHODS

The Piano Sonata with all of its movements by Ludwig van Beethoven (No.1 in F minor, Op.2-1 to No.32 in C minor, Op.111, composed 1795–1822) was used in the present study. Using a scorewriter (Finale version 25, MI Seven Japan, Inc.), electronic scoring data of the sequences of highest pitch were extracted from the XML files. The highest pitches were chosen based on the following definitions: the highest pitches that can be played at a given point in time, the pitches with slurs can be counted as one, and the grace notes were excluded. Although melody is sometimes not highest pitches e.g., bass melodies), the present study only analyzed the highest pitch because different melodies could concurrently appear in some titles of music, and melody is often played in highest pitches. Using all the sequences of highest pitches in a movement of a Sonata, sequential patterns based on uni- to four-grams were extracted. For each type of the sequential patterns, all pitches were numbered so that the first pitch was 0 in each transition, and an increase or decrease in a semitone was 1 and −1 based on the first pitch, respectively. The representative examples were shown in **Figure 2**. This revealed interval patterns but not pitch pattern. This procedure was employed to eliminate the effects of the change of key on transitional patterns. The interpretation of the change of key depends on musicians, and it is difficult to define in an objective manner. Thus, the results in the present study may represent a variation of statistics associated with relative pitch rather than absolute pitch. Then, the TPs of the sequential patterns were calculated based on 0th- to 3rd-order Markov chains. Furthermore, TPs of all the movements in each piece of sonata (No.1 to No.32) were weighted averaged: an average in which probability of each phrase is multiplied by a weight before summing to a single average value. That is weightings are the equivalent of having that many like items with the same value involved in the average. The *n*th-order Markov chain is the conditional probability of an event $e_{n+1}$, given the preceding event $e_n$:

$$P\left(e_{n+1}|e_n\right) = \frac{P(e_{n+1} \cap e_n)}{P(e_n)} \quad (1)$$

The ICs ($I$) and conditional entropy ($H$) in the nth-order TP distribution (hereafter, Markov entropy) were calculated using TPs in the framework of information theory:

$$I\left(e_{n+1}|e_n\right) = -\log_2 P\left(e_{n+1}|e_n\right) \ (bit) \quad (2)$$

$$H\left(B|A\right) = -\sum\nolimits_i \sum_j P(ai)P\left(bj|ai\right)\log_2 P\left(bj|ai\right) \ (bit) \quad (3)$$

where *P(*bj|ai*)* is a conditional probability of sequence *"ai bj."* Then, MI (*I(X;Y)*) were calculated in 1st-, 2nd-, and 3rd-order Markov models. MI is an information theoretic measure of dependency between two variables. From entropy values, the MI can also be expressed as

$$\begin{aligned} I\left(X;\,Y\right) &= \sum_{x,y} p\left(x,y\right)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \\ &= \sum_{x,y} p\left(x,y\right)\log\left(\frac{p(x,y)}{p(x)}\right) - \sum_{x,y} p\left(x,y\right)\log p(y) \\ &= \sum_{x,y} p\left(x\right)p(y|x)\log p(y|x) - \sum_{x,y}\log p\left(y\right)p(x,y) \\ &= \sum_{x} p\left(x\right)\left(\sum_{y} p\left(y|x\right)\log p\left(y|x\right)\right) \end{aligned}$$

$$
\begin{aligned}
&- \sum_{y} \log p(y)\left(\sum_{x} p(x,y)\right) \\
&= -\sum_{x} p\,(x)\, H\,(Y|X=x) - \sum_{y} p\,(y) \log p(y) \\
&= -H\,(Y|X) + H\,(Y) \\
&= H\,(Y) - H\,(Y|X)\ \ (bit) \qquad (4)
\end{aligned}
$$

where p(x,y) is the joint probability function of X and Y, p(x), and p(y) are the marginal probability distribution functions of X and Y respectively, H(X) and H(Y) are the marginal entropies, H(X|Y) and H(Y|X) are the conditional entropies, and H(X,Y) is the joint entropy of X and Y (**Figure 1**) (Daikoku, 2018a). Based on psychological and information-theoretical concepts, the Equation (4) can be regarded that the amount of entropy (uncertainty) remaining about Y after X is known. That is, the MI is corresponding to reduction in entropy (uncertainty). In each order of Markov models, the sequential patterns that ubiquitously appear in all 32 sonatas (hereafter, familiar phrase) were extracted. Then, TPs of the familiar phrases were averaged (0th: 20 phrases, 1st: 37 phrases, 2nd: 12 phrases, and 3rd: 3 phrases) (**Table 1**). The 32 sonatas were divided based on the well-known 3 periods: early (No.1 to 12, No19, and No20), middle (No.13 to 18 and No. 21 to 27), and late (No. 28 to 32). Then, I conducted analysis of variances (ANOVAs) with a within-subject factor order (0th vs. 1st vs. 2nd vs. 3rd) and a between-subjects factor composition period of the sonatas (early vs. middle vs. late) for the TPs of familiar phrases and entropy of whole music, and an ANOVA with a within-subject factor order (1st vs. 2nd vs. 3rd) and a between-subjects factor composition period (early vs. middle vs. late) for the mutual information. When we detected significant effects, Bonferroni-corrected *post-hoc* tests were conducted for further analysis. Then, in each order of Markov models, the TPs of familiar phrase and the uncertainty of whole music were compared by Pearson's correlation analysis. Statistical significance levels were set at $p = 0.05$ for all analyses.

**TABLE 1 |** Sequential patterns that appear in all 32 sonatas (i.e., phrases) in each order of Markov models.

| Order | Sequential patterns |
|---|---|
| 0th | [−2], [−1], [1], [0], [2], [−3], [3], [5],[−4],[4],[−5], [12], [−7], [7], [9], [−12], [8], [−6], [6], [−9] |
| 1st | [1,1], [−1,2], [−1,−1], [1,−2], [0,1], [3,3], [0,2], [1,−4], [0,3], [0,5], [1,−1], [−2,2], [−2,5] |
| 2nd | [−2,−4,−5], [0,0,0], [−1,−3,−5], [2,4,5], [−2,−3,−5], [1,3,5], [2,3,5], [2,0,−1], [−2,−3,−2], [−1,0,2], [1,3,1], [−1,0,−1] |
| 3rd | [−2,−3,−5,−7], [−2,−4,−5,−7], [−1,−3,−5,−6] |

# RESULTS

## TPs of Familiar Phrases

In the TPs of familiar phrases, An ANOVA with a within-subject factor order (0th vs. 1st vs. 2nd vs. 3rd) and a between-subjects factor composition period (early vs. middle vs. late) was conducted. As a result, the main effect of period was significant [$F(2, 29) = 6.02$, $p = 0.007$, partial$\eta^2 = 0.29$, early > late, $p = 0.005$; middle > late, $p = 0.032$] (**Figure 3D**). The period-order interaction was also significant [$F(6) = 6.82$, $p < 0.001$, partial$\eta^2 = 0.32$] (**Figure 3A**). The 3rd-order TPs in late period were significantly lower than those in early ($p < 0.001$) and middle periods ($p = 0.003$). That is to say, the 3rd-order TPs of familiar phrases in the late period only decrease during lifetime. The main effect of order was significant [$F(3, 87) = 1108.35$, $p < 0.001$, partial$\eta^2 = 0.98$]. The 0th-order TPs were significantly lower than the 1st-, 2nd-, and 3rd-order TPs (all: $p < 0.001$). The 1st-order TPs were significantly lower than the 2nd-, 3rd-order TPs (all: $p < 0.001$). The 2nd-order TPs were significantly lower than the 3rd-order TPs ($p = 0.007$).

## Entropy and Uncertainty

In entropy of whole TP distribution, ANOVA with a within-subject factor order and a between-subjects factor composition period of sonatas was performed. The main effect of period was significant [$F_{(2,29)} = 7.58$, $p = 0.002$, partial$\eta^2 = 0.34$, early < middle, $p = 0.005$; early < late, $p = 0.002$] (**Figure 3E**). The period-order interaction was also significant [$F(6) = 6.68$, $p < 0.001$, partial$\eta^2 = 0.32$] (**Figure 3B**). The entropies of 0th-order TPs in late period were significantly lower than those in the middle periods ($p = 0.034$). The entropies of 1st-order TPs in late period were significantly higher than those in the early ($p = 0.004$) and middle periods ($p = 0.014$). The entropies of 2nd-order TPs in late period were significantly higher than those in the early ($p = 0.001$) and middle periods ($p < 0.001$). The entropies of 3rd-order TPs in late period were significantly higher than those in the middle periods ($p = 0.017$). The main effect of order was significant [$F(1.73, 50.30) = 2329.84$, $p < 0.001$, partial$\eta^2 = 0.99$]. The entropies of 0th-order TPs were significantly lower than the 1st-, 2nd-, and 3rd-order TPs (all: $p < 0.001$). The entropies of 1st-order TPs were significantly lower than the 2nd-, 3rd-order TPs (all: $p < 0.001$). The entropies of 2nd-order TPs were significantly lower than the 3rd-order TPs ($p = 0.007$).

## Hierarchy of Statistics: Mutual Information

In the mutual information, an ANOVA with a within-subject factor order and a between-subjects factor composition period was conducted. The main effect of period was significant [$F(2, 29) = 9.08$, $p = 0.001$, partial$\eta^2 = 0.39$, early > late, $p = 0.020$; middle > late, $p = 0.001$] (**Figure 3F**). The period-order interaction was also significant [$F_{(4)} = 2.80$, $p = 0.034$, partial$\eta^2 = 0.16$] (**Figure 3C**). The mutual information of 1st- and 2nd-order TPs in late period were significantly lower than those in the early (1st: $p = 0.004$; 2nd: $p = 0.012$) and middle periods (1st: $p < 0.001$; 2nd: $p < 0.001$). The mutual information of 3rd-order TPs in late period was significantly higher than those in the middle periods ($p = 0.008$). The main effect of order was significant [$F_{(1.32,38.16)} = 2350.56$, $p < 0.001$, partial$\eta^2 = 0.99$]. The mutual information of 1st-order TPs were significantly lower than the 2nd-, 3rd-order TPs (all: $p < 0.001$). The 2nd-order TPs were significantly lower than the 3rd-order TPs ($p < 0.001$).

# DISCUSSION

Brain encodes probability distributions and the entropy/uncertainty of musical information (Koelsch et al.,

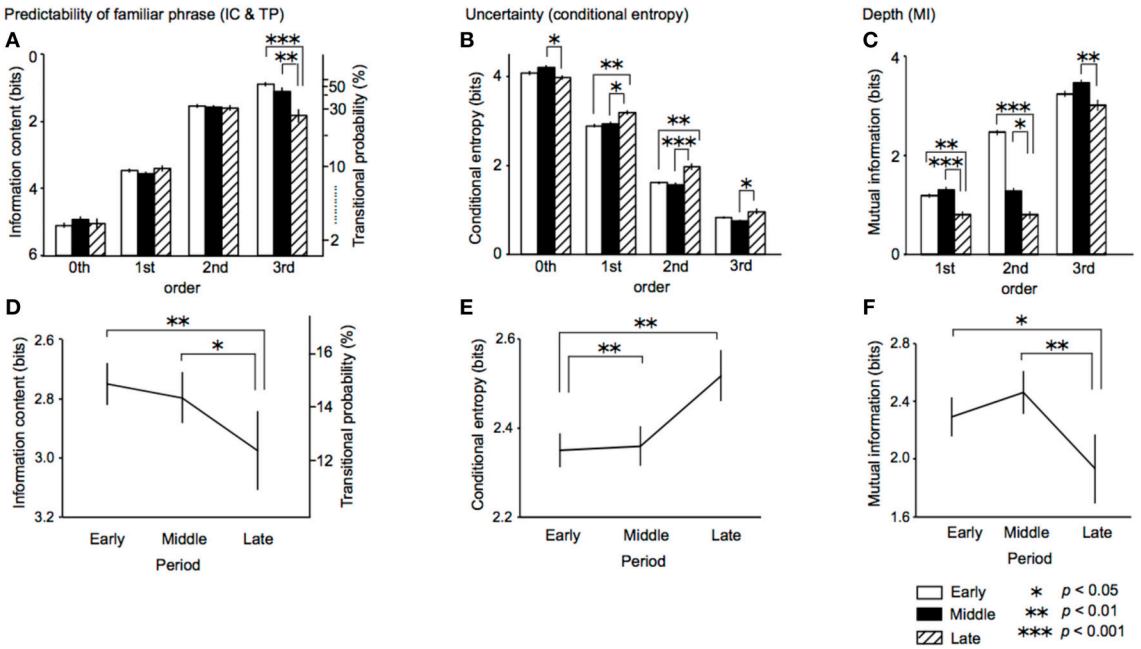


**FIGURE 3 |** The period-order interactions **(A–C)** and main effects of period **(D–F)** in the ANOVA of ICs and TPs of familiar phrases, conditional entropy of TP distribution, and the depth of implicit knowledge (MI) in the early (opened bars), middle (filled bars), and late (dashed bars) periods. IC, information content; TP, transitional probability; MI, mutual information.

2018) and mandates the suppression of prediction errors and uncertainty by updating the internal probabilistic model of music that generates predictions and the belief (Kanai et al., 2015). In other words, through musical experience over lifetime, a human generally tries to optimize the internalized probabilistic model for efficient information processing and the uncertainty resolution, which has been regarded as rewords. On the other hand, to explore the maximum rewards of uncertainty resolution, human's behavior appears to be not necessarily directing to efficiency, but sometimes be drove by unefficient, uncertain, and novelty information, which is thought as curiosity (Kagan, 1972; Wittmann et al., 2008; Krebs et al., 2009; Schwartenbeck et al., 2013). Thus, although brain typically minimizes uncertainty of music structure, we sometimes derive pleasure from music with uncertain structure due to curiosity for novelty-seeking behavior by which we anticipate further rewords by uncertainty resolution. The present study, using all the Beethoven's piano sonatas over his lifetime, examined how the probabilistic model and the uncertainty in music fluctuate over a composer's lifetime. The transitional probability and information content (TP), information content (IC), conditional entropy, and mutual information (MI) can be calculated based on *n*th-order Markov models. Based on psychological and neurophysiological studies on SL (Harrison et al., 2006; Pearce et al., 2010b; de Zubicaray et al., 2013; Daikoku et al., 2015; Monroy et al., 2017c), these three information can be translated to psychological indices: a tone with lower IC (i.e., higher TPs) may be one that a composer is more likely to choose as a next tone, compared to tones with higher IC, whereas the entropy and MI are interpreted as uncertainty and the order of the SL, respectively. It was hypothesized that probability, uncertainty, and the order of SL models is fluctuated over Beethoven's lifetime. If so, it may suggest that his curiosity for uncertain and novelty seeking behavior modulate optimization and creativity in human's brain.

The TPs of familiar phrase (i.e., sequences that appear in all 32 sonatas) were decreased in the late periods (**Figure 3D**), whereas the entropies in music were increased in the late periods (**Figure 3E**). In other words, there was no significant difference between early and middle periods, while there was significant difference between middle and late periods. Particularly, the 3rd-order TPs in the late period decrease during lifetime (**Figure 3A**). According to musicological studies, his works in the early period were strongly influenced by his predecessors in classicism whereas his works in the late period show his personal character and experience (Sullivan, 1927). It is believed that he always explored new directions of musical composition and gradually expanded his scope of music over his lifetime (Dahlhaus, 1991; Adorno-Wiesengrund, 1993). The findings of the present study may suggest the hypothesis that the psychological variations over lifetime are reflected in the statistical structure in music. The decreasing of the TPs of familiar phrase and increasing of the entropies may imply that, in the late period, he tried novel composition strategies in which he avoided familiar sequences in the early period, and tried various transitional patterns by which nth-order TPs are broadly distributed. The previous study detected time-course variation of predictability of familiar phrases over his lifetime (Daikoku, 2018d). The present study,

furthermore, suggested that there seems to be interactions between prediction and uncertainty.

The decreasing of TPs of familiar phrase over Beethoven's lifetime was obvious in the higher-, but not lower-, order models (**Figure 3A**). This may suggest that a higher-, rather than lower-, order statistical structure reflects specific statistical knowledge that is susceptible to experience and novelty seeking behavior. The entropy (i.e., uncertainty) of TP distribution may also support the hypothesis (**Figure 3B**). The entropies of higher-order (1st to 3rd), but not lower-order (0th) models in late period increased compared with those in early period. Furthermore, MI in late periods was lower compared with those early and middle periods (**Figure 3F**). This suggests that, in the late period, each event of tone hardly depends on preceding successive events of tones. Typical Western-classical music has strict syntactic rules based on music theory. Therefore, a forthcoming tone can partially be predicted from preceding successive tones based on the rules. According to previous studies, syntax of musical sequences is partially expressed by conditional probabilities (Rohrmeier and Cross, 2008), although it is not sufficient to account for all of the music syntax. The findings on MI in this study may suggest that, in the late period, the composer avoided a tone that can easily be predicted based on typical transition rules involved in music syntax.

In sum, the present study detected time-course variation of predictability of familiar phrases, uncertainty of whole music, and the depth of SL in music that were composed over Beethoven's lifetime. According to corpus studies, the historical characteristics of music can be extracted based on the era (e.g., Albrecht and Huron, 2014; Gjerdingen, 2014; White, 2014). This indicates that strategies of composition and musical knowledge depend on the era. The present study also suggests that the characteristics of music can be extracted based on the periods within a composer's lifetime. In addition, the higher-order hierarchical structure showed larger time-course variations of both predictability of familiar phrases and uncertainty of whole music. From the psychological perspective, it would be interesting if the higher- (i.e., deep), rather than lower-order statistical knowledge was susceptible to experience and novelty seeking behavior. The present study also suggested that there are interactions between prediction and uncertainty. It is of note, however, that the present study did not directly investigate the composer's statistical knowledge of music, as only the statistics of musical scores were analyzed. Furthermore, the present study only analyzed one composer, therefore could not discuss universal phenomenon on SL. This suggests that there may be other possible explanations for the findings of the present study. For instance, it might have been Beethoven's plan to compose the sonatas from familiar and lower entropy to unfamiliar and larger entropy based on the statistical structure of music. Future study should investigate SL of music from many composers using interdisciplinary approaches in parallel.

## CONCLUSION

The present study investigated how predictability of familiar phrases that was used in all of music, uncertainty of whole structure, and the order of the probabilistic models in music fluctuates over a composer's lifetime, and discussed the results from psychological perspective within SL framework. The results suggest that the higher-, rather than lower-order statistical knowledge may be susceptible to experience and novelty seeking behavior. The present study also suggested that there might be interactions between prediction and uncertainty. The present study first suggested that uncertainty may be increased in a composer's lifetime, and that the higher-order probabilistic model may be susceptible to experience and novelty seeking behavior over the composer's lifetime. It is suggested that human's curiosity for uncertain and novelty seeking behavior may modulate creativity in human's brain, and that the fluctuations of uncertainty could reflect aesthetic appreciation of music. To more understand brain's predictive function, future study is needed to examine relationships between prediction of familiar phrases and uncertainty perception, using both modeling and experimental approaches in parallel.

## AUTHOR CONTRIBUTIONS

The methodology of the present study and were considered by the author. The author analyzed all of the data and prepared the figures, and wrote the manuscript text.

## FUNDING

This work was supported by Grant-in-Aid for Nakayama Foundation for Human Science. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## REFERENCES

Adorno-Wiesengrund, T. (1993). *Beethoven: The Philosophy of Music Fragments and Texts*. Cambridge: Polity Press.

Albrecht, J. D., and Huron, D. (2014). A statistical approach to tracing the historical development of major and minor pitch distributions, 1400–1750. *Music Percept. Interdiscipl. J.* 31, 223–243. doi: 10.1525/mp.2014.31.3.223

Altmann, G. T. (1999). Rule learning by seven-month-old infants and neural networks. *Science* 284, 875a. doi: 10.1126/science.284.5416.875a

Boucourechliev, A. (1963). *Beethoven (in French)*. Seuil.

Carreiras, M., Armstrong, B. C., Perea, M., and Frost, R. (2014). The what, when, where, and how of visual word recognition. *Trends Cogn. Sci.* 18, 90–98. doi: 10.1016/j.tics.2013.11.005

Cleeremans, A., Destrebecqz, A., and Boyer, M. (1998). Implicit learning: news from the front. *Trends Cogn. Sci.* 2, 406–416. doi: 10.1016/S1364-6613(98)01232-7

Dahlhaus, C. (1991). *Ludwig van Beethoven: Approaches to His Music*. Translated by Mary Whittall. New York, NY: Oxford University Press.

Daikoku, T. (2018a). Entropy, uncertainty, and the depth of implicit knowledge on musical creativity : computational study of improvisation in melody and rhythm. *Front. Comput. Neurosci.* 12:97. doi: 10.3389/fncom.2018.00097

Daikoku, T. (2018b). Musical creativity and depth of implicit knowledge: spectral and temporal individualities in improvisation. *Front. Comput. Neurosci.* 12:89. doi: 10.3389/fncom.2018.00089

Daikoku, T. (2018c). Neurophysiological markers of statistical learning in music and language: hierarchy, entropy, and uncertainty. *Brain Sci.* 8:114. doi: 10.3390/brainsci8060114

Daikoku, T. (2018d). Time-course variation of statistics embedded in music: corpus study on implicit learning and knowledge. *PLoS ONE* 13:e0196493. doi: 10.1371/journal.pone.0196493

Daikoku, T., Ogura, H., and Watanabe, M. (2012). The variation of hemodynamics relative to listening to consonance or dissonance during chord progression. *Neurol. Res.* 34, 557–563. doi: 10.1179/1743132812Y.0000000047

Daikoku, T., Okano, T., and Yumoto, M. (2017c). "Relative difficulty of auditory statistical learning based on tone transition diversity modulates chunk length in the learning strategy," in *Proceedings of the Biomagnetic* (Sendai), 75.

Daikoku, T., Takahashi, Y., Futagami, H., Tarumoto, N., and Yasuda, H. (2017a). Physical fitness modulates incidental but not intentional statistical learning of simultaneous auditory sequences during concurrent physical exercise. *Neurol. Res.* 39, 107–116. doi: 10.1080/01616412.2016.1273571

Daikoku, T., Takahashi, Y., Tarumoto, N., and Yasuda, H. (2018). Auditory statistical learning during concurrent physical exercise and the tolerance for pitch, tempo, and rhythm changes. *Motor Control* 22, 233–244. doi: 10.1123/mc.2017-0006

Daikoku, T., Yatomi, Y., and Yumoto, M. (2014). Implicit and explicit statistical learning of tone sequences across spectral shifts. *Neuropsychologia* 63, 194–204. doi: 10.1016/j.neuropsychologia.2014.08.028

Daikoku, T., Yatomi, Y., and Yumoto, M. (2015). Statistical learning of music- and language-like sequences and tolerance for spectral shifts. *Neurobiol. Learn. Mem.* 118, 8–19. doi: 10.1016/j.nlm.2014.11.001

Daikoku, T., Yatomi, Y., and Yumoto, M. (2016). Pitch-class distribution modulates the statistical learning of atonal chord sequences. *Brain Cogn.* 108, 1–10. doi: 10.1016/j.bandc.2016.06.008

Daikoku, T., Yatomi, Y., and Yumoto, M. (2017b). Statistical learning of an auditory sequence and reorganization of acquired knowledge: a time course of word segmentation and ordering. *Neuropsychologia* 95, 1–10. doi: 10.1016/j.neuropsychologia.2016.12.006

Daikoku, T., and Yumoto, M. (2017). Single, but not dual, attention facilitates statistical learning of two concurrent auditory sequences. *Sci. Rep.* 7:10108. doi: 10.1038/s41598-017-10476-x

Daikoku, T., and Yumoto, M. (2019). Concurrent statistical learning of ignored and attended sound sequences: an MEG study. *Front. Hum. Neurosci.* 13:102. doi: 10.3389/fnhum.2019.00102

de Zubicaray, G., Arciuli, J., and Mcmahon, K. (2013). Putting an "End" to the motor cortex representations of action words. *J. Cogn. Neurosci.* 25, 1957–1974. doi: 10.1162/jocn_a_00437

Dienes, Z., Altmann, G., and Gao, S. (1999). Mapping model across domains a neural feedback : network of implicit of transfer of implicit knowledge. *Cogn. Sci.* 23, 53–82. doi: 10.1207/s15516709cog2301_3

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. N. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding.* Cambridge, MA: MIT Press.

Elman, J. L. (1990). Finding structure in time. *Cogn. Sci.* 14, 179–211. doi: 10.1207/s15516709cog1402_1

Feher, O., Ljubičić, I., Suzuki, K., Okanoya, K., and Tchernichovski, O. (2016). Statistical learning in songbirds : from self- tutoring to song culture1. *Philos. Trans. R. Soc. B Biol. Sci.* 372:20160053. doi: 10.1098/rstb.2016.0053

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Friston, K. (2013). Life as we know it. *J. R. Soc. Interface* 10:20130475. doi: 10.1098/rsif.2013.0475

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci.* 6, 187–224. doi: 10.1080/17588928.2015.1020053

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2014). *The Anatomy of Choice: Dopamine and Decision-Making Subject Collections The Anatomy of Choice: Dopamine and Decision-Making.* Available online at: http://rstb.royalsocietypublishing.org/content/369/1655/20130481.full.html#related-urls%5Cnhttp://rstb.royalsocietypublishing.org/content/369/1655/20130481.full.html#ref-list-1%5Cnhttp://dx.doi.org/10.1098/rstb.2013.0481

Furl, N., Kumar, S., Alter, K., Durrant, S., Shawe-Taylor, J., and Griffiths, T. D. (2011). Neural prediction of higher-order auditory sequence statistics. *Neuroimage* 54, 2267–2277. doi: 10.1016/j.neuroimage.2010.10.038

Gjerdingen, R. O. (2014). "Historically Informed" corpus studies. *Music Percept. Interdiscipl. J.* 31, 192–204. doi: 10.1525/MP.2014.31.3.192

Hansen, N. C., and Pearce, M. T. (2014). Predictive uncertainty in auditory sequence processing. *Front. Psychol.* 5:1052. doi: 10.3389/fpsyg.2014.01052

Harrison, L. M., Duggins, A., and Friston, K. J. (2006). Encoding uncertainty in the hippocampus. *Neural Netw.* 19, 535–546. doi: 10.1016/j.neunet.2005.11.002

Hasson, U. (2017). The neurobiology of uncertainty : implications for statistical learning. *Philos. Trans. R. Soc. B* 372:20160048. doi: 10.1098/rstb.2016.0048

Hochreiter, S., and Urgen Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Johnson, D., Tyson, A., and Wnter, R. (1985). *The Beethoven Sketchbooks.* Oxford: Clarendon.

Jones, M. N., and Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychol. Rev.* 114, 1–37. doi: 10.1037/0033-295X.114.1.1

Kagan, J. (1972). Motives and development. *J. Personal. Soc. Psychol.* 22, 51–66. doi: 10.1037/h0032356

Kanai, R., Komura, Y., Shipp, S., and Friston, K. (2015). Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20140169. doi: 10.1098/rstb.2014.0169

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as bayesian inference. *Annu. Rev. Psychol.* 55, 271–304. doi: 10.1146/annurev.psych.55.090902.142005

Kirchhoff, M., Parr, T., Palacios, E., Friston, K., and Kiverstein, J. (2018). The Markov blankets of life: autonomy, active inference and the free energy principle. *J. R. Soc. Interface* 15:20170792. doi: 10.1098/rsif.2017.0792

Knill, D. C., and Pouget, A. (2004). The bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci.* 27, 712–719. doi: 10.1016/j.tins.2004.10.007

Koelsch, S. (2014). Brain correlates of music-evoked emotions. *Nat. Rev. Neurosci.* 15, 170–180. doi: 10.1038/nrn3666

Koelsch, S., Vuust, P., and Friston, K. (2018). Predictive processes and the peculiar case of music. *Trends Cogn. Sci.* 23, 63–77. doi: 10.1016/j.tics.2018.10.006

Krebs, R. M., Schott, B. H., Schütze, H., and Düzel, E. (2009). The novelty exploration bonus and its attentional modulation. *Neuropsychologia* 47, 2272–2281. doi: 10.1016/j.neuropsychologia.2009.01.015

Landauer, T. K., and Dumais, S. T. (1997). A solution to Platos problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychol. Rev.* 104, 211–240. doi: 10.1037/0033-295X.104.2.211

Loewenstein, G. (1994). The psychology of curiosity: a review and reinterpretation. *Psychol. Bull.* 116, 75–98. doi: 10.1037/0033-2909.116.1.75

Lund, K., and Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behav. Res. Methods Instrum. Comput.* 28, 203–208. doi: 10.3758/BF03204766

Monroy, C., Meyer, M., Gerson, S., and Hunnius, S. (2017b). Statistical learning in social action contexts. *PLoS ONE* 12:e0177261. doi: 10.1371/journal.pone.0177261

Monroy, C. D., Gerson, S. A., Domínguez-Martínez, E., Kaduk, K., Hunnius, S., and Reid, V. (2017a). Sensitivity to structure in action sequences: an infant event-related potential study. *Neuropsychologia* 126, 92–101. doi: 10.1016/j.neuropsychologia.2017.05.007

Monroy, C. D., Meyer, M., Schröer, L., Gerson, S. A., and Hunnius, S. (2017c). The infant motor system predicts actions based on visual statistical learning. *Neuroimage* 185, 947–954. doi: 10.1016/j.neuroimage.2017.12.016

Nastase, S., Iacovella, V., and Hasson, U. (2014). Uncertainty in visual and auditory series is coded by modality-general and modality-specific neural systems. *Hum. Brain Mapp.* 35, 1111–1128. doi: 10.1002/hbm.22238

O'Reilly, J. X., Jbabdi, S., and Behrens, T. E. (2012). How can a Bayesian approach inform neuroscience? *Eur. J. Neurosci.* 35, 1169–1179. doi: 10.1111/j.1460-9568.2012.08010.x

Parr, T., and Friston, K. J. (2018). The anatomy of inference: generative models and brain structure. *Front. Comput. Neurosci.* 12:90. doi: 10.3389/fncom.2018.00090

Parr, T., Rees, G., and Friston, K. J. (2018). Computational neuropsychology and bayesian inference. *Front. Hum. Neurosci.* 12:61. doi: 10.3389/fnhum.2018.00061

Pearce, M., and Wiggins, G. A. (2006). Expectation in melody: the influence of context and learning. *Music Precept.* 23, 377–405. doi: 10.1525/mp.2006.23.5.377

Pearce, M. T. (2005). *The Construction and Evaluation of Statistical Models of Melodic Structure in Music Perception and Composition.* Unpublished Doctoral Thesis, City University London.

Pearce, M. T., Müllensiefen, D., and Wiggins, G. A. (2010b). The role of expectation and probabilistic learning in auditory boundary perception: a model comparison. *Perception* 39, 1367–1391. doi: 10.1068/p6507

Pearce, M. T., Ruiz, M. H., Kapasi, S., Wiggins, G. A., and Bhattacharya, J. (2010a). Unsupervised statistical learning underpins computational, behavioural, and neural manifestations of musical expectation. *Neuroimage* 50, 302–313. doi: 10.1016/j.neuroimage.2009.12.019

Pearce, M. T., and Wiggins, G. A. (2012). Auditory expectation: the information dynamics of music perception and cognition. *Top. Cogn. Sci.* 4, 625–652. doi: 10.1111/j.1756-8765.2012.01214.x

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference SanMateo. Morgan Kaufmann Series in Representation and Reasoning.* San Francisco, CA: Morgan Kaufmann Publishers Inc.

Perruchet, P., and Pacton, S. (2006). Implicit learning and statistical learning: one phenomenon, two approaches. *Trends Cogn. Sci.* 10, 233–238. doi: 10.1016/j.tics.2006.03.006

Perruchet, P., and Vinter, A. (1998). PARSER: A model for word segmentation. *J. Mem. Lang.* 39, 246–263.

Pezzulo, G., Rigoli, F., and Friston, K. (2015). Active inference, homeostatic regulation and adaptive behavioural control. *Progr. Neurobiol.* 134, 17–35. doi: 10.1016/j.pneurobio.2015.09.001

Przysinda, E., Zeng, T., Maves, K., Arkin, C., and Loui, P. (2017). Jazz musicians reveal role of expectancy in human creativity. *Brain Cogn.* 119, 45–53. doi: 10.1016/j.bandc.2017.09.008

Rogers, T. T., and McClelland, J. L. J. (2004). Semantic cognition: a parallel distributed processing approach. *Attent. Perform.* 425:439. doi: 10.7551/mitpress/6161.001.0001

Rohrmeier, M., and Cross, I. (2008). "Statistical Properties of Tonal Harmony in Bach's Chorales," in *Proc 10th Intl Conf on Music Perception and Cognition* (Sapporo). 6, 123–1319.

Rohrmeier, M., and Rebuschat, P. (2012). Implicit learning and acquisition of music. *Top. Cogn. Sci.* 4, 525–553. doi: 10.1111/j.1756-8765.2012.01223.x

Roux, F., and Uhlhaas, P. J. (2014). Working memory and neural oscillations: α-gamma versus θ-gamma codes for distinct WM information? *Trends Cogn. Sci.* 18, 16–25. doi: 10.1016/j.tics.2013.10.010

Saffran, J., Aslin, R., and Newport, E. (1996). Statistical learning by 8 mbox{{}-{}}{{}m{}}onth-old infants. *Science* 274, 1926–1928. doi: 10.1126/science.274.5294.1926

Salimpoor, V. N., Zald, D. H., Zatorre, R. J., Dagher, A., and McIntosh, A. R. (2015). Predictions and the brain: how musical sounds become rewarding. *Trends Cogn. Sci.* 19, 86–91. doi: 10.1016/j.tics.2014.12.001

Schmidhuber, J. (2006). Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connect. Sci.* 18, 173–187. doi: 10.1080/09540090600768658

Schwartenbeck, P., FitzGerald, T., Dolan, R. J., and Friston, K. (2013). Exploration, novelty, surprise, and free energy minimization. *Front. Psychol.* 4:710. doi: 10.3389/fpsyg.2013.00710

Servan-Schreiber, E., and Anderson, J. R. (1990). Learning artificial grammars with competitive chunking. *J. Exp. Psychol. Learn. Mem. Cogn.* 16, 592–608. doi: 10.1037/0278-7393.16.4.592

Shen, L., Fishbach, A., and Hsee, C. K. (2015). The motivating-uncertainty effect: uncertainty increases resource investment in the process of reward pursuit. *J. Consumer Res.* 41, 1301–1315. doi: 10.1086/679418

Strange, B. A., Duggins, A., Penny, W., Dolan, R. J., and Friston, K. J. (2005). Information theory, novelty and hippocampal responses: unpredicted or unpredictable? *Neural Netw.* 18, 225–230. doi: 10.1016/j.neunet.2004.12.004

Sullivan, J. (1927). *Beethoven: His Spiritual Development*. New York, NY: A.A. Knopf.

Thiessen, E. D., Kronstein, A. T., and Hufnagle, D. G. (2013). The extraction and integration framework: a two-process account of statistical learning. *Psychol. Bull.* 139, 792–814. doi: 10.1037/a0030801

Tishby, N., and Polani, D. (2011). "Information theory of decisions and actions," in *Perception-action Cycle,* eds V. Cutsuridis, A. Hussain, and J. G. Taylor (New York, NY: Springer), 601–636.

Turk-browne, N. B., Scholl, B. J., Chun, M. M., and Johnson, M. K. (2009). Neural evidence of statistical learning: efficient detection of visual regularities without awareness. *J. Cogn. Neurosci.* 21, 1934–1945. doi: 10.1162/jocn.2009.21131

White, C. W. (2014). Changing styles, changing corpora, changing tonal models. *Music Percept. Interdiscipl. J.* 31, 244–253. doi: 10.1525/mp.2014.31.3.244

Wiggins, G. A. (2018). Creativity, information, and consciousness: the information dynamics of thinking. *Phys. Life Rev.* 1, 1–39. doi: 10.1016/j.plrev.2018.05.001

Wiggins, G. A., and Sanjekdar, A. (in press). Consolidation as re-representation: revising the meaning of memory. *Front. Psychol.* doi: 10.3389/fpsyg.2019.00802

Wittmann, B. C., Daw, N. D., Seymour, B., and Dolan, R. J. (2008). striatal activity underlies novelty-based choice in humans. *Neuron* 58, 967–973. doi: 10.1016/j.neuron.2008.04.027

Yumoto, M., and Daikoku, T. (2016). "IV Auditory system. 5 basic function," in *Clinical Applications of Magnetoencephalography,* eds S. Tobimatsu and R. Kakigi (Springer).

Yumoto, M., and Daikoku, T. (2018). Neurophysiological Studies on Auditory Statistical Learning [in Japanese] **聴覚**刺激列の統計学習の**神経**生理学的研究 *Jpn. J. Cogn. Neurosci. 認**知神経科学***. 20, 38–43. doi: 10.11253/ninchishinkeikagaku.20.38

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

frontiers in Computational Neuroscience

SYSTEMATIC REVIEW
published: 24 May 2019
doi: 10.3389/fncom.2019.00031

# The Application of Unsupervised Clustering Methods to Alzheimer's Disease

*Hany Alashwal[1*†], Mohamed El Halaby[2†], Jacob J. Crouse[3], Areeg Abdalla[2] and Ahmed A. Moustafa[4]*

[1] Department of Computer Science and Software Engineering, College of Information Technology, United Arab Emirates University, Al-Ain, United Arab Emirates, [2] Department of Mathematics, Faculty of Science, Cairo University, Giza, Egypt, [3] Brain and Mind Centre, The University of Sydney, Sydney, NSW, Australia, [4] School of Social Sciences and Psychology, Western Sydney University, Sydney, NSW, Australia

Clustering is a powerful machine learning tool for detecting structures in datasets. In the medical field, clustering has been proven to be a powerful tool for discovering patterns and structure in labeled and unlabeled datasets. Unlike supervised methods, clustering is an unsupervised method that works on datasets in which there is no outcome (target) variable nor is anything known about the relationship between the observations, that is, unlabeled data. In this paper, we focus on studying and reviewing clustering methods that have been applied to datasets of neurological diseases, especially Alzheimer's disease (AD). The aim is to provide insights into which clustering technique is more suitable for partitioning patients of AD based on their similarity. This is important as clustering algorithms can find patterns across patients that are difficult for medical practitioners to find. We further discuss the implications of the use of clustering algorithms in the treatment of AD. We found that clustering analysis can point to several features that underlie the conversion from early-stage AD to advanced AD. Furthermore, future work can apply semi-clustering algorithms on AD datasets, which will enhance clusters by including additional information.

**Keywords: clustering, neurological diseases, Alzheimer's disease, unsupervised learning, machine learning techniques**

OPEN ACCESS

***Edited by:***
*Carlo Laing,*
*Massey University, New Zealand*

***Reviewed by:***
*Xiaofeng Zhu,*
*Massey University, New Zealand*
*Tuo Zhang,*
*Northwestern Polytechnical University, China*

***Correspondence:**
*Hany Alashwal*
*halashwal@uaeu.ac.ae*

*† These authors have contributed equally to this work*

***Received:** 20 January 2019*
***Accepted:** 29 April 2019*
***Published:** 24 May 2019*

***Citation:***
*Alashwal H, El Halaby M, Crouse JJ, Abdalla A and Moustafa AA (2019) The Application of Unsupervised Clustering Methods to Alzheimer's Disease. Front. Comput. Neurosci. 13:31. doi: 10.3389/fncom.2019.00031*

## INTRODUCTION

There has been an increasing interest in the medical community to use machine learning techniques for disease diagnosis (Kononenko, 2001). This is due to the increases in availability of medical datasets, such as Twinanda et al. (2017), Srivastav et al. (2018), Alzheimer's Disease Neuroimaging Initiative (ADNI), and UC Irvine Machine Learning Repository, among others. The accumulation of large datasets has become more feasible recently due to the advancements in hardware (fast, cheap computers), the availability of public and private medical and healthcare datasets, and machine learning classification and clustering methods.

Supervised learning is the process of learning (approximating) a mapping function from a set of input variables to a target variable. The term "supervised" here refers to the training process of the algorithm being supervised by having the correct answers (i.e., we know what the target outcome is). However, when one only has a set of variables and no corresponding output variables (i.e., the data are unlabeled), then the learning process is called unsupervised. Thus, in unsupervised learning, there are no correct answers for the training procedure to learn from and the learning

algorithm is left to discover the structures in the datasets. One of the most important unsupervised learning techniques is clustering, which is the process of partitioning a set of data points according to some measure of similarity (e.g., distance). The goal of clustering is to reveal subgroups within heterogeneous data such that each individual cluster has greater homogeneity than the whole (Eick et al., 2004). **Table 1** summarizes the different types of machine learning methods and some of their real-world applications. In many applications, obtaining labeled data is often difficult, costly, and/or time-consuming, while collecting unlabeled data may be relatively easy. Such cases result in a dataset consisting of a large number of unlabeled variables and a small set of labeled variables. Semi-supervised learning uses both labeled and unlabeled data to improve the accuracy of the learning model.

Several studies have used clustering methods to facilitate the diagnosis of several disorders (Vogt and Nagel, 1992; Nugent and Meila, 2010; Li and Zhu, 2013; Nithya et al., 2013; Wiwie et al., 2015). For example, clustering techniques have been applied to the diagnosis of breast cancer (Chen, 2014), Parkinson's disease (Polat, 2012; Nilashi et al., 2016), headache (Wu et al., 2015), mental health and psychiatric disorders (Trevithick et al., 2015), heart and diabetes diseases (Yilmaz et al., 2014), and Huntington's disease (Nikas and Low, 2011), among many others.

Alzheimer's disease (AD) is one of the most common neurodegenerative diseases, particularly in old age (Ryu et al., 2010), and is among the most common causes of dementia in senior individuals (Ryu et al., 2010; Cuingnet et al., 2011). AD leads to structural and functional loss of neurons in the cortex and hippocampal regions, among other brain areas. A number of studies in the past 20 years have pointed out possible biomarkers for the diagnosis of AD, including brain atrophy revealed by magnetic resonance imaging (Mueller et al., 2006; Seppi and Poewe, 2010).

## METHOD

In this paper, we summarize prior studies that use clustering methods on AD datasets to gain more insights into the disease's nature, diagnosis, and progression. In the following sections, we describe the most common clustering algorithms and their application on AD datasets in the literature. A computer search was carried out, containing the clustering and AD. This search was performed in PubMed and Google Scholar.

## CLUSTERING ALGORITHMS

### k-Means

The k-Means clustering algorithm (Forgy, 1965) is a classical unsupervised learning method. This algorithm takes $n$ observations and an integer $k$. The output is a partition of the $n$ observations into $k$ sets such that each observation belongs to the cluster with the nearest mean. The following steps summarize the operations of k-Means.

Initialize $k$ cluster centers. In practice, this can be done by either randomly selecting $k$ center

1. points from the $n$ observations or random generation of $k$ center points.
2. Calculate the distance between each observation and the cluster centers.
3. Assign each point to the cluster whose distance from its center is minimum of all the cluster centers.
4. Recompute the positions of the $k$ centers as the cluster mean.
5. Recompute the distance between each data point and the newly computed centers. Repeat steps 3 and 4 until all data points are assigned to the same cluster (data points do not move).

The choice of $k$ is usually influenced by prior knowledge regarding the nature of the data or by using clustering validity measures.

Escudero et al. (2011) investigated how applying k-Means clustering to a subject's medical history may shed light on the likelihood of conversion from mild cognitive impairment (MCI) to AD. The dataset used was obtained from the ADNI database and consists of 375 subjects. The selected features included the number of ApoE s4 alleles, ADAS-Cog (Alzheimer's Disease Assessment Scale-Cog), Mini-Mental State Examination (MMSE) scores, MRI (magnetic resonance imaging), and CSF (cerebrospinal fluid) data from cognitively normal (CN), MCI, and AD individuals. The authors tested the potential of how having the following five sets of features can better diagnose AD: (1) ADAS-Cog, MMSE, and ApoE genotype obtained from a blood sample; (2) CSF; (3) MRI; (4) CSF and MRI; and (5) all

**TABLE 1 |** Types of machine learning methods.

| Learning type | Supervised | Unsupervised | Semi-supervised |
|---|---|---|---|
| Type of data | Data points have labels. | Data points do not have corresponding labels. | A subset of the data points is labeled. |
| Learning process | Analyzing the training data to learn a function that can be used for predicting the labels of new examples. | Modeling the structure or the distribution of the data in order to find patterns and gain new insights from the data. | Utilizing unlabeled data with labeled data to learn better models. |
| Applications | Fraud detection, detecting spam emails, predicting real estate prices. | Clustering customers' data and market segmentation, learning rule associations, image segmentation, gene clustering. | When it is expensive to annotate every data point (e.g., using humans), this type of learning is suitable. Examples: web content classification, medical predictions. |

*Firstly, the nature of the data is stated, then the objective of the type of learning is discussed, and finally some real-world examples are mentioned.*

of the above features. The first analysis involved clustering the subjects according to each of the five scenarios (i.e., using only a subset of the variables based on the set of features described above) using k-Means and approximating the occurrence of the medical history of AD in each set. More than 69% of the AD subjects and about half of the MCI individuals were always assigned to the pathological bioprofile.

In the second analysis, k-Means was applied to the CN and AD subjects, and the obtained clusters were used to split the MCI subjects into CN-like and AD-like, that is, which MCI subjects may stay as healthy individuals and which may convert to AD. Next, the rate of decline to AD was used to evaluate the utility of this clustering algorithm in the early diagnosis of AD at the MCI stage. The fifth set of features (which included all features) provided larger differences between the evolution of CN-like and AD-like subjects at the 12-month follow-up. The number of subjects assigned to CN-like and AD-like was 82 and 96, respectively. This indicates that the combination of all clinical tests and biomarkers outperformed using any of them in isolation.

In a recent study, Tosto et al. (2016) applied k-Means clustering algorithm on a dataset of 3,502 patients with AD with longitudinal assessments from the National Alzheimer's Coordinating Center database, with 394 providing neuropathological data. The authors were interested in examining subgroups of patients with variable trajectories of extrapyramidal sign progression (which include movement disorders such as postural instability, tremors and rigidity, body restlessness, and abnormal gait, among others) and their clinical and neuropathological correlates. Tosto et al. (2016) observed the following three clusters of extrapyramidal sign progression: no/low ($n = 1{,}583$), medium ($n = 1{,}259$), and high ($n = 660$) extrapyramidal burden. The high extrapyramidal cluster had greater cognitive and neuropsychiatric impairment (particularly hallucinations), relative to the other clusters. Moreover, despite the three clusters having similar AD pathology, the high extrapyramidal burden cluster had a significantly greater number of patients diagnosed with dementia with Lewy bodies.

In another recent study, Price et al. (2015) recruited participants with AD or vascular dementia and collected MRI measures of infarction, whole brain volume, and leukoaraiosis (LA), as well as neurocognitive measures in all participants. A k-Means cluster analysis derived three cluster-groups characterized by single-domain amnestic ($n = 41$), single-domain dysexecutive ($n = 26$), and multi-domain ($n = 26$) phenotypes. The multi-domain patients scored worse on language measures than the other clusters, yet they were equally impaired on tests of memory when compared to the amnestic group. The three cluster-groups were relatively dissociable in neuroradiological parameters, in which the amnestic and multi-domain clusters had smaller hippocampal volume than the third cluster, while the single-domain dysexecutive cluster had greater deep periventricular (i.e., between periventricular and infracortical regions) and whole brain LA. The volume of the caudate and lacunar infarction did not differ between the three clusters. There was a negative association between the volume of the caudate nucleus and total LA in the dysexecutive and multi-domain clusters. These results suggest the existence of neuroradiological heterogeneity between patients diagnosed with AD/vascular dementia spectrum dementia.

## k-Means-Mode

This algorithm can deal with both numeric (continuous) and categorical data. Each cluster center is an array of means and modes for continuous and categorical attributes, respectively. The steps of the algorithm is similar to that of the classical k-Means; the means and modes are calculated for each cluster as previously stated, and then each point is moved to the cluster with minimum distance. For continuous features, Euclidean distance is often used, and for discrete features, Hamming distance is often used.

Paul and Hoque (2010) have applied the k-Means-Mode clustering algorithm to medical datasets to predict the likelihood of diseases. The likelihood of the disease in a cluster is defined as the number of patients that have the disease divided by the total number of points in the clusters. In other words, it is the probability of finding the disease in the cluster. The average likelihood of all clusters is the actual probability of the disease in the data, which can be found by brute-force methods. Accuracy is the ratio between average likelihood and actual likelihood. Experimental results show that when the algorithm was applied on the Zoo dataset from the University of California at Irvine (UCI) Machine Learning Repository and a diabetes dataset, an accuracy of about 95% is achieved. Other algorithms like k-Means and k-Mode achieved lower than 65% accuracy, suggesting that the k-Means-Mode algorithm is better at clustering data than k-Means and k-Mode algorithms.

## Multi-Layer Clustering

The first step of the multi-layer clustering process is to determine the similarity between each pair of examples. This is done by creating an artificial binary classification problem having the original patient records as the positive example, while negative examples are generated by randomly mixing the values of the attributes of the original examples among themselves. Next, a predictive model is built to distinguish between the positive and negative examples to determine the similarities between each pair of examples. The Random rules algorithm (Pfahringer et al., 2004; Almeida et al., 2013) is applied for each pair of records to construct an example similarity table (EST) where the number of rules covering the pair is calculated. An entry $e_{i,j}$ in the table holds the similarity value between the $i$th and the $j$th example. The second step is to calculate the clustering-related variability (CRV) measure for all examples. The single-layer clustering algorithm starts by assigning each example to a single cluster. It then keeps merging the most similar clusters in terms of the cluster CRV score. The procedure stops when no further merge operations are possible; that is, further merges do not result in a smaller CRV score. In situations having more than one attribute layer (multi-layer attributes), the artificial binary classification problem is constructed for each attribute layer and the ESTs are built. As for the algorithm, for each pair of clusters, the potential variability reduction for all attribute layers is computed and the smallest value for each pair is selected. Merging occurs if this value is

positive, and if the value is positive for more than one pair, the pair with the largest minimal value is chosen and these clusters are merged.

Gamberger et al. (2016a) applied a multi-clustering method to an AD dataset of both male and female patients comprising 243 biological and clinical features. The clusters obtained showed differences between male and female patient groups, including the existence of two male subpopulations with changes to intracerebral and whole brain volumes. The multi-layer clustering technique was used to deal with layers of attributes; that is, a set of attributes is partitioned into several subsets according to a criterion (e.g., laboratory data features and clinical data features). The multi-layer clustering technique was carried out independently on two groups of 317 female and 342 male patients. The first layer consisted of 56 biological measurements and the second consisted of 187 symptoms and clinical descriptors. The authors reported key differences between male and female populations of patients. For example, in the female population, there were two clusters, while in the male population, there were four, two for patients having major issues with dementia (denoted M1 and M2) and two for patients having mild or no dementia (denoted M0A and M0B). There was one large cluster in the female population, denoted F1, with patients having significant problems with dementia, while patients in the other cluster had mild dementia symptoms (denoted F0). Patients in cluster M2 were found to have higher than average intracranial volume (ICV) and whole brain volumes when compared to cognitively normal male patients. Such a cluster was not observed in the female population. The M0A cluster was similar to cluster F0 in the female population in terms of increased ICV values and biological features, while cluster M0B had smaller than average ICV values. This analysis showed that there are significant gender-specific differences in AD patients and suggests that taking gender into account may have important implications for the treatment of AD.

The same multi-layer clustering algorithm used by Gamberger et al. (2016a) was also used on a dataset of 218 female and 344 male individuals with MCI. The algorithm first builds an EST for each attribute layer and then the tables are used by a bottom-up method to merge similar clusters together until no further merging of clusters is possible. The goal of this study is to find homogeneous groups of MCI individuals in terms of baseline and prognostic features and to discover gender differences within the groups. The algorithm produced a cluster of "slow decliners" (i.e., individuals with MCI that slowly develop dementia symptoms) consisting of 184 subjects that included a subset of MCI individuals that had favorable baseline data and prognosis. Another cluster given by the algorithm, termed "rapid decliners" (i.e., individuals with MCI that rapidly develop dementia symptoms; $n = 240$), consisted of a subset of MCI subjects with a more impaired baseline status and a rapidly progressing longitudinal cognitive course. Moreover, 138 subjects did not fit in either of the two clusters. Males and females in the "rapid decliners" cluster had worse baseline cognitive status and smaller brain volumes than those in the "slow decliners" cluster. The rate of progression from MCI to dementia for females and males in the "rapid decliners" cluster was 69 and 61%, respectively. Conversely, the rate of progression from MCI to dementia for females and males in the slow decliners cluster was 9 and 16%, respectively.

Gamberger et al. (2016b) applied the multi-layer clustering method used by Gamberger et al. (2016a) and Gamberger et al. (2017) to an AD dataset obtained from ADNI. The dataset consists of 187 cognitively normal (CN) subjects, 106 patients with significant memory concern (SMC), 311 patients with early MCI (EMCI), 164 patients with late MCI (LMCI), and 148 AD patients (916 subjects in total). There are two layers that make up the features: layer 1 consists of 10 biological features and layer 2 consists of 23 clinical features. The goal of this study was to find clusters that are as large and homogeneous as possible regarding both biological and clinical features. Three clusters were identified having patients with different levels of dementia. The first cluster, A, contained patients with low volumes of hippocampus, entorhinal cortex, fusiform gyrus, and middle temporal gyrus, as well as small intracerebral and whole brain volumes. The number of subjects in that cluster diagnosed with AD, LMCI, and EMCI were 30, 4, and 1, respectively. Compared to CN subjects, patients in cluster A had 20% lower mean values for fusiform and midtemporal gyrus. Moreover, patients in cluster A had, on average, a 30% smaller entorhinal volume than the CN group. The authors regarded it odd that patients with LMCI and EMCI were assigned to this cluster, yet offered no explanation for this discrepancy. It is quite possible that these individuals may be at risk for converting to AD; this hypothesis should be tested in future work. Further, patients in cluster A showed high Clinical Dementia Rating Sum of Boxes (CDRSB), Alzheimer's Disease Assessment Scale (ADAS13), and Functional Assessment Questionnaire (FAQ) scores and low Mini-Mental State Examination (MMSE) and Montreal Cognitive Assessment (MoCA) scores, which is consistent with patients suffering from acute dementia. Importantly, the number of AD, LMCI, and EMCI patients in the second cluster, B, was 10, 9, and 2, respectively. Patients in this cluster have, to some extent, had smaller volumes of entorhinal, hippocampus, fusiform, and midtemporal gyrus that are about 20, 20, 10, and 10% (respectively) lower than mean values for CN subjects. However, the intracranial volume and whole brain volume were normal. Subjects in this cluster had a moderate or mild type of AD, which is indicated by a score above 3 in the CDRSB. An interesting feature of patients in cluster B was that the values for cognitive functions self-reported by the patients were higher than those of the other clusters and of the mean values of the entire AD population.

The third cluster, C, included patients with the lowest degenerative changes in the hippocampus, entorhinal, fusiform, and midtemporal gyrus. Moreover, patients in this cluster had high scores of ventricular and whole brain volumes. Cluster C patients had larger mean ventricle volume than CN subjects. The values for the scales of the MoCA, FAQ, fluorodeoxyglucose imaging (FDG), MMSE, and ADAS13 were all intermediate between those of clusters A and B. Cluster C patients also showed impairment, performing the Rey's Auditory Verbal Learning Test (RAVLT), and divided attention.

This study shows that the nature of the cluster of patients having problems with dementia is non-homogeneous. Moreover, cognitively normal subjects are even more non-homogeneous as a population, as the clustering algorithm reported here shows that there are many clusters of controls as well. The number of AD patients assigned to clusters A, B, and C is <50% of the entire AD population. Another important finding of the current study is the correlation between cognitive impairment and brain atrophy. The presence of degenerative changes of the brain was found in the three derived clusters. The greatest degeneration was found in cluster A and the second greatest degeneration was found in cluster B. The results obtained from cluster C indicate that brain changes are responsible for a significant number of problems with dementia; however, they are not sufficient for AD development.

## HIERARCHAL AGGLOMERATIVE CLUSTERING

Hierarchal agglomerative clustering is a bottom-up approach such that each data point begins in a separate cluster, and pairs of clusters at the bottom are merged together as we go up the hierarchy. This method can be summarized as follows:

1. Assign each object to a separate cluster.
2. For each pair of clusters, calculate the pairwise distance. Then, build a matrix whose elements are the distance values computed.
3. Find the pair of clusters with the shortest distance.
4. Merge the identified pair after removing both clusters from the distance matrix.
5. Calculate all distances from this new cluster to all other clusters and update the distance matrix.
6. Repeat these steps until the matrix is reduced to a single element.

There are several distance metrics that can be used (e.g., Euclidean and Manhattan distances); however, the choice of a metric determines the shape of the clusters produced. This is because two clusters can be close to each other according to one metric, but far from each other according to another metric. It is recommended that an exploratory study be conducted on several distance measures and the one that yields the best results according to chosen performance measures is selected. Unlike k-Means, the number of clusters is not determined by the user, and generally, smaller clusters are generated, which can be helpful in many domains.

Noh et al. (2014) collected high-resolution T1-weighted volumetric MRIs from 152 patients in the early stages of AD. A hierarchical agglomerative clustering analysis was applied to measures of cortical thickness in these patients. Three emergent clusters were compared with an age- and sex-matched control group. The first cluster (A) was characterized by bilateral medial temporal-dominant atrophy predominantly involving anterior and posterior cingulate cortices ($n = 52$, 32.4%); the second cluster (B) was characterized by parietal-dominant atrophy involving bilateral parietal areas, precuneus, and bilateral dorsolateral frontal areas ($n = 28$, 18.4%); and the third cluster (C) was characterized by diffuse atrophy, in which almost all association cortices demonstrated atrophy (except for orbitofrontal and occipital areas) ($n = 72$, 47.4%). Patients in the parietal-dominant cluster (B) were younger, had a younger age at onset, and had the highest years of education. Patients in the diffuse atrophy cluster (C) had the lowest mean cortical thickness. Patients in the parietal-dominant cluster scored the poorest across all neurocognitive tests (attention, visuospatial function, memory, and frontal executive tasks) except for language function measures. These results suggest that there is considerable anatomical heterogeneity evident even in early stages of AD, which may indicate multiple disease processes.

Hwang et al. (2016) conducted several analyses on a dataset that includes 77 patients with AD recruited via the ADNI. Patients underwent 3-T MRI, [$^{18}$F]-fluorodeoxyglucose PET, [$^{18}$F]-florbetapir PET, and cerebrospinal (CSF) tests. Hierarchical agglomerative cluster analysis was applied to measures of cortical thickness, and the remaining measures were compared across groups. Consistent with the study by Noh et al. (2014) and Hwang et al. (2016) observed three clusters, dominated by medial–temporal atrophy (19.5%), parietal atrophy (24.7%), and diffuse atrophy (55.8%). The parietal-dominant cluster was younger and showed greater glucose hypometabolism in parietal and occipital cortices, as well as pronounced amyloid-beta accumulation in most brain regions. The medial–temporal dominant cluster had greater glucose metabolism in the left hippocampus and bilateral frontal cortices and poorer performance on memory tests. There were no significant differences in CSF tests between cluster-groups.

Racine et al. (2015) studied a sample of 103 asymptomatic adults with genetic risk and parental family history of AD. Participants underwent [C-11] Pittsburgh Compound B (PiB) amyloid imaging, MRI, lumbar puncture, and neurocognitive assessment at baseline, with 79 participants also undergoing follow-up PiB imaging 2 years later. The hierarchical agglomerative cluster analysis derived four cluster-groups based on three biomarkers, including CSF total-tau, CSF $A\beta_{42}$, and average PiB burden across 8 AD-sensitive regions of interest. All clusters were compared on amyloid accumulation (controlling for PiB baseline, age, sex, and APOE4 status) as well as on cognitive changes on tests of memory and executive control (controlling for baseline scores, age, sex, APOE4 status, education, and duration between testing visits). Cluster 4 showed the greatest AD-like characteristics (low CSF $A\beta_{42}$ and high PiB), with greater amyloid accumulation over 2 years relative to the other three clusters in regions affected by AD (precuneus, posterior cingulate, and lateral temporal and parietal cortices). Moreover, individuals in cluster 4 scored worse than those in cluster 1 on immediate recall and worse than all three clusters on delayed recall. Individuals in cluster 2 scored better than individuals in cluster 3 on delayed recall and better than both clusters 1 and 2 on total recall. These results suggest that clustering at-risk individuals across validated biomarkers may provide novel insights into those at greatest risk for amyloid accumulation and cognitive decline.

Cappa et al. (2014) recruited 23 patients with posterior cortical atrophy (PCA) and 16 patients with dementia of

Alzheimer's type (AD). First, a principal component analysis was used to reduce 15 neurocognitive variables to the following five factors: memory, language, perceptual processes, visuospatial processes, and calculation (addition, subtraction, and multiplication). These factors were then entered into a hierarchical agglomerative cluster analysis. Four clusters were derived and were characterized by visuospatial/perceptual, memory, perceptual/calculation, and language performance. Four clusters were derived, Cluster 1 ($n = 9$, 100% PCA), Cluster 2 ($n = 10$, 20% PCA), Cluster 3 ($n = 6$, 50% PCA), and Cluster 4 ($n = 14$, 64% PCA). The authors noted that AD pathology appears to produce multiple distinct syndromal subtypes involving impairment in memory (classically associated with AD) and visuospatial deficits (classically associated with PCA), as well as in visual perception and language, which may indicate heterogeneity in vulnerability of specific functional networks.

Armstrong and Wood (1994) applied hierarchical cluster analysis to a group of 78 patients with AD. The dataset consisted of 47 neuropathological measures, including the density and distribution of senile plaques and neurofibrillary tangles. The analyses indicated that an initial splitting of the sample could be made, characterizing one large group (68%) who had a relatively small distribution of senile plaques and neurofibrillary tangles across the brain and a second smaller cluster (15%) who had more diffusely spread lesions throughout the neocortex. These clusters could be further divided based on the extent of capillary amyloid angiopathy. Moreover, patients with a limited development of senile plaques, neurofibrillary tangles, and capillary amyloid angiopathy could be further split into an early- and a late-onset group. Patients with familial AD were not assigned to a single cluster; rather, they were distributed across four of the five groups. Some patients with familial AD had unique combinations of pathological features that did not closely resemble the other clusters.

McCurry et al. (1999) recruited a population-based sample of 205 patients with AD from the Alzheimer's Disease Patient Registry to investigate patterns of sleep problems. The authors applied hierarchical cluster analysis (Lance and Williams, 1967) to patients who were reported to have awakened their caregivers from sleep. They identified one cluster with daytime inactivity but few behavioral problems, another cluster with higher levels of fearfulness, fidgeting and occasional sadness, and a third cluster with multiple behavioral problems that included frequency bouts of sadness, fearfulness, inactivity, fidgeting, and hallucinations. The results demonstrate the heterogeneity of sleep disturbances in AD, which may have implications for the direction of interventions to homogeneous subgroups experiencing similar patterns of sleep problems.

## DISCUSSION

In this study, we were able to identify and review 13 articles that applied clustering methods on mainly AD datasets. To our knowledge, these are the only existing studies on clustering AD datasets. The distribution of these articles over time is presented in **Figure 1**.

Across all of these studies, there are four clustering algorithms used: k-Means, k-Means-Mode, multi-layer clustering, and hierarchical agglomerative clustering (see above sections for description of these clustering algorithms). As **Figure 2** shows, hierarchical agglomerative was the most commonly used method throughout the reviewed papers, followed by k-Means and multi-layer clustering and finally k-Means-Mode.

The reviewed studies vary across various dimensions including the clustering algorithm used, the dataset used, variables included in the dataset, and groups included in the datasets (i.e., AD, controls, MCI). Some of the studies have highlighted differences among males and females with AD (Gamberger et al., 2016a,b). Noting that AD is more common in females than in males (Viña and Lloret, 2010; Mazure and Swendsen, 2016), it is possible that there are gender-specific factors underlying the progression of AD in females. The Gamberger et al. studies have highlighted several neural changes between females and males with AD, suggesting that these neural changes may be the underlying reason behind AD being more common in females than in males. Some clustering analyses have shown that AD is not a homogeneous disorder and there are subtypes of AD patients. For example, Noh et al. (2014) have shown that there are three clusters of AD patients that differ in their neural damage. This is important as it may suggest different treatment for each subgroup of patients. Similar findings were also reported in Hwang et al. (2016), thus confirming the existence of subtypes of AD patients. Unlike other clustering studies, Racine et al. (2015) conducted clustering analysis on a dataset that includes individuals at risk for developing AD. The study was able to find several features that explain why some individuals may convert to AD while others do not. These features include low CSF Aβ42 and impaired immediate recall. Cappa et al. (2014) also reported the existence of several subtypes of AD patients that differ in memory and visuospatial impairment. Price et al. (2015) found that there were three groups of AD patients that are characterized by memory, executive dysfunction, or multiple impairments. Similarly, Tosto et al. (2016) found that there are three clusters of AD patients that vary in their extrapyramidal symptoms. According to Armstrong and Wood (1994), AD patients can be subdivided into several groups based on the distribution of senile plaques and neurofibrillary tangles in their brains. McCurry et al. (1999) also reported that there are subtypes of AD patients depending on their sleep disturbances. One problem with the abovementioned studies is that they subtyped AD patients based on very different features varying from neural, cognitive, and clinical variables. Accordingly, it is thus unclear what the subtypes of AD patients are, given the different features reported in every study.

Further, to our knowledge, there were only three studies that have used an MCI population in the clustering analysis (Escudero et al., 2011; Gamberger et al., 2016a,b). Gamberger et al. (2017) found that converting to dementia in individuals with MCI is related to worse baseline cognitive dysfunction as well as having smaller brain volumes. In another study, Gamberger et al. (2016a) found that few individuals with EMCI and some with LMCI were assigned to the same cluster as most AD patients. While the authors did not explain these results, it is possible that these MCI individuals may be at risk of developing AD, and thus were assigned to the AD cluster.

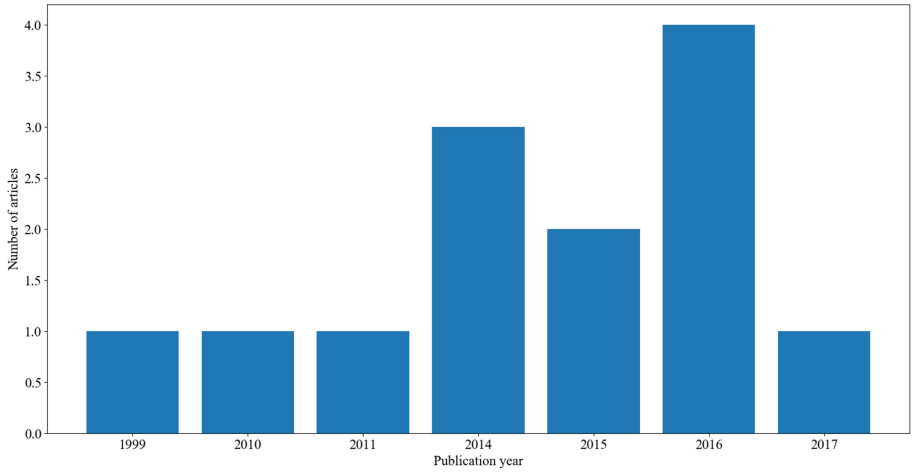


**FIGURE 1 |** A summary of the number of articles and their corresponding year of publication.

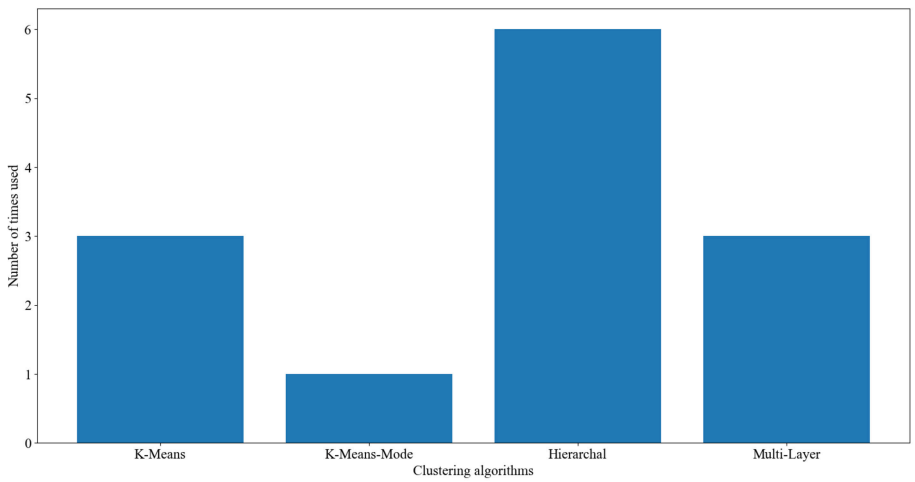


**FIGURE 2 |** The frequency of usage of clustering algorithms on Alzheimer's disease data.

Escudero et al. (2011) evaluated several analytic approaches for determining which MCI individuals are likely to convert to AD. They found that by using a large dataset that includes clinical tests and biomarkers in the clustering algorithms, greater accuracy is achieved compared to using smaller numbers of variables in isolation.

Further, to our knowledge, none of the existing studies on clustering analysis have used a dataset that includes early-stage vs. late-stage AD patients. Several experimental studies have shown that these two groups differ profoundly in terms of clinical, cognitive, and neural damage (Kauer-Sant'Anna et al., 2009). Like MCI conversion to AD, clustering analysis can point to several features that underlie the conversion from early-stage AD to advanced AD.

Importantly, while some other medical studies have used semi-clustering algorithms, to our knowledge, there are no studies on using semi-clustering algorithms in AD. While traditional clustering algorithms (as described in this article) work on datasets in which there is no outcome (target) variable nor is anything known about the relationship between the observations (i.e., unlabeled data), semi-clustering enhances clustering by using additional information as constraints in the clustering process. This is helpful in identifying clusters that are linked to a particular target variable. Such additional information is often existent in the dataset or provided by neurologists/clinicians to guide the clustering process. Future work should apply semi-clustering methods on AD.

## FUTURE RESEARCH

As mentioned above, only three studies have used an MCI population in the clustering analysis (Escudero et al., 2011;

Gamberger et al., 2016a,b). Future research should use more than three populations: healthy controls, individuals with MCI, and AD patients. For example, none of the clustering used subpopulations with MCI, such as amnestic vs. non-amnestic MCI. Such populations are increasingly being studied in the literature, as patients with amnestic MCI are more likely to develop AD than patients with non-amnestic MCI (Mauri et al., 2012; Monacelli et al., 2015).

Another type of clustering is known as fuzzy clustering, in which the classification function causes the class members to become a relative one and an object can belong to several classes at the same time but with different degrees (Ahmadi et al., 2018). Fuzzy clustering has many applications to health sciences, as some individuals may or may not be diagnosed with a certain disorder, depending on different conditions. This is quite relevant to AD. Fuzzy clustering can help us understand the nature of MCI, as some of these individuals may convert to AD, but others may stay healthy.

Further, to our knowledge, different kinds of clustering methods, such as latent profile analysis, were rarely applied to AD datasets. These algorithms do not use a distance function, but instead attempt to produce normally distributed clusters. The latent profile analysis has been applied to several disorders with some success. In one study, Aldridge and Roesch (2008) used latent profile analysis to classify subgroups of adolescents and examine rates of depression and anxiety in these different groups. They observed three clusters of adolescents who vary greatly in their depressive and anxiety symptoms. As another example, Mitchell et al. (2007) used latent profile analysis to subgroup individuals with eating disorders. The analysis revealed five subtypes that have very different profiles. Future research should use latent profile analysis clustering methods to better understand the nature of MCI and their conversion to AD.

## AUTHOR CONTRIBUTIONS

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

## ACKNOWLEDGMENTS

HA received financial support from the United Arab Emirates University (grant no. CIT 31T085).

## REFERENCES

Ahmadi, H., Gholamzadeh, M., Shahmoradi, L., Nilashi, M., and Rashvand, P. (2018). Diseases diagnosis using fuzzy logic methods: A systematic and meta-analysis review. *Comput.. Methods Prog. Biomed.* 161, 145–172. doi: 10.1016/j.cmpb.2018.04.013

Aldridge, A. A., and Roesch, S. C. (2008). Developing coping typologies of minority adolescents: a latent profile analysis. *J. Adolesc.* 31, 499–517. doi: 10.1016/j.adolescence.2007.08.005

Almeida, E., Kosina, P., and Gama, J. (2013). "Random rules from data streams," in *Proceedings of the 28th Annual ACM Symposium on Applied Computing (ACM)* (Coimbra),813–4.

Armstrong, R. A., and Wood, L. (1994). The identification of pathological subtypes of Alzheimer's disease using cluster analysis. *Acta Neuropathol.* 88, 60–66. doi: 10.1007/BF00294360

Cappa, A., Ciccarelli, N., Baldonero, E., Martelli, M., and Silveri, M. C. (2014). Posterior ad-type pathology: cognitive subtypes emerging from a cluster analysis. *Behav. Neurol.* 2014:259358. doi: 10.1155/2014/259358

Chen, C.-H. (2014). A hybrid intelligent model of analyzing clinical breast cancer data using clustering techniques with feature selection. *Appl. Soft. Comput.* 20, 4–14. doi: 10.1016/j.asoc.2013.10.024

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehe'ricy, S., Habert, M.-O., et al. (2011). Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *Neuroimage* 56, 766–81 329. doi: 10.1016/j.neuroimage.2010.06.013

Eick, C. F., Zeidat, N., and Zhao, Z. (2004). "Supervised clustering-algorithms and benefits," in *Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on (IEEE)* (Boca Raton, FL), 774–776.

Escudero, J., Zajicek, J. P., and Ifeachor, E. (2011). Early detection and characterization of Alzheimer's disease in clinical scenarios using Bioprofile concepts and K-means. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2011, 6470–3. doi: 10.1109/IEMBS.2011.6091597

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics* 21, 768–9.

Gamberger, D., Lavrac, N., Srivatsa, S., Tanzi, R. E., and Doraiswamy, P. M. (2017). Identification of clusters of rapid and slow decliners among subjects at risk for Alzheimer's disease. *Sci. Rep.* 7:6763. doi: 10.1038/s41598-017-06624-y

Gamberger, D., Ženko, B., Mitelpunkt, A., Lavrač, N., and The Alzheimer's Disease Neuroimaging Initiative (2016b). Homogeneous clusters of Alzheimer's disease patient population. *Biomed. Eng. Online* 15:78. doi: 10.1186/s12938-016-0183-0

Gamberger, D., Ženko, B., Mitelpunkt, A., Shachar, N., and Lavrac, N. (2016a). Clusters of male and female Alzheimer's disease patients in the Alzheimer's disease neuroimaging initiative (ADNI) database. *Brain Inform.* 3, 169–179. doi: 10.1007/s40708-016-0035-5

Hwang, J., Kim, C. M., Jeon, S., Lee, J. M., Hong, Y. J., Roh, J. H., et al. (2016). Prediction of Alzheimer's disease pathophysiology based on cortical thickness patterns. *Alzheimers Dement.* 2, 58–67. doi: 10.1016/j.dadm.2015.11.008

Kauer-Sant'Anna, M., Kapczinski, F., Andreazza, A. C., Bond, D. J., Lam, L., Young, T., et al. (2009). Brain-derived neurotrophic factor and inflammatory markers in patients with early- vs. late-stage bipolar disorder. *Int. J. Neuropsychopharmacol.* 12, 447–458. doi: 10.1017/S1461145708009310

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artif. Intell. Med.* 23, 89–109. doi: 10.1016/S0933-3657(01)00077-X

Lance, G. N., and Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. *Comput. J.* 9, 373–380. doi: 10.1093/comjnl/9.4.373

Li, X., and Zhu, F. (2013). On clustering algorithms for biological data. *Engineering* 5:549. doi: 10.4236/eng.2013.510B113

Mauri, M., Sinforiani, E., Zucchella, C., Cuzzoni, M. G., and Bono, G. (2012). Progression to dementia in a population with amnestic mild cognitive impairment: clinical variables associated with conversion. *Funct. Neurol.* 27, 49–54. Retrieved from: https://www.functionalneurology.com/common/php/portiere.php?ID=7e9d18f89fd466375df486c577ef3819

Mazure, C. M., and Swendsen, J. (2016). Sex differences in Alzheimer's disease and other dementias. *Lancet Neurol.* 15, 451–452. doi: 10.1016/S1474-4422(16)00067-3

McCurry, S. M., Logsdon, R. G., Teri, L., Gibbons, L. E., Kukull, W. A., Bowen, J. D., et al. (1999). Characteristics of sleep disturbance in community-dwelling Alzheimer's disease patients. *J. Geriatr. Psychiatry Neurol.* 12, 53–59. doi: 10.1177/089198879901200203

Mitchell, J. E., Crosby, R. D., Wonderlich, S. A., Hill, L., Le Grange, D., Powers, P., et al. (2007). Latent profile analysis of a cohort of patients with eating disorders not otherwise specified. *Int. J. Eat. Disord.* 40, S95–S98. doi: 10.1002/eat.20459

Monacelli, F., Borghi, R., Cammarata, S., Nencioni, A., Piccini, A., Tabaton, M., et al. (2015). Amnestic mild cognitive impairment and conversion to Alzheimer's disease: insulin resistance and glycoxidation as early biomarker clusters. *J. Alzheimers Dis.* 45, 89–95. doi: 10.3233/JAD-142511

Mueller, S., Schuff, N., and Weiner, M. (2006). Evaluation of treatment effects in Alzheimer's and other neurodegenerative diseases by MRI and MRS. *NMR Biomed.* 19, 655–668. doi: 10.1002/nbm.1062

Nikas, J. B., and Low, W. C. (2011). Application of clustering analyses to the diagnosis of Huntington's disease in mice and other diseases with well-defined group boundaries. *Comput. Methods Programs Biomed.* 104, e133–e147. doi: 10.1016/j.cmpb.2011.03.004

Nilashi, M., Ibrahim, O., and Ahani, A. (2016). Accuracy improvement for predicting Parkinson's disease progression. *Sci. Rep.* 6:34181. doi: 10.1038/srep34181

Nithya, N., Duraiswamy, K., and Gomathy, P. (2013). A survey on clustering techniques in medical diagnosis. *Int. J. Comput. Sci. Trends Technol.* 1, 17–23. Retrieved from: http://www.ijcstjournal.org/volume-1/issue-2/IJCST-V1I2P4.pdf

Noh, Y., Jeon, S., Lee, J. M., Seo, S. W., Kim, G. H., Cho, H., et al. (2014). Anatomical heterogeneity of Alzheimer's disease based on cortical thickness on MRIs. *Neurology* 83, 1936–1944. doi: 10.1212/WNL.0000000000001003

Nugent, R., and Meila, M. (2010). "An overview of clustering applied to molecular biology," in *Statistical Methods in Molecular Biology,* eds H. Bang, X. K. Zhou, H. L. van Epps, and M. Mazumdar (Springer), 369–404.

Paul, R., and Hoque, A. S. M. L. (2010). "Clustering medical data to predict the likelihood of diseases," in *Digital Information Management (ICDIM), 2010 Fifth International Conference on (IEEE)* (Thunder Bay, ON), 44–9.

Pfahringer, B., Holmes, G., and Wang, C. (2004). "Millions of random rules," in *Proceedings of the Workshop on Advances in Inductive Rule Learning, 15th European Conference on Machine Learning (ECML),* (Pisa), 365.

Polat, K. (2012). Classification of Parkinson's disease using feature weighting method on the basis of fuzzy 366 c-means clustering. *Int. J. Syst. Sci.* 43, 597–609. doi: 10.1080/00207721.2011.581395

Price, C. C., Tanner, J. J., Schmalfuss, I. M., Brumback, B., Heilman, K. M., and Libon, D. J. (2015). Dissociating statistically-determined Alzheimer's disease/vascular dementia neuropsychological syndromes using white and gray neuroradiological parameters. *J. Alzheimers Dis.* 48, 833–847. doi: 10.3233/JAD-150407

Racine, A. M., Nicholas, C. R., Clark, L. R., Koscik, R. L., Okonkwo, O. C., Hillmer, A. T., et al. (2015). Alzheimer's disease biomarker-based clusters predict amyloid accumulation and cognitive decline in a preclinical cohort: findings from the Wisconsin registry for Alzheimer's prevention (wrap). *Alzheimers Dement.* 11, P47–P49. doi: 10.1016/j.jalz.2015.06.084

Ryu, S.-Y., Kwon, M. J., Lee, S.-B., Yang, D. W., Kim, T.-W., Song, I.-U., et al. (2010). Measurement of precuneal and hippocampal volumes using magnetic resonance volumetry in Alzheimer's disease. *J. Clin. Neurol.* 6, 196–203. doi: 10.3988/jcn.2010.6.4.196

Seppi, K., and Poewe, W. (2010). Brain magnetic resonance imaging techniques in the diagnosis of Parkinsonian syndromes. *Neuroimaging Clin.* 20, 29–55. doi: 10.1016/j.nic.2009.08.016

Srivastav, V., Issenhuth, T., Kadkhodamohammadi, A., de Mathelin, M., Gangi, A., and Padoy, N. (2018). MVOR: A multi-view RGB-D operating room dataset for 2D and 3D human pose estimation. *arXiv[Preprint].arXiv:1808.08180.* Retrieved from: https://arxiv.org/pdf/1808.08180.pdf

Tosto, G., Monsell, S. E., Hawes, S. E., Bruno, G., and Mayeux, R. (2016). Progression of extrapyramidal signs in Alzheimer's disease: clinical and neuropathological correlates. *J. Alzheimers Dis.* 49, 1085–1093. doi: 10.3233/JAD-150244

Trevithick, L., Painter, J., and Keown, P. (2015). Mental health clustering and diagnosis in psychiatric in-patients. *BJPsych Bull.* 39, 119–123. doi: 10.1192/pb.bp.114.047043

Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., and Padoy, N. (2017). Endonet: 386 388 387 A deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans. Med. Imaging* 36, 86–97. doi: 10.1109/TMI.2016.2593957

Viña, J., and Lloret, A. (2010). Why women have more Alzheimer's disease than men: gender and mitochondrial toxicity of amyloid-beta peptide. *J. Alzheimers Dis.* 20, S527–S533. doi: 10.3233/JAD-2010-100501

Vogt, W., and Nagel, D. (1992). Cluster analysis in diagnosis. *Clin. Chem.* 38, 182–198.

Wiwie, C., Baumbach, J., and Rottger, R. (2015). Comparing the performance of biomedical clustering methods. *Nat. Methods* 12:1033. doi: 10.1038/nmeth.3583

Wu, Y., Duan, H., and Du, S. (2015). Multiple fuzzy c-means clustering algorithm in medical diagnosis. *Technol. Health Care* 23, S519–S527. doi: 10.3233/THC-150989

Yilmaz, N., Inan, O., and Uzer, M. S. (2014). A new data preparation method based on clustering algorithms for diagnosis systems of heart and diabetes diseases. *J. Med. Syst.* 38:48. doi: 10.1007/s10916-014-0048-7

**Conflict of Interest Statement:** 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.

frontiers in Computational Neuroscience

METHODS
published: 20 August 2019
doi: 10.3389/fncom.2019.00053

# Multi-method Fusion of Cross-Subject Emotion Recognition Based on High-Dimensional EEG Features

*Fu Yang*[1,2], *Xingcong Zhao*[1,2], *Wenge Jiang*[1,2], *Pengfei Gao*[1,2] *and Guangyuan Liu*[1,2]*

[1] College of Electronic Information and Engineering, Southwest University, Chongqing, China, [2] Chongqing Key Laboratory of Nonlinear Circuit and Intelligent Information Processing, Chongqing, China

Emotion recognition using electroencephalogram (EEG) signals has attracted significant research attention. However, it is difficult to improve the emotional recognition effect across subjects. In response to this difficulty, in this study, multiple features were extracted for the formation of high-dimensional features. Based on the high-dimensional features, an effective method for cross-subject emotion recognition was then developed, which integrated the significance test/sequential backward selection and the support vector machine (ST-SBSSVM). The effectiveness of the ST-SBSSVM was validated on a dataset for emotion analysis using physiological signals (DEAP) and the SJTU Emotion EEG Dataset (SEED). With respect to high-dimensional features, the ST-SBSSVM average improved the accuracy of cross-subject emotion recognition by 12.4% on the DEAP and 26.5% on the SEED when compared with common emotion recognition methods. The recognition accuracy obtained using ST-SBSSVM was as high as that obtained using sequential backward selection (SBS) on the DEAP dataset. However, on the SEED dataset, the recognition accuracy increased by ∼6% using ST-SBSSVM from that using the SBS. Using the ST-SBSSVM, ∼97% (DEAP) and 91% (SEED) of the program runtime was eliminated when compared with the SBS. Compared with recent similar works, the method developed in this study for emotion recognition across all subjects was found to be effective, and its accuracy was 72% (DEAP) and 89% (SEED).

**Keywords: EEG, emotion recognition, cross-subject, multi-method fusion, high-dimensional features**

OPEN ACCESS

***Edited by:***
*Abdelmalik Moujahid,*
*University of the Basque Country, Spain*

***Reviewed by:***
*Houtan Jebelli,*
*Pennsylvania State University, United States*
*Dat Tran,*
*University of Canberra, Australia*
*Xiang Li,*
*National Supercomputer Center, China*

***Correspondence:**
*Guangyuan Liu*
*liugy@swu.edu.cn*

***Received:*** *20 May 2019*
***Accepted:*** *19 July 2019*
***Published:*** *20 August 2019*

***Citation:***
*Yang F, Zhao X, Jiang W, Gao P and Liu G (2019) Multi-method Fusion of Cross-Subject Emotion Recognition Based on High-Dimensional EEG Features. Front. Comput. Neurosci. 13:53. doi: 10.3389/fncom.2019.00053*

## 1. INTRODUCTION

Emotion is essential to humans, as it contributes to the communication between people and plays a significant role in rational and intelligent behavior (Picard et al., 2001; Nie et al., 2011), which is critical to several aspects of daily life. Therefore, research on emotion recognition is necessary. It is difficult to define and classify emotion due to the complex nature and genesis of emotion (Ashforth and Humphrey, 1995; Horlings et al., 2008; Hwang et al., 2018). To classify and represent emotion, several models have been proposed. Moreover, there are two main models. The first assumes that all emotions can comprise primary emotions, similar to how all colors can comprise primary colors. Plutchik (1962) related eight basic emotions to evolutionarily valuable properties, and then reported the following primary emotions: anger, fear, sadness, disgust, surprise, curiosity, acceptance, and joy. Ekman (Power and Dalgleish, 1999; Horlings et al., 2008) reported other

emotions as a basis set and found that these primary emotions, in addition to their expressions, are universal. The Ekman list of primary emotions is as follows: anger, fear, sadness, happiness, disgust, and surprise. The second main model is composed of multiple dimensions, and each emotion is on a multi-dimensional scale. Russell (1980) divided human emotions into two dimensions: arousal and valence. Arousal represents the strength of the emotion with respect to arousal and relaxation and valence represents positive and negative levels. Among several emotional models, the Russell model (Russell, 1980) is generally adopted, in which two dimensions are represented by a vertical arousal axis and horizontal valence axis (Choi and Kim, 2018). In both dimensions of emotion, the ability to measure valence levels is essential, as the valence level is a more critical dimension for distinguishing between positive emotions (e.g., excitement, happiness, contentment, or satisfaction) and negative emotions (e.g., fear, anger, frustration, mental stress, or depression; Hwang et al., 2018). It is necessary to effectively classify and identify positive and negative emotions. For example, the accurate identification of the mental stress (a negative emotion) or emotional state of construction workers can help reduce construction hazards and improve production efficiency (Chen and Lin, 2016; Jebelli et al., 2018a). The focus of this study was on improving the classification accuracy of positive and negative emotions. In daily human life, communication and decision-making are influenced by emotional behavior. For many years, the brain-computer interface (BCI) has been a critical topic with respect to biomedical engineering research, allowing for the use of brain waves to control equipment (Nijboer et al., 2009). To achieve accurate and smooth interactions, computers and robots should be able to analyze emotions (Pessoa and Adolphs, 2010; Zheng et al., 2019). Researchers in the fields of psychology, biology, and neuroscience have directed significant attention toward emotional research. Emotional research has a preliminary development trend in the field of computer science, such as task workload assessments and operator vigilance (Shi and Lu, 2013; Zheng and Lu, 2017b). The automatic emotion recognition system simplifies the computer interface and renders it more convenient, more efficient, and more user-friendly. Human emotion recognition can be studied using questionnaires (Mucci et al., 2015; Jebelli et al., 2019), facial images, gestures, speech signals and other physiological signals (Jerritta et al., 2011). However, the questionnaire method interfered with this study. In addition, it exhibited a significant deviation and yielded inconsistent results (Jebelli et al., 2019). There was an ambiguity with respect to emotion recognition from facial images, gestures, or speech signals, as real emotions can be mimicked. To overcome the ambiguity, an electroencephalogram (EEG) could be employed for emotion recognition, as it is more accurate and more objective than emotional evaluation based on facial image and gesture-based methods (Ahern and Schwartz, 1985). Therefore, EEG has attracted significant research attention. Moreover, EEG signals can be used to effectively identify different emotions (Sammler et al., 2007; Mathersul et al., 2008; Knyazev et al., 2010; Bajaj and Pachori, 2014). For effective medical care, it is essential to consider emotional states (Doukas and Maglogiannis, 2008; Petrantonakis and Hadjileontiadis, 2011). Due to the objectivity of physiological data and the ability to model learning principles from heterogeneous features to emotional classifiers, the use of machine learning methods for the analysis of EEG signals has attracted significant attention in the field of human emotion recognition. To improve the satisfaction and reliability of the people who interact and collaborate with machines and robots, a smart human-machine (HM) system that can accurately interpret human communication capabilities is required (Koelstra et al., 2011). Human intentions and commands mostly convey emotions in a linguistic or non-verbal manner; thus, the accurate response to human emotional behavior is critical to the realization of machine and computer adaptation (Zeng et al., 2008; Fanelli et al., 2010). At this stage, the majority of HM systems cannot accurately recognize emotional cues. Emotional classifiers were developed based on facial/sound expressions or physiological signals (Hanjalic and Xu, 2005; Kim and André, 2008). Emotion classifiers can provide temporal predictions of specific emotional states. Emotional recognition requires appropriate signal preprocessing techniques, feature extraction, and machine learning-based classifiers to carry out automatic classification. An EEG, which captures brain waves, can effectively distinguish between emotions. The EEG directly detects brain waves from the central nervous system activities (i.e., brain activities), whereas other responses (e.g., EDA, HR, and BVP) are based on peripheral nervous system activities (Zhai et al., 2005; Chanel et al., 2011; Hwang et al., 2018). In particular, central nervous system activities are related to several aspects of emotions (e.g., from displeasure to pleasure, and from relaxation to excitement); however, the peripheral nervous system activities are only associated with arousal and relaxation (Zhai et al., 2005; Chanel et al., 2011). Therefore, the EEG can provide more detailed information on emotional states than other methods (Takahashi et al., 2004; Lee and Hsieh, 2014; Liu and Sourina, 2014; Hou et al., 2015). Moreover, EEG-based emotion recognition has a greater potential with respect to research than facial and speech-based methods, given that internal nerve fluctuations cannot be deliberately masked or controlled. However, the improvement of the performance of cross-subject emotional recognition has been the focus of several studies, including this study. In previous studies, cross-subject emotion recognition was difficult to achieve when compared with intra-subject emotion recognition. In Kim (2007), the method of bimodal data fusion was investigated, and a linear discriminant analysis (LDA) was conducted to classify emotions. The best recognition accuracy across the three subjects was 55%, which was significantly lower than the 92% achieved using the intra-subject emotion recognition method (Kim, 2007). In Zhu et al. (2015), the authors employed differential entropy (DE) as features, and a linear dynamic system (LDS) was applied to carry out feature smoothing. The average cross-subject classification accuracy was 64.82%, which was significantly lower than the 90.97% of the intra-subject emotion recognition method (Zhu et al., 2015). In Zhuang et al. (2017), a method for feature extraction and emotion recognition based on empirical mode decomposition (EMD) was introduced. Using EMD, the EEG signals were automatically decomposed into intrinsic mode functions (IMFs). Based on the results, IMF1 demonstrated

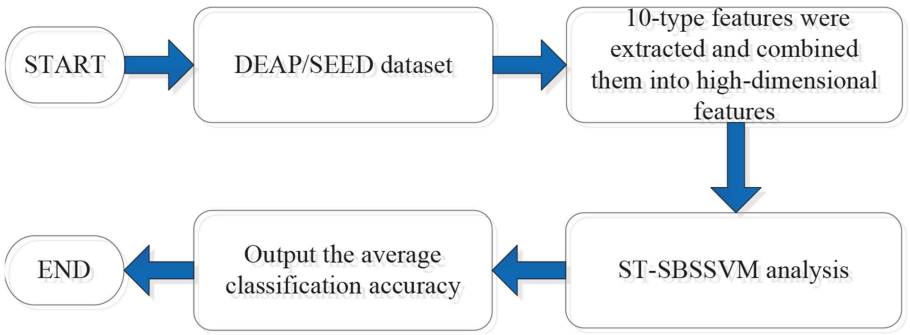


**FIGURE 1 |** Analysis process.

the best performance, which was 70.41% for valence (Zhuang et al., 2017). In Candra et al. (2015), an accuracy of 65% was achieved for valence and arousal using the wavelet entropy of signal segments with periods of 3–12 s. In Mert and Akan (2018), the advanced properties of EMD and its multivariate extension (MEMD) for emotion recognition were investigated. The multichannel IMFs extracted by MEMD were analyzed using various time- and frequency-domain parameters such as the power ratio, power spectral density, and entropy. Moreover, Hjorth parameters and correlation were employed as features of the valence and arousal scales of the participants. The proposed method yielded an accuracy of 72.87% for high/low valences (Mert and Akan, 2018). In Zheng and Lu (2017a), deep belief networks (DBNs) were trained using differential entropy features extracted from multichannel EEG data, and the average accuracy was 86.08%. In Yin et al. (2017), cross-subject EEG feature selection for emotion recognition was carried out using transfer recursive feature elimination. The classification accuracy was 78.75% in the valence dimension, which was higher than those reported in several studies that used the same database. However, from the calculation times of all the classifiers, it was found that the accuracy of the *t*-test/recursive feature extraction (T-RFE) increased at the expense of the training time. In Li et al. (2018), 18 linear and non-linear EEG features were extracted. In addition, the support vector machine (SVM) method and the leave-one-subject-out verification strategy were used to evaluate the recognition performance. With the automatic feature selection method, the recognition accuracy rate using the dataset for emotion analysis using physiological signals (DEAP) was a maximum of 59.06%, and the recognition accuracy using the SEED dataset was a maximum of 83.33% (Li et al., 2018). In Gupta et al. (2018), the aim of the study was to comprehensively investigate the channel specificity of EEG signals and provide an effective emotion recognition method based on the flexible analytic wavelet transform (FAWT). The average classification accuracy obtained using this method was 90.48% for positive/neutral/negative (SEED) emotion classification, and 79.99% for high valence (HV)/low valence (LV) emotion classification using EEG signals (Gupta et al., 2018). In Li et al. (2019), the accuracy of multisource supervised STM (MS-S-STM) for emotion recognition accuracy was 88.92%, and the multisource semi-supervised selective transfer machine (STM) (MS-semi-STM) experimental data was used in a transmissive manner, with a maximum accuracy of 91.31%. The methods of emotion recognition across subjects, as employed in the previous studies, require improvements. A method for improving the accuracy of emotion classification is therefore necessary, which requires only a small computational load when applied to the analysis of high-dimensional features. In this study, multiple types of features were extracted. In addition, a two-category emotion recognition method across subjects is proposed. In particular, 10 types of linear and non-linear EEG features were first extracted, and then combined into high-dimensional features. With respect to high-dimensional features, a method for improving the emotion recognition performance across subjects based on high-dimensional features was proposed. Moreover, the significant test/sequential backward selection/support vector machine (ST-SBSSVM) fusion method was proposed and then used to identify and classify the high-dimensional EEG features of the cross-subject emotions.

## 2. MATERIALS AND METHODS

**Figure 1** presents the analysis process in this study. First, 10 types of high-dimensional features were extracted from both the DEAP and SEED. The features were then combined into high-dimensional features, as follows:

$$DEAP, 1280(trials, rows) \times 320(features, cols) \quad (1)$$

$$SEED, 450(trials, rows) \times 620(features, cols) \quad (2)$$

For further details, refer to section 2.2. Furthermore, the proposed method (ST-SBSSVM) was used to analyze the high-dimensional features and output the classification and recognition accuracy of positive and negative emotions.

### 2.1. DEAP Dataset and SEED Dataset

Two publicly accessible datasets were employed for the analysis, namely, the DEAP and SEED. The DEAP dataset (Koelstra et al., 2011) consisted of 32 subjects. Each subject was exposed to 40 1-min long music videos as emotional stimuli

**TABLE 1 |** Structure of the DEAP dataset.

| Array name | Array shape | Array contents |
|---|---|---|
| Data | $40 \times 40 \times 8,064$ | (*Videos*/*trials*) × *channels* × *data* (128 *Hz* × 63 *s*) |
| Labels | $40 \times 4$ | (*videos*/*trials*) × *labels* (*Valence*, *arousal*, *dominance*, *liking*) |

**TABLE 2 |** Structure of the SEED dataset.

| Array name | Array shape | Array contents |
|---|---|---|
| Data | $3 \times 15 \times 62 \times 48,000$ | (*Experiments*) × (*Videos*/*trials*) × *channels* × *data* (200 *Hz* × 240 *s*) |
| Labels | $3 \times 15 \times 3$ | (*Experiments*) × (*videos*/*trials*) × *labels* (*positive*, *neutral*, *negative*) |

while their physiological signals were recorded. The resulting dataset includes 32 channels of EEG signals, four-channel electrooculography (EOG), four-channel electromyography (EMG), respiration, plethysmography, galvanic skin response and body temperature. Each subject underwent 40 EEG trials, each of which corresponded to an emotion triggered by a music video. After watching each video, the participants were asked to score their real emotions on a five-level scale: (1) valence (related to the level of pleasure), (2) arousal (related to the level of excitement), (3) dominance (related to control), (4) like (related to preference), and (5) familiarity (related to the awareness of stimuli). The score ranged from 1 (weakest) to 9 (strongest), with the exception of familiarity, which ranged from 1 to 5. The EEG signal was recorded using Biosemi ActiveTwo devices at a sampling frequency of 512 Hz and down-sampling frequency of 128 Hz. The data structure of DEAP is shown in **Table 1**. The SEED (Zheng and Lu, 2017a) consisted of 15 subjects. Movie clips were selected to induce (1) positive emotions, (2) neutral emotions, and (3) negative emotions; each of which were distributed over five segments of each movie. All subjects underwent three EEG recordings, with two consecutive recordings conducted at a two-week interval. At each stage, each subject was exposed to 15 movie clips, each of which was ∼4 min long, to induce specific emotions. The same 15-segment movie clip was used in all three recording sessions. The resulting data contained 15 EEG trials. Each subject underwent 15 trials with 5 trials per emotion. The EEG signals were recorded using a 62-channel NeuroScan electric source imaging (ESI) device at a sampling rate of 1,000 Hz and down-sampling rate of 200 Hz. The data structure of SEED is shown in **Table 2** (the duration of the SEED videos varied: each video was about 4 min = 240 s; thus, the data were about $200Hz \times 240s = 48,000$). In this study, only experiments with positive emotions and negative emotions were carried out to evaluate the ability of the proposed method to distinguish between these two emotions. For consistency with the DEAP, 1 min of data extracted at the middle of each trial was employed using the SEED.

**TABLE 3 |** Ten-type EEG features.

| Feature type | Extracted features | | |
|---|---|---|---|
| The linear features | 1. Hjorth activity | 2. Hjorth mobility | 3. Hjorth complexity |
| | 4. The standard deviation | 5. PSD-Alpha | 6. PSD-Beta |
| | 7. PSD-Gamma | 8. PSD-Theta | |
| The non-linear features | 9. Sample entropy | 10. Wavelet entropy | |

## 2.2. Data Processing

### 2.2.1. Data Preprocessing

The EEG signal considered in this study was a neurophysiological signal with a high dimensionality, redundancy, and noise. After the EEG data were collected, the original data were pre-processed, i.e., the removal of EOG, EMG artifacts, and down-sampling; to reduce the computational overhead of feature extraction. For the DEAP, the default pre-processing technique was as follows: (1) the data was down-sampled to 128 Hz; (2) the EOG artifacts were removed, as achieved in Koelstra et al. (2011); (3) a bandpass filter with a throughput frequency range of 4.0–45.0 Hz was applied; (4) the data were averaged to the common reference; and (5) the data were segmented into 60-s trials and a 3-second pre-trial baseline. For the SEED, the default preprocessing technique was applied as follows: (1) the data was down-sampled to 200 Hz; (2) a bandpass filter with a throughput frequency range of 0–75 Hz was applied; and (3) the EEG segments corresponding to the duration of each movie were extracted. Prior to the extraction of the power spectral density (PSD) features, four rhythms were extracted, namely, theta (3–7 Hz), alpha (8–13 Hz), beta (14–29 Hz), and gamma (30–47 Hz) (Koelstra et al., 2011). Other features were extracted on the data preprocessed by the dataset.

### 2.2.2. Label Processing

For label processing using the DEAP, the subjects were divided into two categories according to the corresponding scores of the subjects with respect to valence. A score higher than 5 was set as 1, which represented positive emotions; and a score below 5 was set as 1, which represented negative emotions. In the SEED, the trials were divided into positive emotions, neutral emotions, and negative emotions. However, for consistency with the DEAP, only positive and negative emotion samples were investigated using the SEED. Moreover, binary classification tasks were employed to carry out emotional recognition across the subjects.

## 2.3. Feature Extraction

Ten types of linear and non-linear features were extracted, as shown in **Table 3**. Several features [Hjorth activity, Hjorth mobility, Hjorth complexity, standard deviation, sample entropy (SampEn), and wavelet entropy (WE)] were directly extracted from the dataset pre-processed EEG signals. The extraction processes of the remaining features (the four PSD frequency domain features) were divided into two steps. First, four types

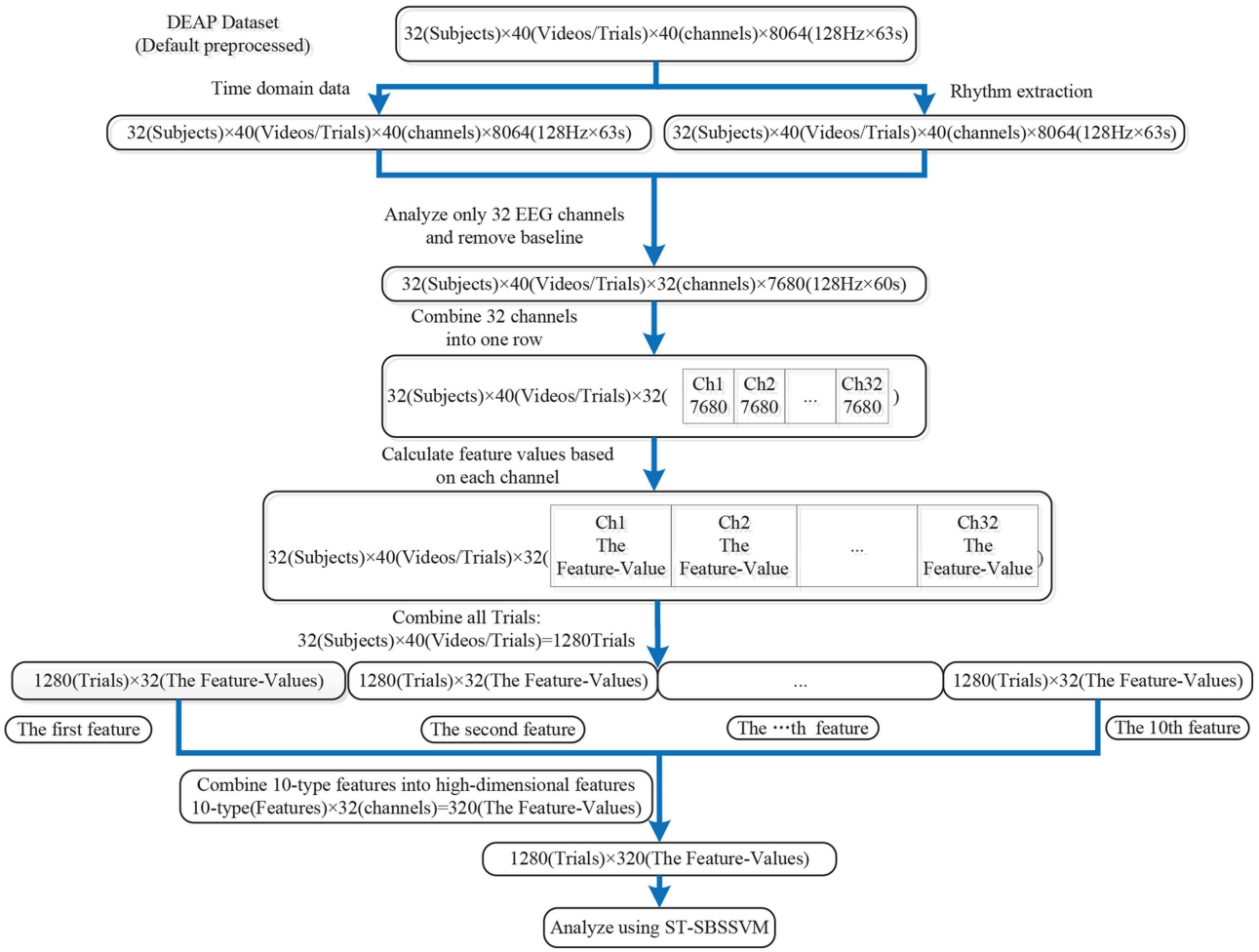


**FIGURE 2 |** DEAP data analysis and feature extraction process.

of rhythms were extracted from the EEG signals pre-processed using the dataset, and the PSD features were then extracted from the four rhythms. The detailed analysis of the data and feature extraction is shown in **Figures 2**, **3**.

### 2.3.1. The Linear Feature

Hjorth parameters were indicators of statistical properties used in signal processing in the time domain, as introduced by Hjorth (1970). The parameters are as follows: activity, mobility, and complexity. They were commonly used in the analysis of electroencephalography signals for feature extraction. The parameters are normalized slope descriptors (NSDs) used in EEGs. The standard deviation feature was the standard value of the EEG time-series signal. The four PSD Features were extracted as follows: PSD-alpha was extracted from the alpha rhythm, PSD-beta was extracted from the beta rhythm, PSD-gamma was extracted from the gamma rhythm, and PSD-theta was extracted from the theta rhythm. The power spectrum $S_{xx}(f)$ of a time series $x(t)$ describes the power distribution with respect to the frequency components that compose that signal (Fanelli et al., 2010). According to Fourier analysis, any physical signal can be decomposed into several discrete frequencies, or a spectrum of frequencies over a continuous range. The statistical average of a certain signal or signal type (including noise), as analyzed with respect to its frequency content, is referred to as its spectrum. When the energy of the signal is concentrated around a finite time interval, especially if its total energy is finite, the energy spectral density can be computed. Moreover, the PSD (power spectrum) is more commonly used, which applies to signals existing over a sufficiently large time period (especially in relation to the duration of a measurement) that can be considered as an infinite time interval. The PSD refers to the spectral energy distribution per unit time, given that the total energy of such a signal over an infinite time interval would generally be infinite. The summation or integration of the spectral components yield the total power (for a physical process) or variance (in a statistical process), which correspond to the values that are obtained by integrating $x^2(t)$ over the time domain, as dictated by Parseval's theorem (Snowball, 2005). For continuous signals over a quasi-infinite time interval, such as stationary processes, the PSD)

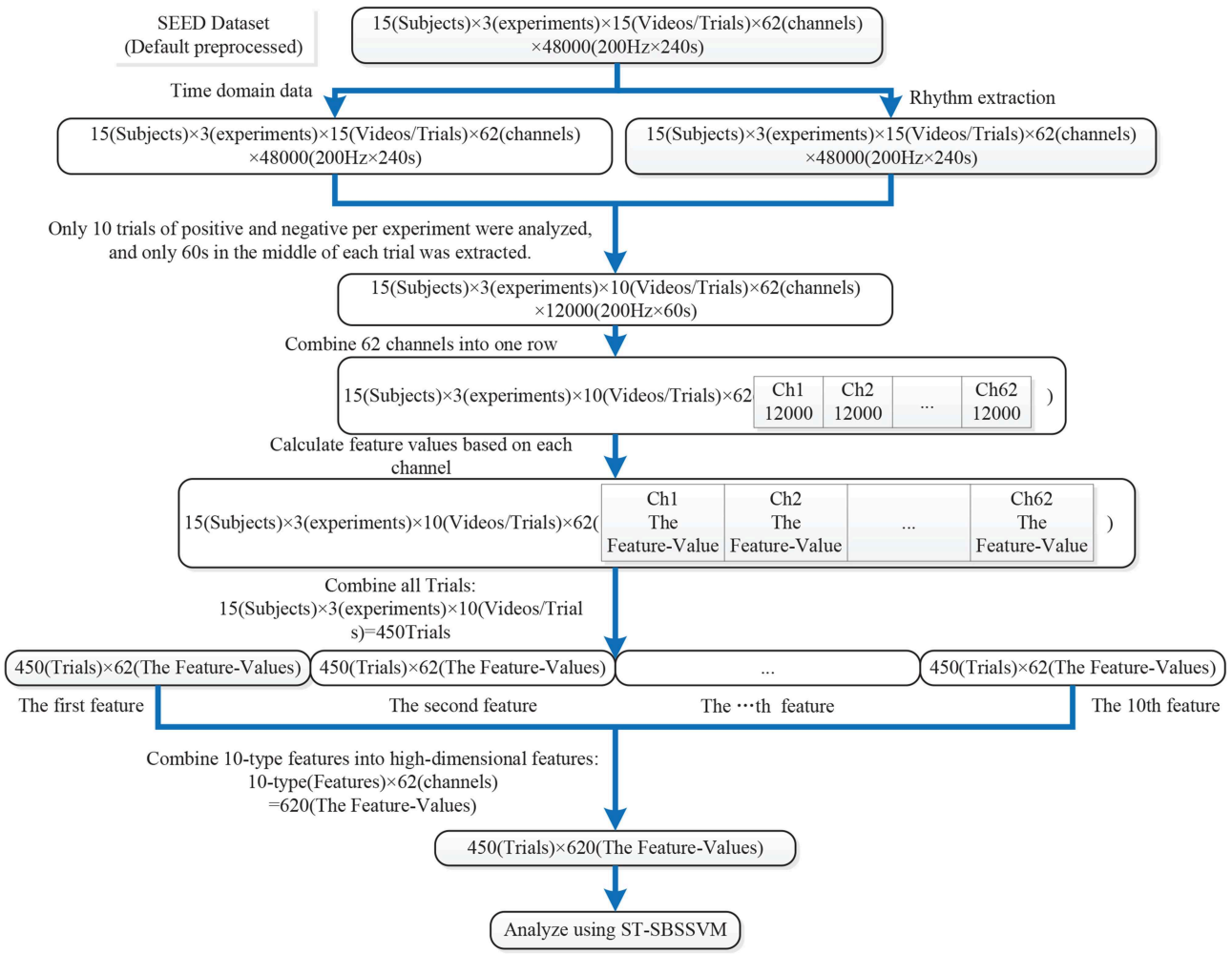


**FIGURE 3 |** SEED data analysis and feature extraction process.

should be defined, which describes the power distribution of a signal or time-series with respect to frequency.

### 2.3.2. The Non-linear Feature

The SampEn is a modification of the approximate entropy (ApEn), and it is used for assessing the complexity of physiological time-series signals in addition to the diagnosis of diseased states (Richman and Moorman, 2000). Moreover, SampEn has two advantages over ApEn, namely, data length independence and a relatively simple implementation. Similar to ApEn, SampEn is a measure of complexity (Richman and Moorman, 2000). The Shannon entropy provides a useful criterion for the analysis and comparison of probability distributions, which can act as a measure of the information of any distribution; namely, the wavelet entropy (WE) (Blanco et al., 1998). In this study, the total WE was defined as follows:

$$S_{WT} \equiv S_{WT}(p) = \sum_{j<0} p_j \cdot In[p_j] \quad (3)$$

The WE can be used as a measure of the degree of order/disorder of the signal; thus, it can provide useful information on the underlying dynamical process associated with the signal.

## 2.4. ST-SBSSVM

The ST-SBSSVM method is a combination of the significance test, sequential backward selection, and support vector machine. In this study, the SVM based on the radial basis function (RBF) kernel was employed. The detailed fusion process is shown in **Figure 4**. Ten types of features from both public datasets were extracted, and high-dimensional features [$DEAP, 1280(trials, rows) \times 320(features, cols)$; and $SEED, 450(trials, rows) \times 620(features, cols)$] were formed. If the sequential backward selection (SBS) method was directly employed for the analysis of the high-dimensional EEG features and SVM was used to determine the accuracy of the emotion classification of each feature combination, the computational overhead would be significantly large. Therefore, a method was developed to achieve a higher emotional

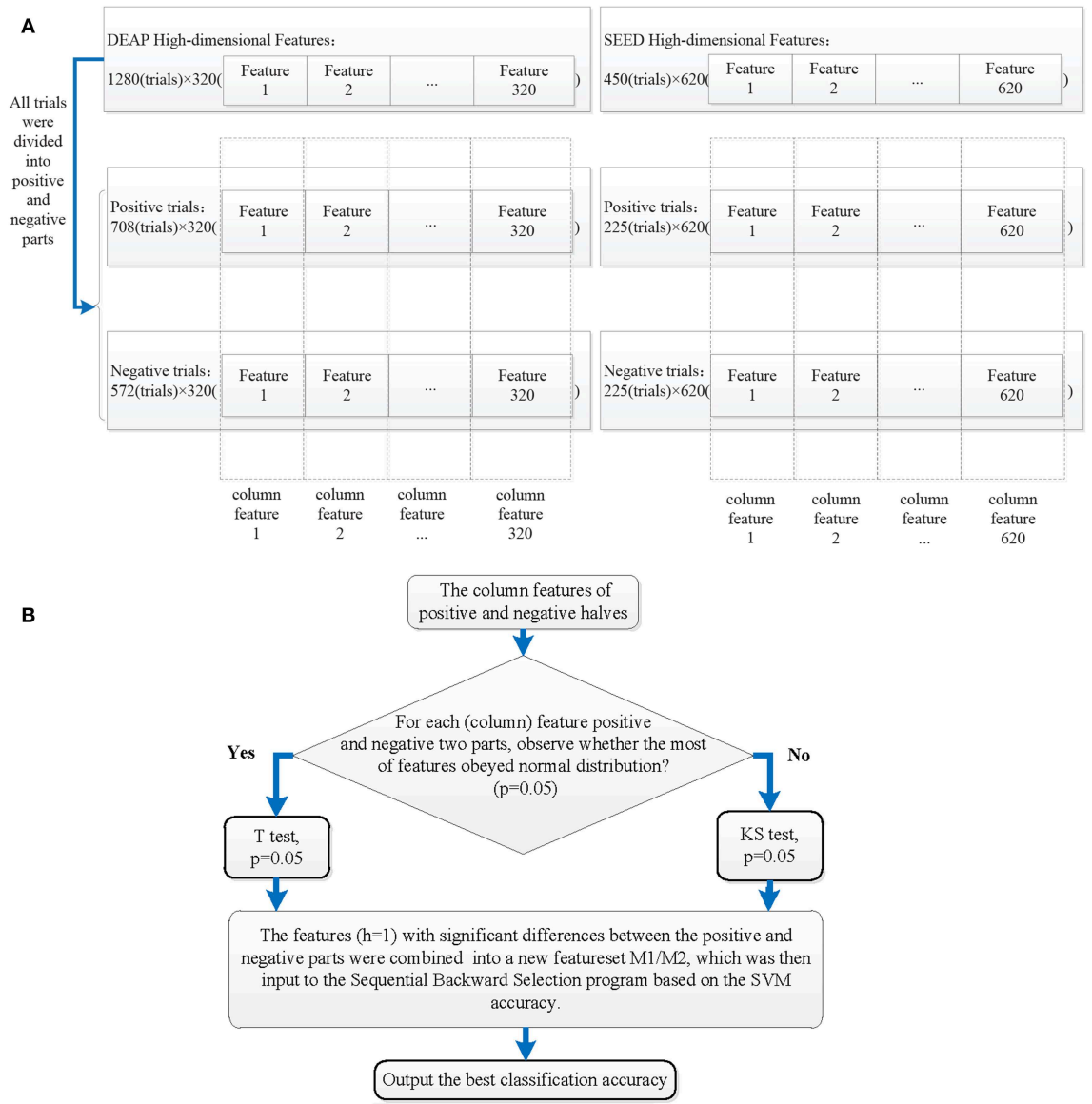


**FIGURE 4 |** Processes of ST-SBSSVM method: **(A)** generation of column features of positive and negative halves along with the division of trials (rows), and **(B)** the ST-SBSSVM analysis of all column features.

recognition accuracy across the subjects than the SBS, namely the ST-SBSSVM. Moreover, the proposed method requires a significantly lower computational overhead for the analysis of high-dimensional EEG features. As shown in step 1 of **Figure 4**, each trial (row) of $1280(trials, rows) \times 320(features, cols)$ and $450(trials, rows) \times 620(features, cols)$ was in one-to-one correspondence with the positive and negative emotion labels. In step 2, according to the labels, all the trials (rows) were divided into two parts. The objective was to simultaneously divide each column feature into two parts. In step 3, the significance test was carried out from the first column feature to the final column feature for the column features that were divided into two positive and negative parts (the last column of the DEAP feature was the 320th column, and the last column of the SEED feature was the 620th column). It was then determined whether the majority of EEG column features, which were divided, were in accordance with the normal distribution. If the majority of EEGs were subject to the normal distribution, the student's $t$-test ($T$-test) was used for the divided column features; otherwise, the Kolmogorov–Smirnov (KS) test was used. The corresponding column features of the positive and negative significant difference (h = 1) were then selected. In step 4, after the significance test, the high-dimensional feature set was simplified, and the following was obtained:

$$M_1 = 1280(trials, rows) \times 68(features, cols) \quad (4)$$

$$M_2 = 450(trials, rows) \times 227(features, cols) \quad (5)$$

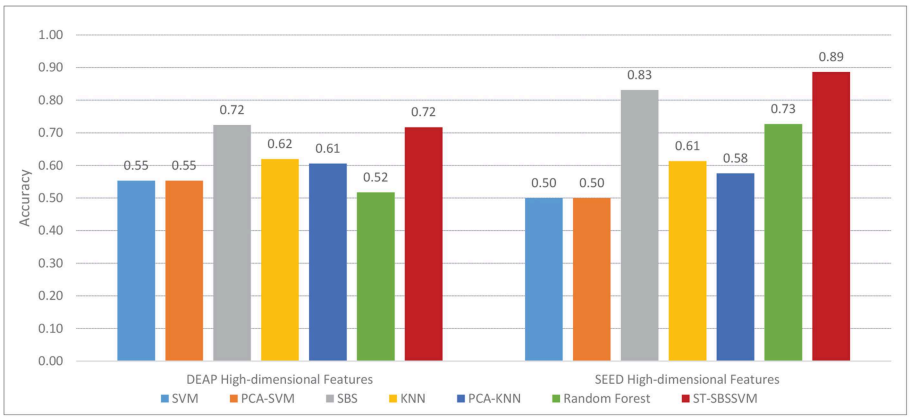


**FIGURE 5 |** Accuracy results of valence classification using DEAP and SEED.

**TABLE 4 |** Comparison of Valence classification accuracy between ST-SBSSVM and common methods.

| | Difference from ST-SBSSVM accuracy (DEAP) | Difference from ST-SBSSVM accuracy (SEED) |
|---|---|---|
| SVM | +17% | +39% |
| PCA-SVM | +17% | +39% |
| SBS | –0.42% | +6% |
| KNN | +10% | +28% |
| PCA-KNN | +11% | +31% |
| RF | +20% | +16% |
| The average difference from ST-SBSSVM accuracy | +12.4% | +26.5% |

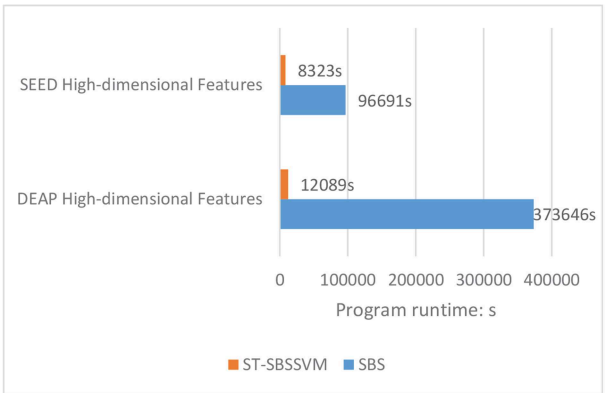


**FIGURE 6 |** Comparison of runtime between ST-SBSSVM and SBS methods.

In step 5, M1 and M2 were inputted into the SVM-based SBS program. Sequential backward selection is a process that decreases the number of features, in which a feature is repeatedly eliminated until a final feature is remaining. In this manner, all the feature combinations were separately classified by the SVM. The data was normalized prior to the use of SVM modeling for emotion classification recognition, which helped to improve the convergence rate and accuracy of the model. In the SVM-based SBS program, a "leave-one-subject-out" verification strategy was employed. During each process, the data of one subject was considered as the test set, and the data of the other subjects were considered as the training set. The feature selection was carried out on the training set, and the performance was then evaluated on the test set. This procedure was iterated until the data of each subject had been tested. Moreover, this strategy can eliminate the risk of "overfitting". In step 6, the average classification accuracy of the employed "leave-one-subject-out" verification strategy and SVM-based SBS program was outputted.

## 3. RESULTS

**Figure 5** and **Table 4** present a comparison between the valence classification recognition results of the ST-SBSSVM and those of common methods using the DEAP and SEED. For the consistency of the analysis of the two datasets, only cross-subject emotional recognition was carried out for the valence classification. The ST-SBSSVM method is an improvement of the SBS method; thus, the two methods were compared. **Figure 5** and **Table 4** present a comparison of the recognition accuracies of the two emotions. Therefore, **Figure 6** presents a runtime comparison between the ST-SBSSVM and corresponding SBS program (using the corresponding method for emotion recognition on the same computer "DELL, intel(R) Core (TM) i5-4590 CPU @ 3.30 GHz, RAM-8.00 GB"). As shown in **Figure 5** and **Table 4**, with respect to high-dimensional features, the accuracy of the ST-SBSSVM was improved by 12.4% (DEAP) and 26.5% (SEED) when compared with the common emotion recognition methods. From **Table 4**, it can be seen

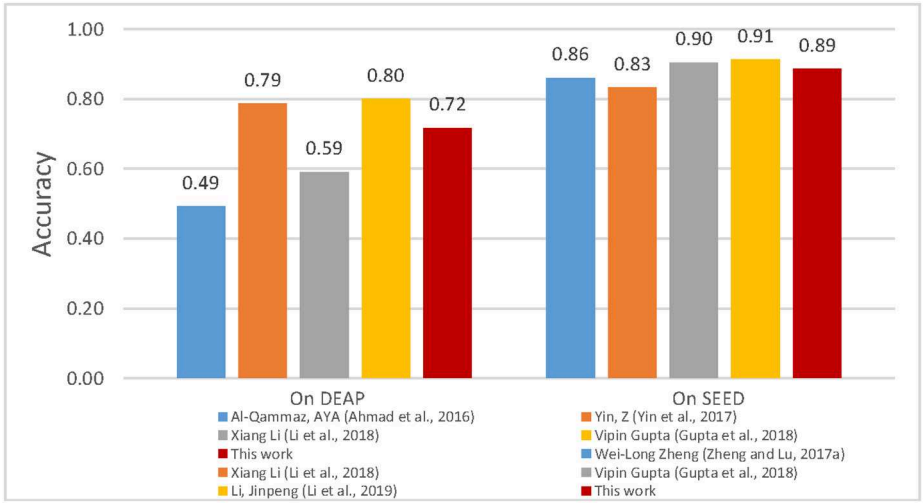


**FIGURE 7 |** Comparison between valence classification accuracies of similar studies.

that with respect to high-dimensional features, the cross-subject emotion recognition accuracy of the ST-SBSSVM decreased by 0.42% (almost unchanged) using the DEAP, and it improved by 6% using the SEED. **Figure 6** shows that the ST-SBSSVM decreased the program runtime by ∼97 and 91% when compared with the SBS method.

## 4. DISCUSSIONS

In this paper, a method that can effectively promote emotion recognition is proposed, namely, the ST-SBSSVM method. The proposed method was used to effectively analyze the high-dimensional EEG features extracted from the DEAP and SEED. The results of this study confirmed that the ST-SBSSVM method offers two advantages. First, the ST-SBSSVM can classify and identify emotions, with an improved emotion recognition accuracy. Because ST-SBSSVM performed the Significant Test by comparing the same column feature that had significant difference between positive and negative trials, a "leave-one-subject-out" verification strategy and SVM-based SBS program were then employed to carry out feature selection for those features with significant differences, and the best emotion classification accuracy was obtained. Second, the ST-SBSSVM and SBS methods exhibited similar emotion recognition results to those of the common emotion classification methods. Moreover, when using ST-SBSSVM and SBS to analyze high-dimensional features, ST-SBSSVM decreased the program runtime by ∼90% when compared with SBS. The limitations of this study were as follows. The features extracted were relatively common, and these features were not the new features that significantly promoted emotion recognition in the most recent studies. In future work, the new features combined with ST- SBSSVM will be employed to investigate emotion recognition among subjects. In recent years, several EEG devices and data technologies were developed, such as using wearable EEG devices, for the collection of data in actual working environments (Jebelli et al., 2017b, 2018b; Chen et al., 2018). High quality brainwaves can then be extracted from the data collected by wearable EEG devices (Jebelli et al., 2017a). A stress recognition framework was proposed, which can effectively process and analyze EEG data collected from wearable EEG devices in real work environments (Jebelli et al., 2018c). These new developments comprise the scope of future research. Similar works are as follows. In (Ahmad et al., 2016), the empirical results revealed that the proposed genetic algorithm (GA) and least squares support vector machine (LS-SVM) (GA-LSSVM) increased the classification accuracy to 49.22% for valence using the DEAP. In Zheng and Lu (2017a), DBNs were trained using differential entropy features extracted from multichannel EEG data. As shown in **Figure 7**, the proposed method demonstrated a good performance and its accuracy was similar to those of achieved in similar studies with respect to the emotion classification recognition of the cross subjects using the same datasets. In summary, when compared with the most recent studies, this method developed in this study was found to be effective for emotional recognition across subjects.

## 5. CONCLUSIONS

For emotion recognition, a method is proposed in this paper that can significantly enhance the two-category emotion recognition effect; with a small computational overhead when using the corresponding program to analyze high-dimensional features. In this study, 10 types of EEG features were extracted to form high-dimensional features, and the proposed ST-SBSSVM method was employed, which can rapidly analyze high-dimensional features and effectively improve the accuracy of cross-subject emotion recognition, namely the ST-SBSSVM. The results of this work revealed that ST-SBSSVM demonstrates better accuracy with respect to emotion recognition than common classification methods. Compared with the SBS method, the ST-SBSSVM exhibited a higher accuracy of emotion recognition and significantly decreased the program runtime. In comparison

to recent similar methods, the method proposed in this study is effective for emotional recognition across subjects. In summary, the proposed method can effectively promote the emotional recognition across subjects. This method can therefore contribute to the research of health therapy and intelligent human-computer interactions.

## DATA AVAILABILITY

The data set generated for this study can be provided to the corresponding author upon request.

## AUTHOR CONTRIBUTIONS

FY developed the ST-SBSSVM method, performed all the data analysis, and wrote the manuscript. XZ, WJ, PG, and GL advised data analysis and edited the manuscript.

## FUNDING

This work was supported by the National Natural Science Foundation of China (61472330, 61872301, and 61502398).

## REFERENCES

Ahern, G. L., and Schwartz, G. E. (1985). Differential lateralization for positive and negative emotion in the human brain: Eeg spectral analysis. *Neuropsychologia* 23, 745–755.

Ahmad, F. K., Al-Qammaz, A. Y. A., and Yusof, Y. (2016). Optimization of least squares support vector machine technique using genetic algorithm for electroencephalogram multi-dimensional signals. *Jurnal Teknologi* 78, 5–10. doi: 10.11113/jt.v78.8842

Ashforth, B. E., and Humphrey, R. H. (1995). Emotion in the workplace: a reappraisal. *Hum. Rel.* 48, 97–125.

Bajaj, V., and Pachori, R. B. (2014). "Human emotion classification from eeg signals using multiwavelet transform," in *2014 International Conference on Medical Biometrics* (Indore: IEEE), 125–130.

Blanco, S., Figliola, A., Quiroga, R. Q., Rosso, O., and Serrano, E. (1998). Time-frequency analysis of electroencephalogram series. iii. wavelet packets and information cost function. *Phys. Rev. E* 57:932.

Candra, H., Yuwono, M., Chai, R., Handojoseno, A., Elamvazuthi, I., Nguyen, H. T., et al. (2015). Investigation of window size in classification of eeg-emotion signal with wavelet entropy and support vector machine. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2015, 7250–7253. doi: 10.1109/EMBC.2015.7320065

Chanel, G., Rebetez, C., Bétrancourt, M., and Pun, T. (2011). Emotion assessment from physiological signals for adaptation of game difficulty. *IEEE Trans. Syst. Man Cybernet. Part A. Syst. Hum.* 41, 1052–1063. doi: 10.1109/TSMCA.2011.2116000

Chen, J., and Lin, Z. (2016). "Assessing working vulnerability of construction labor through eeg signal processing," in *16th International Conference on Computing in Civil and Building Engineering* (Hong Kong), 1053–1059.

Chen, J., Lin, Z., and Guo, X. (2018). "Developing construction workers'mental vigilance indicators through wavelet packet decomposition on eeg signals," in *Construction Research Congress 2018: Safety and Disaster Management, CRC 2018* (Hong Kong: American Society of Civil Engineers), 51–61.

Choi, E. J., and Kim, D. K. (2018). Arousal and valence classification model based on long short-term memory and deap data for mental healthcare management. *Healthcare Inform. Res.* 24, 309–316. doi: 10.4258/hir.2018.24.4.309

Doukas, C., and Maglogiannis, I. (2008). Intelligent pervasive healthcare systems, advanced computational intelligence paradigms in healthcare. *Stud. Comput. Intell.* 107, 95–115. doi: 10.1007/978-3-540-77662-8_5

Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., and Van Gool, L. (2010). A 3-d audio-visual corpus of affective communication. *IEEE Trans. Multim.* 12, 591–598. doi: 10.1109/TMM.2010.2052239

Gupta, V., Chopda, M. D., and Pachori, R. B. (2018). Cross-subject emotion recognition using flexible analytic wavelet transform from eeg signals. *IEEE Sensors J.* 19, 2266–2274. doi: 10.1109/JSEN.2018.2883497

Hanjalic, A., and Xu, L.-Q. (2005). Affective video content representation and modeling. *IEEE Trans. Multim.* 7, 143–154. doi: 10.1109/TMM.2004.840618

Hjorth, B. (1970). Eeg analysis based on time domain properties. *Electroencephal. Clin. Neurophysiol.* 29, 306–310.

Horlings, R., Datcu, D., and Rothkrantz, L. J. (2008). "Emotion recognition using brain activity," in *Proceedings of the 9th International Conference on Computer Systems and Technologies and Workshop for PhD Students in Computing* (Delft: ACM), 6.

Hou, X., Liu, Y., Sourina, O., and Mueller-Wittig, W. (2015). "Cognimeter: Eeg-based emotion, mental workload and stress visual monitoring," in *2015 International Conference on Cyberworlds (CW)* (Sydney, NSW: IEEE), 153–160.

Hwang, S., Jebelli, H., Choi, B., Choi, M., and Lee, S. (2018). Measuring workers'emotional state during construction tasks using wearable eeg. *J. Constr. Eng. Manage.* 144:04018050. doi: 10.1061/(ASCE)CO.1943-7862.0001506

Jebelli, H., Hwang, S., and Lee, S. (2017a). Eeg signal-processing framework to obtain high-quality brain waves from an off-the-shelf wearable eeg device. *J. Comput. Civil Eng.* 32:04017070. doi: 10.1061/(ASCE)CP.1943-5487.0000719

Jebelli, H., Hwang, S., and Lee, S. (2017b). Feasibility of field measurement of construction workers'valence using a wearable eeg device. *Comput. Civil Eng.* 99–106. doi: 10.1061/9780784480830.013

Jebelli, H., Hwang, S., and Lee, S. (2018a). Eeg-based workers'stress recognition at construction sites. *Autom. Construct.* 93, 315–324. doi: 10.1016/j.autcon.2018.05.027

Jebelli, H., Khalili, M. M., Hwang, S., and Lee, S. (2018b). A supervised learning-based construction workers'stress recognition using a wearable electroencephalography (eeg) device. *Constr. Res. Congress* 2018, 43–53. doi: 10.1061/9780784481288.005

Jebelli, H., Khalili, M. M., and Lee, S. (2018c). A continuously updated, computationally efficient stress recognition framework using electroencephalogram (eeg) by applying online multi-task learning algorithms (omtl). *IEEE J. Biomed. Health Inform.* doi: 10.1109/JBHI.2018.2870963. [Epub ahead of print].

Jebelli, H., Khalili, M. M., and Lee, S. (2019). "Mobile EEG-based workers' stress recognition by applying deep neural network," in *Advances in Informatics and Computing in Civil and Construction Engineering*, eds I. Mutis and T. Hartmann (Cham: Springer), 173–180.

Jerritta, S., Murugappan, M., Nagarajan, R., and Wan, K. (2011). "Physiological signals based human emotion recognition: a review," in *2011 IEEE 7th International Colloquium on Signal Processing and its Applications* (Arau: IEEE), 410–415.

Kim, J. (2007). "Bimodal emotion recognition using speech and physiological changes," in *Robust Speech Recognition and Understanding* (Augsburg: IntechOpen).

Kim, J., and André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Trans. Pattern Analy. Mach. Intell.* 30, 2067–2083. doi: 10.1109/TPAMI.2008.26

Knyazev, G. G., Slobodskoj-Plusnin, J. Y., and Bocharov, A. V. (2010). Gender differences in implicit and explicit processing of emotional facial expressions as revealed by event-related theta synchronization. *Emotion* 10:678. doi: 10.1037/a0019175

Koelstra, S., Muhl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2011). Deap: a database for emotion analysis; using physiological signals. *IEEE Trans. Affect. Comput.* 3, 18–31. doi: 10.1109/T-AFFC.2011.15

Lee, Y.-Y., and Hsieh, S. (2014). Classifying different emotional states by means of eeg-based functional connectivity patterns. *PLoS ONE* 9:e95415. doi: 10.1371/journal.pone.0095415

Li, J., Qiu, S., Shen, Y.-Y., Liu, C.-L., and He, H. (2019). Multisource transfer learning for cross-subject eeg emotion recognition. *IEEE Transact. Cyber*. 1–13. doi: 10.1109/TCYB.2019.2904052

Li, X., Song, D., Zhang, P., Zhang, Y., Hou, Y., and Hu, B. (2018). Exploring eeg features in cross-subject emotion recognition. *Front. Neurosci.* 12:162. doi: 10.3389/fnins.2018.00162

Liu, Y., and Sourina, O. (2014). "Real-time subject-dependent eeg-based emotion recognition algorithm," in *Transactions on Computational Science XXIII*, eds M. L. Gavrilova, C. J. Kenneth Tan, X. Mao, and L. Hong (Singapore: Springer), 199–223.

Mathersul, D., Williams, L. M., Hopkinson, P. J., and Kemp, A. H. (2008). Investigating models of affect: Relationships among eeg alpha asymmetry, depression, and anxiety. *Emotion* 8:560. doi: 10.1037/a0012811

Mert, A., and Akan, A. (2018). Emotion recognition from eeg signals by using multivariate empirical mode decomposition. *Pattern Analy. Appl.* 21, 81–89. doi: 10.1007/s10044-016-0567-6

Mucci, N., Giorgi, G., Cupelli, V., Gioffrè, P. A., Rosati, M. V., Tomei, F., et al. (2015). Work-related stress assessment in a population of italian workers. the stress questionnaire. *Sci. Total Environ.* 502, 673–679. doi: 10.1016/j.scitotenv.2014.09.069

Nie, D., Wang, X.-W., Shi, L.-C., and Lu, B.-L. (2011). "Eeg-based emotion recognition during watching movies," in *2011 5th International IEEE/EMBS Conference on Neural Engineering* (Shanghai: IEEE), 667–670.

Nijboer, F., Morin, F. O., Carmien, S. P., Koene, R. A., Leon, E., and Hoffmann, U. (2009). "Affective brain-computer interfaces: psychophysiological markers of emotion in healthy persons and in persons with amyotrophic lateral sclerosis," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops* (Donostia-San Sebastian: IEEE), 1–11.

Pessoa, L., and Adolphs, R. (2010). Emotion processing and the amygdala: from a'low road'to'many roads' of evaluating biological significance. *Nat. Rev. Neurosci.* 11:773. doi: 10.1038/nrn2920

Petrantonakis, P. C., and Hadjileontiadis, L. J. (2011). A novel emotion elicitation index using frontal brain asymmetry for enhanced eeg-based emotion recognition. *IEEE Trans. Inform. Techn. Biomed.* 15, 737–746. doi: 10.1109/TITB.2011.2157933

Picard, R. W., Vyzas, E., and Healey, J. (2001). Toward machine emotional intelligence: analysis of affective physiological state. *IEEE Trans. Pattern Analy. Mach. Intell.* 23, 1175–1191. doi: 10.1109/34.954607

Plutchik, R. (1962). The emotions: facts, theories and a new model. *Am. J. Psychol.* 77:518.

Power, M. J., and Dalgleish, T. (1999). *Handbook of Cognition and Emotion*. Singapore: Wiley.

Richman, J. S., and Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circul. Physiol.* 278, H2039–H2049. doi: 10.1152/ajpheart.2000.278.6.H2039

Russell, J. A. (1980). A circumplex model of affect. *J. Personal. Soc. Psychol.* 39:1161.

Sammler, D., Grigutsch, M., Fritz, T., and Koelsch, S. (2007). Music and emotion: electrophysiological correlates of the processing of pleasant and unpleasant music. *Psychophysiology* 44, 293–304. doi: 10.1111/j.1469-8986.2007.00497.x

Shi, L.-C., and Lu, B.-L. (2013). Eeg-based vigilance estimation using extreme learning machines. *Neurocomputing* 102, 135–143. doi: 10.1016/j.neucom.2012.02.041

Snowball, P. S. (2005). Spectral analysis of signals. *Leber Magen Darm* 13, 57–63.

Takahashi, K. et al. (2004). "Remarks on emotion recognition from bio-potential signals," in *2nd International Conference on Autonomous Robots and Agents* (Kyoto), 1148–1153.

Yin, Z., Wang, Y., Liu, L., Zhang, W., and Zhang, J. (2017). Cross-subject eeg feature selection for emotion recognition using transfer recursive feature elimination. *Front. Neurorobot.* 11:19. doi: 10.3389/fnbot.2017.00019

Zeng, Z., Pantic, M., Roisman, G. I., and Huang, T. S. (2008). A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Trans. Pattern Analy. Mach. Intell.* 31, 39–58. doi: 10.1109/TPAMI.2008.52

Zhai, J., Barreto, A. B., Chin, C., and Li, C. (2005). "Realization of stress detection using psychophysiological signals for improvement of human-computer interactions," in *Proceedings IEEE SoutheastCon, 2005* (Miami, FL: IEEE), 415–420.

Zheng, W.-L., and Lu, B.-L. (2017b). A multimodal approach to estimating vigilance using eeg and forehead eog. *J. Neural Eng.* 14:026017. doi: 10.1088/1741-2552/aa5a98

Zheng, W. L., Liu, W., Lu, Y., Lu, B. L., and Cichocki, A. (2019). Emotionmeter: a multimodal framework for recognizing human emotions. *IEEE Trans. Cybernet.* 49, 1110–1122. doi: 10.1109/TCYB.2018.2797176

Zheng, W. L., and Lu, B. L. (2017a). Investigating critical frequency bands and channels for eeg-based emotion recognition with deep neural networks. *IEEE Trans. Autonom. Ment. Dev.* 7, 162–175. doi: 10.1109/TAMD.2015.2431497

Zhu, J.-Y., Zheng, W.-L., and Lu, B.-L. (2015). "Cross-subject and cross-gender emotion classification from eeg," in *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada* (Shanghai: Springer), 1188–1191.

Zhuang, N., Zeng, Y., Tong, L., Zhang, C., Zhang, H., and Yan, B. (2017). Emotion recognition from eeg signals using multidimensional information in emd domain. *Biomed Res. Int.* 2017:8317357. doi: 10.1155/2017/8317357

**Conflict of Interest Statement:** 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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 05 November 2019
doi: 10.3389/fncom.2019.00077

# Beta-Band Resonance and Intrinsic Oscillations in a Biophysically Detailed Model of the Subthalamic Nucleus-Globus Pallidus Network

*Lucas A. Koelman* and Madeleine M. Lowery*

*Neuromuscular Systems Laboratory, School of Electrical and Electronic Engineering, University College Dublin, Dublin, Ireland*

Increased beta-band oscillatory activity in the basal ganglia network is associated with Parkinsonian motor symptoms and is suppressed with medication and deep brain stimulation (DBS). The origins of the beta-band oscillations, however, remains unclear with both intrinsic oscillations arising within the subthalamic nucleus (STN)—external globus pallidus (GPe) network and exogenous beta-activity, originating outside the network, proposed as potential sources of the pathological activity. The aim of this study was to explore the relative contribution of autonomous oscillations and exogenous oscillatory inputs in the generation of pathological oscillatory activity in a biophysically detailed model of the parkinsonian STN-GPe network. The network model accounts for the integration of synaptic currents and their interaction with intrinsic membrane currents in dendritic structures within the STN and GPe. The model was used to investigate the development of beta-band synchrony and bursting within the STN-GPe network by changing the balance of excitation and inhibition in both nuclei, and by adding exogenous oscillatory inputs with varying phase relationships through the hyperdirect cortico-subthalamic and indirect striato-pallidal pathways. The model showed an intrinsic susceptibility to beta-band oscillations that was manifest in weak autonomously generated oscillations within the STN-GPe network and in selective amplification of exogenous beta-band synaptic inputs near the network's endogenous oscillation frequency. The frequency at which this resonance peak occurred was determined by the net level of excitatory drive to the network. Intrinsic or endogenously generated oscillations were too weak to support a pacemaker role for the STN-GPe network, however, they were considerably amplified by sparse cortical beta inputs and were further amplified by striatal beta inputs that promoted anti-phase firing of the cortex and GPe, resulting in maximum transient inhibition of STN neurons. The model elucidates a mechanism of cortical patterning of the STN-GPe network through feedback inhibition whereby intrinsic susceptibility to beta-band oscillations can lead to phase locked spiking under parkinsonian conditions. These results point to resonance of endogenous oscillations with exogenous patterning of the STN-GPe network as a mechanism of pathological synchronization, and a role for the pallido-striatal feedback loop in amplifying beta oscillations.

**Keywords: basal ganglia, subthalamic nucleus, Parkinson's disease, beta-band oscillations, synchronization, globus pallidus, multi-compartmental neuron model**

OPEN ACCESS

***Edited by:***
*Matjaž Perc,*
*University of Maribor, Slovenia*

***Reviewed by:***
*Ergin Yilmaz,*
*Bulent Ecevit University, Turkey*
*Chen Liu,*
*Tianjin University, China*

****Correspondence:***
*Lucas A. Koelman*
*lucas.koelman@gmail.com*

***Received:*** *27 June 2019*
***Accepted:*** *17 October 2019*
***Published:*** *05 November 2019*

***Citation:***
*Koelman LA and Lowery MM (2019) Beta-Band Resonance and Intrinsic Oscillations in a Biophysically Detailed Model of the Subthalamic Nucleus-Globus Pallidus Network. Front. Comput. Neurosci. 13:77. doi: 10.3389/fncom.2019.00077*

## INTRODUCTION

Pathological oscillations in the basal ganglia-thalamocortical (BGTC) network have long been implicated in the motor symptoms of Parkinson's disease. Beta-band (13–30 Hz) oscillations are consistently strengthened with dopamine depletion both in individuals with Parkinson's disease (PD) and parkinsonian animal models (Sharott et al., 2005; Kuhn et al., 2008; Mallet et al., 2008b), and are reduced by deep brain stimulation (DBS) and pharmacological interventions that alleviate parkinsonian motor symptoms (Kühn et al., 2006; Weinberger et al., 2006; Ray et al., 2008; Eusebio et al., 2011). The magnitude of subthalamic nucleus local field potential beta oscillations is also correlated with the severity and degree of improvement of bradykinetic/akinetic motor symptoms and rigidity (Kühn et al., 2006; Bronte-Stewart et al., 2009). Although beta-band oscillations may not be causal to bradykinetic/akinetic symptoms (Leblois et al., 2007), they offer potential as a biomarker for symptom severity and the underlying network pathophysiology in advanced Parkinson's Disease. The origin of beta-band oscillations in the BGTC network, however, remains unclear. The most prominent hypotheses emphasize the importance of dopamine-modulated strengthening of particular feedback loops within the BGTC network. Computational models have provided a valuable tool with which to explore various hypotheses regarding the mechanisms by which oscillatory activity with the network is generated. Different models have placed the origin of beta and sub-beta band oscillations in the STN-GPe network (Terman et al., 2002; Gillies and Willshaw, 2007; Holgado et al., 2010; Pavlides et al., 2012), in cortical and thalamo-cortical circuits (Pavlides et al., 2015; Sherman et al., 2016; Liu et al., 2017; Reis et al., 2019), in striatal or pallidostriatal circuits (McCarthy et al., 2011; Corbit et al., 2016), or in the full BGTC loop (Leblois, 2006; Kang and Lowery, 2013; Pavlides et al., 2015; Kumaravelu et al., 2016). These models show that under many conditions the network is prone to oscillate, through intrinsic pacemaking or susceptibility to an extrinsic rhythm.

The reciprocally connected subthalamo-pallidal (STN-GPe) network is a key site in the basal ganglia in which beta-band oscillations are manifest in Parkinson's disease (Mallet et al., 2008a,b). This network was an early focus of modeling studies due to its reciprocally connected structure and ability to generate low frequency oscillations in tissue cultures (Plenz and Kital, 1999). Models of the STN-GPe as a pacemaker initially focused on the generation of low frequency oscillations within the frequency range of parkinsonian tremor (Gillies et al., 2002; Terman et al., 2002), with focus shifting to the beta-band with increasing evidence of a link between beta activity and parkinsonian motor symptoms (Holgado et al., 2010; Pavlides et al., 2012).

More recent experimental evidence suggests that, rather than the STN-GPe network operating in a pacemaking mode, patterning by cortex may play a critical role in the generation of pathological beta-band oscillations in Parkinson's disease. This is supported by observations of high functional coupling between cortex and STN (Magill et al., 2004; Sharott et al., 2005; Mallet et al., 2008a; Litvak et al., 2011; Moran et al., 2011), and that oscillatory activity in STN-GPe is contingent on inputs from the cortex and can be abolished by disrupting them (Magill et al., 2001; Drouot et al., 2004; Tachibana et al., 2011). Cortical patterning of the STN-GPe network by means of feedback inhibition provides a proposed mechanism for this functional coupling (Baufreton et al., 2005; Bevan et al., 2006; Mallet et al., 2008a, 2012; Tachibana et al., 2011). According to this hypothesis, weak oscillatory activity arriving via cortico-STN afferents is amplified in the STN-GPe network when feedback inhibition from the GPe is offset in phase with cortical excitation. While such feedback-mediated oscillations have been observed *in vivo* (Paz, 2005) and in slices (Baufreton et al., 2005), the ability of the network to generate autonomous oscillations and its resonant response properties are still poorly understood. Specifically, it is not clear whether the STN-GPe network plays an active part in generating beta-band oscillations, nor whether it amplifies or merely sustains them. Neither is it fully understood how beta-band oscillations relate to other pathological patterns of neural activity in the subthalamic nucleus (STN) and external globus pallidus (GPe) that correlate more strongly with parkinsonian motor symptoms, notably increased neural bursting (Sanders et al., 2013; Sharott et al., 2014). It is clear, however, that interventions in the loop and its afferents that reduce beta-band oscillations (Tachibana et al., 2011) or bursting (Gradinaru et al., 2009; Pan et al., 2016; Sanders and Jaeger, 2016) lead to improvements in motor symptoms. Similarly, the STN (Benabid et al., 2009) and GPe, in non-human primates (Vitek et al., 2012), are effective targets for DBS.

Previous modeling studies have focused on alterations in connection patterns and strength within or between nuclei, typically represented by mean-field or single-compartment spiking neuron models. While such models are computationally efficient, they may not fully capture the role of intrinsic properties of neurons in shaping pathological activity patterns. Although cell-specific ion channels can be used, single-compartment neuron models lump together ion channels and synapses in one isopotential compartment in a way that may not capture the complex dynamics that arise when non-uniformly distributed ion channels (Gillies and Willshaw, 2005) interact with synapses associated with distinct subcellular regions (Bevan et al., 1995; Galvan et al., 2004; Pan et al., 2016). Hence they may not fully account for the mechanisms contributing to pathological activity within the STN and the role that synaptic-ionic current interactions play in sustaining beta-band oscillations and excessive burst firing.

It has recently been demonstrated that following dopamine depletion the balance of excitatory and inhibitory synaptic currents in STN neurons is shifted toward inhibition (Chu et al., 2017; Wang et al., 2018), known to promote burst responses by increasing the availability of $Ca^{2+}$ and $Na^{+}$ channels de-inactivated at hyperpolarized potentials (Baufreton et al., 2005). In the GPe increased inhibition, caused mainly by strengthening of striato-pallidal afferents, is also believed to play a role in generating pathological oscillations as demonstrated in model simulation (Gillies et al., 2002; Terman et al., 2002; Holgado et al., 2010; Kumar et al., 2011). Increased GPe inhibition

has been suggested to cause increased engagement of HCN channels (Chan, 2004), which are involved in phase resetting and controlling the regularity of firing. However, whether functional coupling between BG nuclei is also moderated by the excitation-inhibition balance is not fully understood.

The aim of this study was, therefore, to examine the relative contributions of intrinsic, endogenously generated oscillations and patterning by exogenous oscillatory inputs in the generation of synchronous beta-band oscillatory activity in a biophysically detailed model of the parkinsonian STN-GPe network and the underlying biophysical mechanisms. A second aim was to understand how pathological oscillations and bursting patterns are related to the balance of excitation and inhibition in the STN and GPe. The STN-GPe network was modeled using biophysically detailed multi-compartmental cell models of STN and GPe neurons that capture the interaction between synaptic and intrinsic currents distributed within the dendritic structure and involved in autonomous pacemaking and bursting (Gillies and Willshaw, 2005; Gunay et al., 2008). The generation of oscillations both autonomously within the network and in response to beta frequency inputs from the cortex (CTX) and indirect pathway striatal medium spiny neurons (iMSN) was examined as the balance of excitation and inhibition within the network was systematically varied, and oscillatory inputs with varying phase relationships were added. A better understanding of the relative contribution of these different factors and their interaction has the potential to improve understanding of the mechanism of action of existing anti-parkinsonian therapies, including DBS and to guide the development of more effective circuit interventions.

# METHODS

## Model Architecture

The network model of the STN-GPe network consisted of four populations of neurons (**Figure 1**): the STN and GPe neurons, modeled as multi-compartmental conductance-based models, and their cortical and striatal inputs, modeled as Poisson or bursting spike generators.

Population sizes were chosen to preserve the decrease in population sizes and convergence of projections along the indirect and hyperdirect pathways in the basal ganglia. The STN and GPe populations consisted of 50 and 100 multi-compartmental cells, respectively, to approximate the ratio of 13,000 STN cells to 30,000 GPe prototypic cells (Oorschot et al., 1999; Abdi et al., 2015) unilaterally in the rat. As a source of synaptic noise, an additional 10% of the cells in the STN and GPe populations were modeled as Poisson spike generators firing at a mean rate equal to the experimentally reported rate for the modeled state.

The cortical and striatal populations consisted of 1,000 and 2,000 cells, respectively, modeled as spike generators. These numbers were chosen to have 20 independent pre-synaptic spike generators per post-synaptic cell to model convergence along the hyperdirect CTX-STN and indirect iMSN-GPe projection. For the iMSN-GPe projection, convergence from all medium spiny neurons (MSN) to GPe, ignoring subpopulations, is 2,800,000 MSN cells to 46,000 GPe cells (Oorschot, 1996) resulting in a convergence factor of 60. Assuming that convergence is similar between iMSN and GPe prototypic neurons, our number is an underestimation by a factor three. Because iMSN cells in our model spike independently and since the number of synapses per cell was lower than in reality, this was considered acceptable.

Stochastic connectivity profiles for the connections illustrated in **Figure 1** were generated by randomly selecting a fixed number of afferents from the pre-synaptic population for each post-synaptic cell. The ratios of number of afferents from each source population were determined, where possible, based on the reported number of synaptic boutons per afferent type and the number of contacts per axon (**Table 1**). Each multi-synaptic contact was represented by a single synapse to reduce the number of simulated synapses to a more tractable number.

## Conductance-Based Models

The membrane potential $v_j$ (mV) in each compartment $j$ of a multi-compartmental cable model is governed by:

$$c_m \frac{\delta v_j}{\delta t} = \frac{d}{4R_a}\frac{\delta^2 V}{\delta x^2} - g_m(V - E_m) - \sum I_{ion,j} - \sum I_{syn,j} \quad (1)$$

where $x$ (cm) is the position along the cable, $c_m$ ($\mu F/cm^2$) is the specific membrane capacitance, $d$ (cm) is the cable diameter, $R_a$ is the specific axial resistance ($\Omega cm$), $g_m$ ($S/cm^2$) is the passive membrane conductance, $E_m$ (mV) the leakage reversal potential, $I_{ion,j}$ ($mA/cm^2$) are the ionic currents flowing across the membrane of compartment $j$, and $I_{syn,j}$ ($mA/cm^2$) are the synaptic currents at synapses placed in the compartment. Each ionic current is governed by an equation of the form:

$$I_x = \overline{g}_x m_x^p h_x^q (V - E_x) \quad (2)$$

where $\overline{g}_x$ is the maximum conductance of the channel ($S/cm^2$), $E_x$ is the reversal potential (mV), and $m_x$ and $h_x$ the open fractions of the activation and inactivation gates. The dynamics of the activation and inactivation gates $m$ and $h$ are governed by

$$\frac{dm}{dt} = \frac{m_\infty(v) - m}{\tau_m(v)}, \quad (3)$$

with $m_\infty(v)$ and $\tau_m(v)$ representing the voltage-dependent steady state value and time constant of the gate. For some currents the gating dynamics are described in terms of the opening and closing rates $\alpha_m$ and $\beta_m$ related through $\tau_m = \frac{1}{\alpha_m+\beta_m}$, $m_\infty = \frac{\alpha_m}{\alpha_m+\beta_m}$:

$$\frac{dm}{dt} = \alpha_m(v) \cdot (1 - m) - \beta_m(v) \cdot m. \quad (4)$$

Reversal potentials are assumed constant unless otherwise noted. The reversal potential for $Ca^{2+}$ currents was calculated using the Nernst equation from the intra- and extracellular ion concentrations:

$$E_{Ca} = \frac{RT}{zF} \ln \frac{\left[Ca^{2+}\right]_o}{\left[Ca^{2+}\right]_i} \quad (5)$$

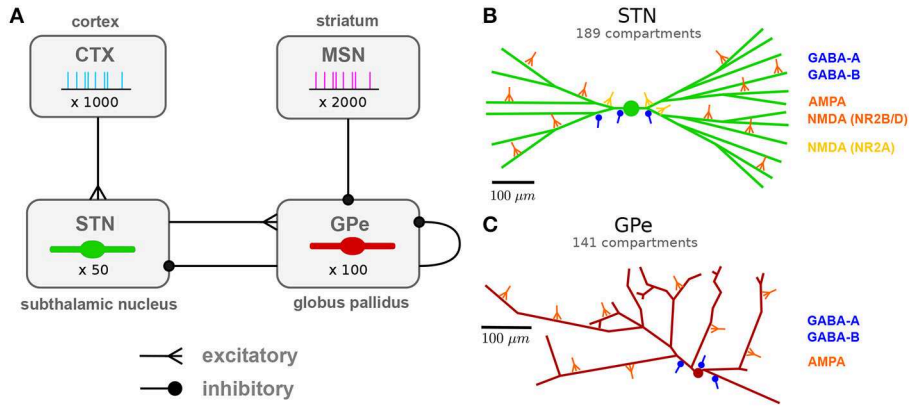


**FIGURE 1 |** Network architecture: population and subcellular connectivity. **(A)** Neuronal populations and their projections modeled in the network. Subthalamic projection neurons (STN) and prototypic neurons of the external globus pallidus (GPe) were modeled using multi-compartmental neuron models. Cortical projection neurons (CTX) and indirect pathway striatal medium spiny neurons (iMSN) were modeled as spike generators. **(B)** Branching structure of the STN neuron model and representative synapses by afferent type, indicating subcellular distribution of synapses. Cortical glutamergic afferents synapse primarily onto thin dendrites, distally relative to the soma, but NMDA receptors with faster NR2A subunits mainly target the soma and proximal areas. Pallidal GABAergic afferents target proximal areas of the cell. **(C)** Branching structure of the GPe neuron model and representative synapses by afferent type. GABAergic GPe-GPe collaterals mainly target somata and proximal dendrites, whereas glutamergic afferents were placed in distal regions. Full details of the model are provided in the section Methods.

where $T$ is the temperature in Kelvin, $R$ is the universal gas constant, $F$ is the Faraday constant, and $z$ is the valence of the calcium ion (+2). Intracellular calcium buffering in a sub-membrane shell is modeled as:

$$\frac{d\left[\mathrm{Ca}^{2+}\right]_i}{dt} = -\left(I_{\mathrm{CaL}} + I_{\mathrm{CaN}} + I_{\mathrm{CaT}}\right)\frac{c}{2Fd} - \frac{\left[\mathrm{Ca}^{2+}\right]_{i0} - \left[\mathrm{Ca}^{2+}\right]_i}{\tau_{\mathrm{Ca}}} \tag{6}$$

where $c$ is a unit conversion constant, $d$ is the thickness of the sub-membrane shell, and $\tau_{\mathrm{Ca}}$ is the time constant of decay.

Synaptic connections between cells were modeled by spike detectors in the somatic compartments, coupled to synapses in the target cells by a time delay. As no interactions between axons and other biophysical processes such as electric fields were required, axonal structures were omitted from the model and represented as delays between connected neurons. This constrained the computational complexity of the model, avoiding the requirement to simulate large number of additional compartments without altering the network behavior. Synapses were modeled by a dual exponential profile with rise and decay times $\tau_{rise}$ and $\tau_{decay}$ modulated by the fraction of synaptic resources in the active state which was governed by Tsodyks-Markram dynamics (Tsodyks et al., 1998):

$$I_{syn} = \bar{g}_{syn}(B - A)(v - E_{syn}) \tag{7}$$

$$\frac{dA}{dt} = \frac{-A}{\tau_{rise}} + f_{peak} \cdot U_{SE} \cdot R \cdot \delta(t - t_{spk}) \tag{8}$$

$$\frac{dB}{dt} = \frac{-B}{\tau_{decay}} + f_{peak} \cdot U_{SE} \cdot R \cdot \delta(t - t_{spk}) \tag{9}$$

$$\frac{dR}{dt} = \frac{1 - R}{\tau_{rec}} - U_{SE} \cdot R \cdot \delta(t - t_{spk}) \tag{10}$$

$$\frac{dU_{SE}}{dt} = \frac{-U_{SE}}{\tau_{facil}} + U_1 \cdot (1 - U_{SE}) \cdot \delta(t - t_{spk}) \tag{11}$$

$$f_{peak} = \frac{1}{\exp(-t_{peak}/\tau_{decay}) - \exp(-t_{peak}/\tau_{rise})} \tag{12}$$

$$t_{peak} = \frac{\tau_{rise} \cdot \tau_{decay}}{\tau_{decay} - \tau_{rise}} \log\left(\frac{\tau_{decay}}{\tau_{rise}}\right) \tag{13}$$

where, $\bar{g}_{syn}$ is the peak synaptic conductance, B-A represents the synaptic gating variable, $f_{peak}$ is a normalization factor so that B-A reaches its maximum at time $t_{peak}$ after the time of spike arrival $t_{spk}$, R is the fraction of vesicles available for release, $U_{SE}$ is the release probability, and $\tau_{rec}$ and $\tau_{facil}$ are the time constants for recovery from short-term depression and facilitation, respectively. The synaptic reversal potentials $E_{syn}$ were 0 mV for AMPA and NMDA, −80 mV for $\mathrm{GABA_A}$, and −95 mV for $\mathrm{GABA_B}$. For NMDA synapses there is an additional voltage-dependent gating variable representing magnesium block (Jahr and Stevens, 1990):

$$m(v) = 1/(1 + exp(-0.062v) * (1/3.57)) \tag{14}$$

The metabotropoc $\mathrm{GABA_B}$ receptor-mediated current was modeled as an intracellular signaling cascade based on the model by Destexhe and Sejnowski (1995). The equations describing G-protein activation and the synaptic current were retained, but the bound receptor fraction including the effects of desensitization was represented by the fraction of resources in the active state in the Tsodyks-Markram scheme (B-A). The equation governing the G-protein production rate thus became

$$\frac{dG}{dt} = K3 * (B - A) - K4 * G \tag{15}$$

**TABLE 1 |** Experimentally reported connection parameters used to calibrate the model.

| Target | Source | Afferent neurons | Synaptic contacts | Subcellular targets | Short-term plasticity | Delay | Effect of dopamine depletion |
|---|---|---|---|---|---|---|---|
| STN | (all) | 300 (Baufreton and Bevan, 2008) | | N.A. | N.A. | N.A. | N.A. |
| | CTX | | | distal (Bevan et al., 1995; Mathai et al., 2015; Pan et al., 2016), proximal (Pan et al., 2016) | depression (Chu et al., 2015) | 5.9 ms (Kita and Kita, 2011) | weakened (Chu et al., 2017; Wang et al., 2018) |
| | GPe | 57 (Atherton et al., 2013) | 883 (Baufreton et al., 2009) | proximal (Smith et al., 1990) | depression (Atherton et al., 2013) | 4 ms (Fujimoto and Kita, 1993) | strengthened (Chu et al., 2015) prolonged decay (Fan et al., 2012) |
| GPe | GPe | | | proximal,somatic (Chan, 2004; Sadek et al., 2007) | depression (Miguelez et al., 2012) | | strengthened (Miguelez et al., 2012) |
| | STN | 135 (Kita and Jaeger, 2016) | | dendritic,distal (Shink and Smith, 1995) | facilitation, depression (Hanson and Jaeger, 2002) | 2 ms (Kita and Kitai, 1991) | strengthened (Hernández et al., 2006) |
| | MSN | | 10622 (Kita and Jaeger, 2016) | dendritic, distal (Chan, 2004) | facilitation (Miguelez et al., 2012) | 5 ms (Kita and Kitai, 1991) | |

where G is the G-protein concentration, and K3 and K4 are the rates of G-protein production and decay, respectively. The G-protein concentration $G$ gates the peak synaptic conductance according to a sigmoid activation function represented by the Hill equation:

$$I_{\mathrm{GABA_B}} = \overline{g}_{syn} \frac{G^n}{G^n + K_d^n} (v - E_{GABA_B}). \tag{16}$$

## STN Cell Model

STN neurons were modeled using the rat subthalamic projection neuron model by Gillies and Willshaw (2005) (ModelDB accession number 74298). The neuron morphology is based on quantitative characterization of the dendritic trees of STN neurons *in vitro*. The model includes 10 intrinsic ionic currents (**Table 2**) :

$$\begin{aligned} I_{ion,j} = {} & I_{NaF} + I_{NaP} \\ & + I_{KDR} + I_{Kv31} + I_{sKCa} \\ & + I_{CaT} + I_{CaL} + I_{CaN} \\ & + I_{HCN} + I_L \end{aligned} \tag{17}$$

where $I_{NaF}$ and $I_{NaP}$ are the transient fast-acting and persistent sodium current, $I_{KDR}$, $I_{Kv31}$, and $I_{sKCa}$ the delayed rectifier, fast rectifier and calcium-activated potassium current, $I_{CaT}$, $I_{CaL}$, and $I_{CaN}$ the low-voltage-activated T-type, high-voltage-activated L-type, and high-voltage-activated N-type calcium currents, $I_{HCN}$ the hyperpolarization-activated cyclic nucleotide (HCN) current, and $I_L$ the leak current. The equations governing the dynamics of the gating variables are listed in **Table 2**. The channel density distributions are described extensively in Gillies and Willshaw (2005). As a source of noise, a current with a Gaussian amplitude distribution, mean zero and standard deviation 0.1 was added to the somatic compartment.

The synaptic currents included an excitatory glutamergic input from cortex, acting through AMPA and NMDA receptors, and an inhibitory GABAergic input from the GPe, acting through $\mathrm{GABA_A}$ and $\mathrm{GABA_B}$ receptors (**Table 3**):

$$\begin{aligned} I_{syn,j} = {} & I_{\mathrm{CTX-STN,AMPA}} + I_{\mathrm{CTX-STN,NMDA}} \\ & + I_{\mathrm{GPE-STN,GABA_A}} + I_{\mathrm{GPE-STN,GABA_B}} \end{aligned} \tag{18}$$

In the control condition STN neurons had 20 excitatory afferents from CTX neurons and 8 inhibitory afferents from GPe neurons. The location of synapses on STN neurons and axonal propagation delays were based on experimental observations (**Table 1**). Cortico-subthalamic (CTX-STN) synapses were modeled as conductance-based synapses with Tsodyks-Markram dynamics (Tsodyks et al., 1998). On each of its target cells, a cortical axon had one synapse located distally in the dendritic tree and one located proximally near the soma. Distal synapses had both an AMPA and slower NMDA conductance component. The latter represented slower-kinetics NMDA receptors with majority NR2B and NR2D subunits that have dendritic punctual expression (Pan et al., 2016). Proximal synapses had only an NMDA component and represented NMDA receptors with fast-kinetics NR2A subunits. Synaptic parameter values are listed in **Table 3**. Synaptic rise and decay time constants $\tau_{rise}$ and $\tau_{decay}$ for AMPA and NMDA NR2A constants were based on traces reported in Chu et al. (2015). For the slower NMDA NR2B synapses, values were based on Flint et al. (1997). The propagation delay $t_d$ was taken from Kita and Kita (2011). Synapses were made to exhibit short-term depression upon high-frequency activation, based on observations by Froux et al. (2018). The ratio of the total AMPA to NMDA conductance were

**TABLE 2 |** STN model intrinsic current equations from Gillies and Willshaw (2005).

| Current | Equation | Gating variables | | Parameters |
|---|---|---|---|---|
| $I_{NaF}$ | $\bar{g}_{NaF} m^2 h (v - E_{Na})$ | $\alpha_m = 0.32 \frac{(13.1-v)}{\exp((13.1-v)/4)-1}$<br>$\alpha_h = 0.128 \exp\left(\frac{17-v}{18}\right)$ | $\beta_m = 0.28 \frac{(v-40,1)}{\exp(v-40.1)-1}$<br>$\beta_h = \frac{4}{\exp((40-v)/5)+1}$ | $\bar{g}_{NaF} = 14.83\text{e}{-3}$ (soma)<br>$\bar{g}_{NaF} = 1\text{e}{-7}$ (dendrite) |
| $I_{NaP}$ | $\bar{g}_{NaP} (v - E_{Na})$ | | | $\bar{g}_{NaP} = 1.11\text{e}{-5}$ (soma)<br>$\bar{g}_{NaP} = 8.10\text{e}{-6}$ (dendrite) |
| $I_{KDR}$ | $\bar{g}_{KDR} n (v - E_K)$ | $\alpha_n = \frac{0.016(35.1-v)}{\exp((35.1-v)/5)-1}$ | $\beta_n = 0.25 \exp((20 - v)/40)$ | $\bar{g}_{KDR} = 3.84\text{e}{-3}$ (soma)<br>$\bar{g}_{KDR} \in [4.22, 9.32] \times 10^5$ (dendrite) |
| $I_{Kv31}$ | $\bar{g}_{Kv31} p (v - E_K)$ | $p_\infty = \frac{1}{1+\exp(-(v+5)/9)}$ | $\tau_\infty = \frac{18.71}{\exp(-(v+28)/6)+\exp((v+4)/16)}$ | $\bar{g}_{Kv31} = 1.34\text{e}{-2}$ (soma)<br>$\bar{g}_{Kv31} \in [8.91, 10] \times 10^4$ (dendrite) |
| $I_{sKCa}$ | $\bar{g}_{sKCa} w (v - E_K)$ | $w_\infty = \frac{0.81}{1+\exp\frac{-\log\left[Ca^{2+}\right]_i - 0.3}{0.46}}$ | $\tau_w = 40$ | $\bar{g}_{sKCa} = 6.84\text{e}{-5}$ (soma)<br>$\bar{g}_{sKCa} = 3.92\text{e}{-5}$ (dendrite) |
| $I_{HCN}$ | $\bar{g}_{HCN} f (v - E_{HCN})$ | $f_\infty = \frac{1}{1+\exp(v+75)/5.5]}$ | $\tau_f = \frac{1}{\exp(-14.59-.086v)+\exp(-1.87+.07v)}$ | $\bar{g}_{HCN} = 1.01\text{e}{-3}$ (soma)<br>$\bar{g}_{HCN} = 5.10\text{e}{-4}$ (dendrite) |
| $I_{CaT}$ | $\bar{g}_{CaT} r^3 s (v - E_{Ca})$ | $\alpha_r = \frac{1}{1.7+\exp(-(v+28.2)/13.5)}$<br>$\alpha_s = \exp[-(v + 160.3)/17.8]$<br>$\alpha_d = \frac{1+\exp\left[\frac{(v+37.4)}{30}\right]}{240\left(0.5+\sqrt{0.25+\exp\left[\frac{(v+83.5)}{6.3}\right]}\right)}$ | $\beta_r = \frac{\exp(-(v+63)/7.8)}{1.7+\exp(-(v+28.8)/13.5)}$<br>$\beta_s = \left(\sqrt{.25 + \exp\frac{v+83.5}{6.3}} - .5\right) k_s$<br>$k_s = \exp[-(v + 160.3)/17.8]$<br>$\beta_d = \left(\sqrt{0.25 + \exp\frac{v+83.5}{6.3}} - 0.5\right) \alpha_d (v)$ | $\bar{g}_{CaT} = 0$ (soma)<br>$\bar{g}_{CaT} \in [1.17, 1.67] \times 10^3$ (dendrite)<br>$\left[Ca^{2+}\right]_{i0} = 1\text{e}{-4}$<br>$\tau_{Ca} = 185.7$ |
| $I_{CaL}$ | $g_{CaT} q^2 h (v - E_{Ca})$ | $h_\infty([Ca^{2+}]_i)=0.53+\frac{0.47}{1+\exp(\frac{[Ca^{2+}]_i-0.7}{0.15})}$<br>$q_\infty (v) = \frac{1}{1+\exp[-(24.6v)/11.3]}$ | $\tau_\infty \left([Ca^{2+}]_i\right) = 1220$<br>$\tau_q (v) = \frac{1.25}{\cosh[-0.03(v+37.1)]}$ | $\bar{g}_{CaL} = 9.50\text{e}{-4}$ (soma)<br>$\bar{g}_{CaL} \in [1.21, 18.7] \times 10^4$ (dendrite) |
| $I_{CaN}$ | $g_{CaN} q^2 (v - E_{Ca})$ | $u_\infty (v_j) = \frac{1}{1+\exp[(v_j+60)/12.5]}$ | $\tau_u (v) = 98 + \cosh[0.021 (10.1 - v)]$ | $\bar{g}_{CaN} = 1.15\text{e}{-3}$ (soma)<br>$\bar{g}_{CaN} = 4.79\text{e}{-4}$ (dendrite) |

based on the ratios reported in Shen and Johnson (2005) for the normal and dopamine-depleted conditions, taking into account the reduction of synaptic terminals reported in Chu et al. (2017). Absolute values for the synaptic conductances were hand-tuned to bring the mean population firing rates into the reported range for the rat in dopamine-depleted condition (Mallet et al., 2008b; Kita and Kita, 2011). Synapses from GPe neurons were located proximally, close to the soma. Synapses of GPe-STN afferents had a fast $GABA_A$ and a slower $GABA_B$ component. Rise and decay time constants for the $GABA_A$ conductance were based on Fan et al. (2012). Short-term plasticity parameters were chosen so that synapses exhibited short-term depression, as shown in Atherton et al. (2013). Parameters for the $GABA_B$ synapse were taken from the model by Destexhe and Sejnowski (1995), and the decay time constant K4 was adapted so that the $GABA_B$ conductance exhibited depression upon continued pre-synaptic stimulation.

## GPe Cell Model

GPe neurons were modeled using the baseline rat GPe neuron model by Gunay et al. (2008) (ModelDB accession number 114639). The model is based on a reconstructed morphology from the adult rat and contains nine types of ion channels with varying densities in the soma, dendrite, and axon initial segment:

$$\begin{aligned} I_{ion,j} = {} & I_{NaF} + I_{NaP} + I_{Kv2} + I_{Kv3} \\ & + I_{Kv4,f} + I_{Kv4,s} + I_{KCNQ} + I_{sKCa} \\ & + I_{CaHVA} + I_{HCN,f} + I_{HCN,s} + I_L \end{aligned} \tag{19}$$

where $I_{NaF}$ and $I_{NaP}$ are the transient fast-acting and persistent sodium current, $I_{Kv2}$ and $I_{Kv3}$ the slow and fast delayed rectifier potassium current, $I_{Kv4f}$ and $I_{Kv4s}$ the fast and slow component of the A-type, transient potassium current, $I_{KCNQ}$ the M-type potassium current, $I_{sKCa}$ the calcium-dependent potassium current, $I_{CaHVA}$ the high-threshold, non-inactivating calcium current (reflecting a mixture of L, N, and P/Q-type calcium channel types), and $I_{HCN,f}$ and $I_{HCN,s}$ the fast and slow component of the HCN channel. The equations governing the dynamics of the gating variables are listed in **Table 4**. The channel density distributions are those described in Gunay et al. (2008) for model *t9842*. As a source of noise, a current that with a Gaussian amplitude distribution, mean zero and standard deviation 0.0075 was added to the somatic compartment, to represent membrane voltage noise of similar amplitude the STN cell model, given the lower somatic input resistance of the STN model.

GPe neurons each had 10 excitatory afferents from STN neurons, 6 inhibitory afferents from GPe-GPe collaterals, and 30 inhibitory afferents from iMSN (**Table 5**). The location of synapses on GPe neurons and axonal propagation delays were based on experimental observations reported in the literature (**Table 1**). Relative magnitudes of synaptic conductances were chosen to bring the population firing rate into the reported range for the rat (Mallet et al., 2008b; Kita and Kita, 2011). Rise and decay time constants for AMPA conductances were set to 1 and 4 ms, respectively, and for $GABA_A$ conductances they were set to 2 and 5 ms, respectively. Synapses from STN neurons were located distally, in the dendritic tree.

**TABLE 3 |** STN model synaptic current equations.

| Current | Equation | Location | Parameters | |
|---|---|---|---|---|
| $I_{CTX-STN,AMPA}$ | $\bar{g}_{syn}s(v - E_{AMPA})$ | distal: $x \geq 100\mu m$ | $\tau_{rise} = 1$ | $\tau_{rec} = 200$ |
| | | | $\tau_{decay} = 4$ | $\tau_{facil} = 1$ |
| | | | $t_d = 5.9$ | $U_1 = 0.2$ |
| | | | $\bar{g}_{syn} = 4.44e{-}3$ | |
| $I_{CTX-STN,NMDA1}$ | $\bar{g}_{syn}ms(v - E_{NMDA})$ | distal: $x \geq 100\mu m$ | $\tau_{rise} = 3.7$ | $\tau_{rec} = 200$ |
| | | | $\tau_{decay} = 212$ | $\tau_{facil} = 1$ |
| | | | $t_d = 5.9$ | $U_1 = 0.2$ |
| | | | $\bar{g}_{syn} = 5.04e{-}3$ | |
| $I_{CTX-STN,NMDA2}$ | $\bar{g}_{syn}ms(v - E_{NMDA})$ | proximal: $x < 120\mu m$ | $\tau_{rise} = 3.7$ | $\tau_{rec} = 200$ |
| | | | $\tau_{decay} = 80$ | $\tau_{facil} = 1$ |
| | | | $t_d = 5.9$ | $U_1 = 0.2$ |
| | | | $\bar{g}_{syn} = 5.04e{-}3$ | |
| $I_{GPE-STN,GABA_A}$ | $\bar{g}_{syn}s(v - E_{GABA_A})$ | proximal: $x < 120\mu m$ | $\tau_{rise} = 2$ | $\tau_{rec} = 400$ |
| | | | $\tau_{decay} = 7$ | $\tau_{facil} = 1$ |
| | | | $t_d = 2.0$ | $U_1 = 0.2$ |
| | | | $\bar{g}_{syn} = 18e{-}3$ | |
| $I_{GPE-STN,GABA_B}$ | $\bar{g}_{syn}\frac{G^n}{G^n+K_d^n}(v - E_{GABA_B})$ | proximal: $x < 120\mu m$ | $\tau_{rise} = 5$ | $\tau_{rec} = 400$ |
| | | | $\tau_{decay} = 25$ | $\tau_{facil} = 1$ |
| | | | $t_d = 2.0$ | $U_1 = 0.2$ |
| | | | $\bar{g}_{syn} = 3.75e{-}3$ | $K_3 = 0.098$ |
| | | | $n = 4$ | $K_4 = 6.25e{-}3$ |
| | | | | $K_d = 1.4$ |

They consisted of an AMPA component and were modeled using Tsodyks-Markram dynamics. The parameters describing short-term plasticity dynamics were chosen to match traces reported in Hanson and Jaeger (2002). Synapses from GPe were located proximally, near the soma and had both a fast $GABA_A$ component with Tsodyks-Markram dynamics, and a slow metabotropic $GABA_B$ component. Short-term plasticity parameters were chosen so that synapses exhibited short-term depression (Miguelez et al., 2012). Synapses from striatal neurons had a $GABA_A$ component and were made to exhibit short-term facilitation based on Miguelez et al. (2012).

## Modeling the Parkinsonian State

To model the parkinsonian state, the biophysical properties of the network and cell models were modified based on experimental observations made in the dopamine depleted and control conditions as reported in the literature. Various biophysical parameters, including synaptic strengths and time constants are affected by dopamine depletion, and were adjusted as detailed below. Scaling factors for synaptic and ionic conductances were set to experimentally reported values where available. Otherwise they were chosen to bring the mean population firing rates into physiological ranges reported for the rat in a state of cortical activation during light anesthesia (Mallet et al., 2008b; Kita and Kita, 2011).

The mean firing rate of STN surrogate spike sources was increased from 14.6 to 29.5 Hz in the parkinsonian state (Mallet et al., 2008b). The peak $GABA_A$ and $GABA_B$ conductance of GPe to STN synapses was increased by 50% and the $GABA_B$ decay time constant increased by 2 ms to model the increase in the number of contacts, vesicle release probability, and decay kinetics of GPe afferents (Fan et al., 2012). To model the reduction in cortico-STN axon terminals and their dendritic targets (Chu et al., 2017; Wang et al., 2018) the number of CTX-STN afferents was reduced to 70% of the normal condition, corresponding to the ratio of vGluT1 expression in the normal and dopamine depleted condition used to label axon terminals (Chu et al., 2017). To model functional strengthening of remaining synapses, the AMPA and NMDA peak conductances of remaining synapses were multiplied by the ratio of the current scaling factors reported in Shen and Johnson (2005) to the fraction of remaining synapses. The effect of functional strengthening and weakening of the CTX-STN projection was further investigated by systematically varying the peak synaptic conductances in the simulations experiments. Finally, HCN currents were reduced by 50% to model reduced depolarization and spontaneous activity after dopamine depletion (Zhu et al., 2002; Cragg et al., 2004) and modulation of HCN current by D2R receptors (Yang et al., 2016).

In GPe neurons the peak AMPA conductance of STN afferents was increased by 50% to model the modulatory effect of dopamine on glutamergic excitatory currents (Johnson and Napier, 1997; Hernández et al., 2006; Kita, 2007). The strengthening of GPe-GPe collaterals (Miguelez et al., 2012; Nevado-Holgado et al., 2014) was modeled by increasing the peak $GABA_A$ and $GABA_B$ conductances by 50%. The mean firing rate of GPe surrogate spike sources was decreased from 33.7 to 14.6 (Mallet et al., 2008a). Finally, the HCN channel conductance was decreased by 50% in accordance with experimental data (Chan et al., 2011).

In simulations without oscillatory inputs, cortical projection neurons were modeled as Poisson spike generators firing at 10 Hz, a multiple of the experimentally reported rate of 2.5 Hz (Li et al., 2012), so that each synapse represented the combined inputs of four pre-synaptic neurons (making use of the additive property of the Poisson distribution). In simulations with oscillatory inputs, oscillatory spike trains were generated as follows: on top of the aforementioned background firing pattern, bursts were added in each period of a regular oscillation at the chosen oscillation frequency. In each period of the oscillation 10% of neurons were selected randomly to emit a burst. The onset time of the burst was the same in each selected neuron, so that bursts occurred in-phase between neurons, but the number of spikes in a burst was variable with inter-spike intervals sampled from the interval $[5, 6]$ ms. All background spikes occurring in a time window centered on a burst were deleted to prevent unrealistically high inter-spike intervals. **Figure 7D** shows a rastergram with representative spike trains generated using this method.

The increase in excitability and spontaneous activity of iMSN (Kita and Kita, 2011; Fieblinger et al., 2014) was modeled by increasing the mean firing rate of the Poisson spike generators from 1.5 to 6.64 Hz. In experiments where iMSN cells fired oscillatory bursts the same algorithm as described for cortical projection neurons was used. The modulation of GABAergic transmission from iMSN to GPe neurons (Cooper and Stanford,

**TABLE 4 |** GPe model intrinsic current equations from Gunay et al. (2008).

| Current | Equation | Gate | $m_0$ | $\theta_{m\infty}$ | $\sigma_{m\infty}$ | $\tau_0$ | $\tau_1$ | $\theta_{m\tau}$ | $\sigma_{m0}$ | $\sigma_{m1}$ | Additional parameters |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $I_{NaF}$ | $\bar{g}_{NaF}m^3hs(v-E_{Na})$ | m | 0 | –39 | 5 | 0.028 | 0.028 | N/A | N/A | N/A | $\bar{g}_{NaF}=0.035$ (soma) |
| | | h | 0 | –48 | –2.8 | 0.025 | 4 | –43 | 10 | –5 | $\bar{g}_{NaF}=0.035$ (dendrite) |
| | | s | 0.15 | –40 | –5.4 | 10 | 1000 | –40 | 18.3 | –10 | $\bar{g}_{NaF}=0.5$ (axon) |
| $I_{NaP}$ | $\bar{g}_{NaP}m^3hs(v-E_{Na})$ | m | 0 | –57.7 | 5.7 | 0.03 | 0.146 | –42.6 | 14.4 | –14.4 | $\bar{g}_{NaP}=10.15e{-}3$ (soma) |
| | | h | 0.154 | –57 | –4 | 10 | 17 | –34 | 26 | –31.9 | $\bar{g}_{NaP}=10.15e{-}3$ (dendrite) |
| | | s | 0 | –10 | –4.9 | N/A | N/A | N/A | N/A | N/A | $\bar{g}_{NaF}=4e{-}3$ (axon) |
| $I_{Kv2}$ | $\bar{g}_{Kv2}m^4h(v-E_K)$ | m | 0 | –33.2 | 9.1 | 0.1 | 30 | –33.2 | 21.7 | –13.9 | $\bar{g}_{Kv2}=0.1e{-}3$ (soma, dendrite) |
| | | h | 0.2 | –20 | –10 | 3400 | 3400 | N/A | N/A | N/A | $\bar{g}_{Kv2}=64e{-}3$ (axon) |
| $I_{Kv3}$ | $\bar{g}_{Kv3}m^4h(v-E_K)$ | m | 0 | –26 | 7.8 | 0.1 | 14 | –26 | 13 | –12 | $\bar{g}_{Kv3}=1e{-}3$ (soma, dendrite) |
| | | h | 0.6 | –20 | –10 | 7 | 33 | 0 | 10 | –10 | $\bar{g}_{Kv3}=128e{-}3$ (axon) |
| $I_{Kv4,f}$ | $\bar{g}_{Kv4,f}m^4h(v-E_K)$ | m | 0 | –49 | 12.5 | 0.25 | 7 | –49 | 29 | –29 | $\bar{g}_{Kv4,f}=2e{-}3$ (soma) |
| | | h | 0 | –83 | –10 | 7 | 21 | –83 | 10 | –10 | $\bar{g}_{Kv4,f}=4e{-}3$ (dendrite) |
| | | | | | | | | | | | $\bar{g}_{Kv4,f}=160e{-}3$ (axon) |
| $I_{Kv4,s}$ | $\bar{g}_{Kv4,s}m^4h(v-E_K)$ | m | 0 | –49 | 12.5 | 0.25 | 7 | –49 | 29 | –29 | $\bar{g}_{Kv4,s}=3e{-}3$ (soma) |
| | | h | 0 | –83 | –10 | 50 | 121 | –83 | 10 | –10 | $\bar{g}_{Kv4,s}=6e{-}3$ (dendrite) |
| | | | | | | | | | | | $\bar{g}_{Kv4,s}=240e{-}3$ (axon) |
| $I_{KCNQ}$ | $\bar{g}_{KCNQ}m^4h(v-E_K)$ | m | 0 | –61 | 19.5 | 6.7 | 100 | –61 | 35 | –25 | $\bar{g}_{KCNQ}=20e{-}5$ (soma, dendrite) |
| | | | | | | | | | | | $\bar{g}_{KCNQ}=4e{-}5$ (axon) |
| $I_{CaHVA}$ | $\bar{g}_{CaHVA}m(v-E_{Ca})$ | m | 0 | –20 | 7 | 0.2 | 0.2 | –20 | N/A | N/A | $\bar{g}_{CaHVA}=3e{-}5$ (soma, thick dendrites) |
| | | | | | | | | | | | $\bar{g}_{CaHVA}=4.5e{-}5$ (medium dendrites) |
| | | | | | | | | | | | $\bar{g}_{CaHVA}=9e{-}5$ (thin dendrites) |
| | | | | | | | | | | | $\left[Ca^{2+}\right]_{i0}=5e{-}5$ |
| | | | | | | | | | | | $\tau_{Ca}=1$ |
| $I_{HCN,f}$ | $\bar{g}_{HCN,f}m(v-E_h)$ | m | 0 | –76.4 | –3.3 | 0 | 3625 | –76.4 | 6.56 | –7.48 | $\bar{g}_{HCN,f}=1e{-}4$ (soma, dendrite) |
| $I_{HCN,s}$ | $\bar{g}_{HCN,s}m(v-E_h)$ | m | 0 | –87.5 | –4 | 0 | 6300 | –87.5 | 8.9 | –8.2 | $\bar{g}_{HCN,f}=2.5e{-}4$ (soma, dendrite) |

**TABLE 5 |** GPe model synaptic current equations.

| Current | Equation | Location | Parameters | |
|---|---|---|---|---|
| $I_{STN-GPE,AMPA}$ | $\bar{g}_{syn}s(v-E_{AMPA})$ | distal: $x \geq 100\mu m$ | $\tau_{rise}=1$ | $\tau_{rec}=200$ |
| | | | $\tau_{decay}=4$ | $\tau_{facil}=800$ |
| | | | $t_d=2$ | $U_1=0.1$ |
| | | | $\bar{g}_{syn}=3.75e{-}4$ | |
| $I_{GPE-GPE,GABA_A}$ | $\bar{g}_{syn}s(v-E_{GABA_A})$ | proximal: $x < 200\mu m$ | $\tau_{rise}=2$ | $\tau_{rec}=400$ |
| | | | $\tau_{decay}=5$ | $\tau_{facil}=1$ |
| | | | $t_d=0.5$ | $U_1=0.2$ |
| | | | $\bar{g}_{syn}=2e{-}4$ | |
| $I_{GPE-GPE,GABA_B}$ | $\bar{g}_{syn}\frac{G^n}{G^n+K_d^n}(v-E_{GABA_B})$ | proximal: $x < 200\mu m$ | $\tau_{rise}=5$ | $K_3=0.098$ |
| | | | $\tau_{decay}=25$ | $K_4=6.25e{-}3$ |
| | | | $t_d=0.5$ | $K_d=1.4$ |
| | | | $\bar{g}_{syn}=0.4e{-}4$ | $n=4$ |
| $I_{iMSN-GPE,GABA_A}$ | $\bar{g}_{syn}s(v-E_{GABA_A})$ | proximal: $x < 200\mu m$ | $\tau_{rise}=2$ | $\tau_{rec}=1$ |
| | | | $\tau_{decay}=5$ | $\tau_{facil}=200$ |
| | | | $t_d=5$ | $U_1=0.3$ |
| | | | $\bar{g}_{syn}=3e{-}4$ | |

2001; Shin et al., 2003) was modeled by increasing the initial release probability and the peak $GABA_A$ conductance of synapses by 50%.

## Simulation Details

The model was simulated in the NEURON simulation environment (Hines and Carnevale, 1997) and implemented in Python. The default fixed time step integrator with a time step of 0.025 ms was used for all simulations. Compartmental membrane voltages were initialized to a random value between −63 and −73 mV in GPe and between −60 and −70 mV in STN cells. Gating variables were initialized to their equilibrium values for the initial membrane voltage. Simulation data for the first 2,000 ms of each simulation were discarded, and the analyzed intervals were of duration 4,000 ms unless otherwise noted. Simulations were run on the UCD Sonic cluster using 8 parallel processes per simulation on a single computing node, consisting of two Intel Ivybridge E5-2660 v2 CPUs (10 cores per CPU).

## Signal Analysis

Signal analyses were performed using the SciPy toolbox (Jones et al., 2001) for Python. Power spectral densities (PSDs) were

calculated using Welch's periodogram method, using overlapping segments of 2 s duration with 50% overlap and a Hanning window. Given the sampling period of 0.05 ms this led to a frequency resolution of 0.5 Hz. The population PSD was calculated as the mean PSD of all somatic membrane voltages. The instantaneous phase of each population was estimated by applying the Hilbert transform to the average somatic membrane voltage of cells in the population, after band-pass filtering using a neutral-phase filter (Butterworth filter, 4th order, command *sosfiltfilt*) in an 8 Hz wide frequency band centered on the dominant oscillation frequency. For populations that were modeled as surrogate spike trains (cortex and striatum), artificial membrane voltage signals were first constructed by convolving the spike trains with a typical action potential waveform. Bursts were detected using a simple algorithm where a burst consisted of a minimum of four spikes with inter-spike intervals (ISIs) $\leq$ 20 ms.

# RESULTS

The STN-GPe pacemaker hypothesis was first investigated by modeling cortical inputs to the STN as Poisson spike generators without any periodic or oscillatory component. Cortical patterning of neural activity in the STN-GPe network via the hyperdirect pathway was then investigated by modeling cortical input to the STN inputs as periodically bursting spike trains. To investigate whether changing the excitation-inhibition balance in STN and GPe contributed to changes in spontaneous synchronization and functional coupling between nuclei, the ratio of excitation and inhibition was systematically increased by altering the strength of individual projections between nuclei. The ratio of total excitatory to inhibitory synaptic currents (E/I ratio) was altered by scaling the peak conductance of all synapses belonging to a given projection known to be strengthened or weakened by dopamine depletion. The role of additional oscillatory inputs entering the STN-GPe network via the indirect striato-pallidal pathway and their phase relationship to cortical inputs were then investigated.

## The Balance of Excitation and Inhibition Balance in the STN Affects the Oscillation Frequency of the STN-GPe Network and Firing Mode of STN Neurons

Increasing the strength of the CTX-STN projection by increasing the conductance of cortico-subthalamic synapses revealed parameter regimes that favored low frequency bursting in STN neurons and phase-locking to an emergent beta-band rhythm in the STN-GPe network (**Figures 2A–D**). For lower values of synaptic conductance the network exhibited synchronous oscillatory activity at 12–13 Hz (**Figure 2A**), with both STN and GPe neurons entrained to the oscillation (**Figures 2Bi–iii**). This high entrainment regimen coincided with low neuronal firing rates (**Figure 2E**) where short spike sequences, mostly singlets and doublets, showed a high phase preference as evidenced by the high population and individual neuronal phase vector lengths (**Figure 2Bii**).

Increasing the synaptic conductance caused a proportional increase in excitatory current to the STN (**Figure 2E**, blue area), with a corresponding increase in inhibition (red area) as a result of the negative feedback structure of the STN-GPe loop. However, because the GPe population exhibited a saturating population firing rate curve (**Figure 2Eii**), feedback inhibition to STN was outpaced by cortical excitation, resulting in a shift to net excitation (E/I $>$ 1). This saturating firing rate curve in the GPe was a result of two negative feedback mechanisms that have a homeostatic effect on the GPE's E/I ratio: reciprocal inhibition through intra-GPe colaterals and short-term depression of STN to GPe synapses (Hanson and Jaeger, 2002). The increase of excitatory drive in the network increased the frequency at which oscillations emerged within the network (**Figure 2Aii**, peak in the PSD is shifted), though the level of synchronization of neurons was relatively weak. This was particularly the case in the STN, as evidenced by the low phase vector lengths (**Figures 2Cii,Dii,Eii**). Despite the lower vector lengths, reflecting more dispersed spike timings within a period of the oscillation, spikes in both STN and GPe neurons showed a consistent phase preference with respect to the ongoing oscillation, as evidenced by the alignment of individual neuronal and population phase vector. The STN population vector led that of GPe by 45 degrees indicating that STN neurons excited GPe neurons which responded with a delay of 10 ms, resulting in a wave of inhibition to the STN with a long recovery period comparable to the oscillation period. Although excitation outpaced inhibition in STN neurons, higher inhibitory currents resulted in increased transient inhibition of STN dendrites, engaging the ion channels underlying burst responses. This brought STN neurons into a slow burst firing mode characterized by sparse, strong bursts (**Figure 2Ci**). These low-frequency fluctuations in firing rate were transmitted to GPe neurons as evident in the power spectra of both nuclei (**Figures 2Ai,ii**).

Increasing the strength of GPe-GPe colaterals (**Figure 3**) similarly increased the level of excitation of STN neurons but by a different mechanism. By increasing self-inhibition within the GPe, and thereby decreasing inhibition of targets in the STN (**Figure 3D**, red area), the E/I ratio in both populations moved in opposite directions. As the E/I ratio in STN increased toward dominant excitation (**Figure 3D**), neural activity shifted from strong low-frequency bursting (characterized by a high intra-burst firing rate and high low-frequency power) (**Figures 3Ai,Bi,F**) toward more regular firing with decreasing coefficient of variation of inter-spike intervals ($CV_{ISI}$) and intra-burst firing rate (**Figures 3Ci,F**). The E/I ratio and population firing rate in the GPe showed a saturating characteristic (**Figure 3Dii**) caused by the negative feedback structures inherent in the loop as before (**Figure 3D**) as it was progressively disinhibited. GPe neurons were more strongly entrained to the emergent oscillation (17–26 Hz) whereas STN spiking showed a weaker phase preference (**Figures 3Bii,Cii,E**). This result was the same whether the instantaneous phase was extracted from the STN or GPe population.

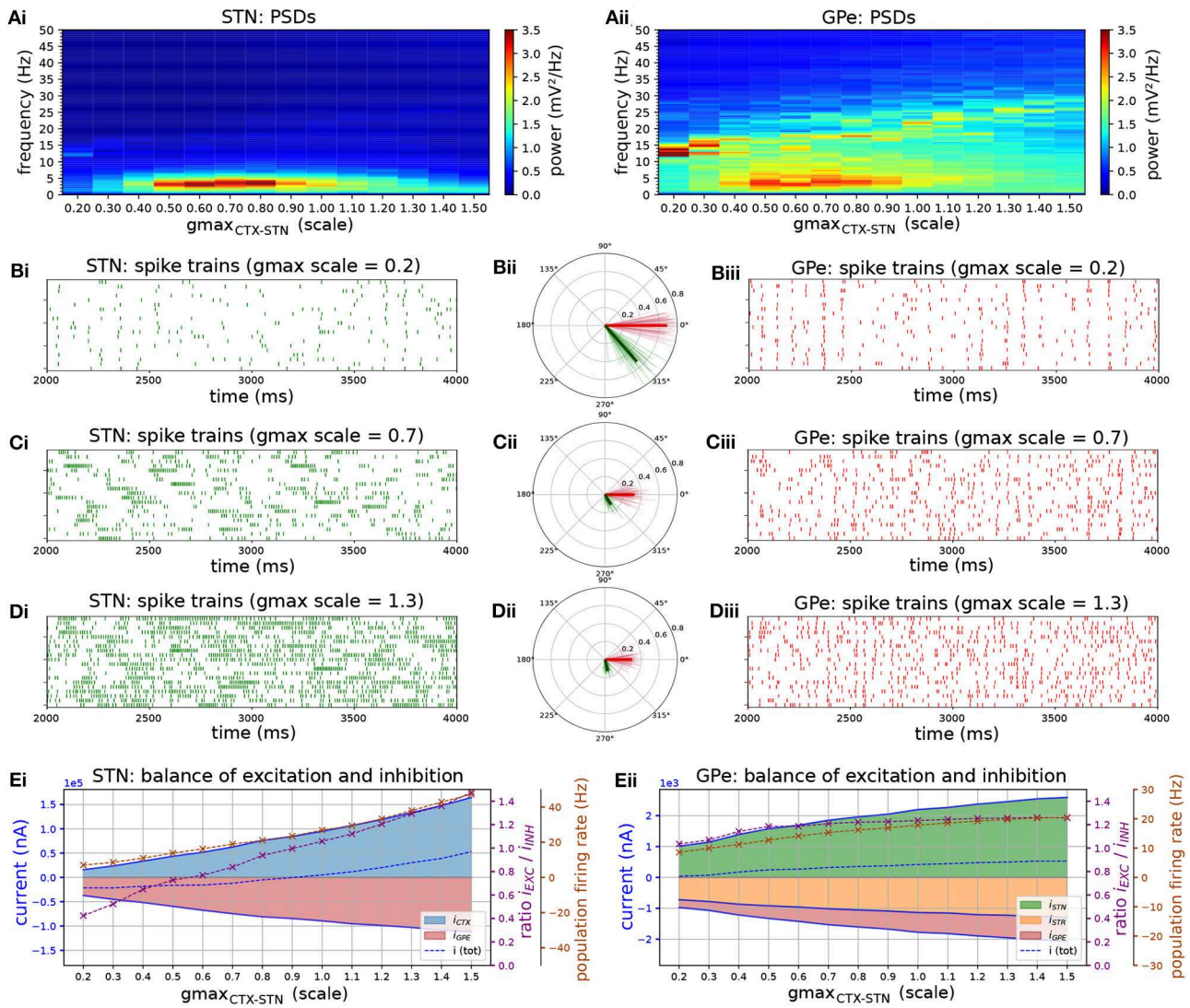


**FIGURE 2 |** The level of excitation by cortex determines firing patterns and oscillation frequency in the autonomous STN-GPe network. Behavior of the autonomous STN-GPe network for increasing values of the CTX to STN synaptic conductance. **(A)** Mean PSD of the somatic membrane voltages of STN **(Ai)** and GPe **(Aii)** neurons. **(B–D)** Representative spike trains and phase vectors for STN (column i, green) and GPe population (column iii, red) for three different scale factors of the CTX to STN conductance [scale 0.2; 0.7; 1.3 in rows **(C–E)**, respectively]. Column ii shows phase vectors of the STN and GPe populations (in green; red, respectively, mean population vectors plotted as thick solid lines and cell vectors as thin transparent lines) reflecting phase locking to the instantaneous GPe phase. Phase vectors were measured with respect to the instantaneous phase of the GPe population extracted using a bandpass filter with passband of 8 Hz centered on the frequency bin with maximum power in the 13–30 Hz band. **(E)** Balance of excitation and inhibition in the STN **(Ei)** and GPe **(Eii)** based on synaptic currents recorded in three neurons. Population firing rate (brown), E/I ratio (purple), and net synaptic current (blue). Shaded areas represent estimated total synaptic current from one pre-synaptic population during a simulation. Total current was estimated by recording all synapses on 3 randomly selected cells in each population and adjusting for the true number of cells.

## Strength and Time-Course of GPe-STN Inhibition Controls Bursting and Phase-Locking in STN Neurons

Following dopamine depletion the inhibitory GPe-STN connection is strengthened by a proliferation of synapses and increased decay kinetics of GABA currents (Fan et al., 2012). Moreover, the expression of both $GABA_A$ (Fan et al., 2012) and $GABA_B$ (Shen and Johnson, 2005) receptors is upregulated leading to larger evoked synaptic currents. To investigate the effects of increased inhibition and altered kinetics of inhibitory post-synaptic currents (IPSC) in STN neurons on network activity patterns, an increase in the $GABA_A$ and $GABA_B$ conductances was simulated and the relative contribution of both currents was altered.

Increasing the conductance of both $GABA_A$ and $GABA_B$ synapses lead to an increase in low-frequency bursting of STN neurons (**Figures 4A–C**). Bursting was periodic at

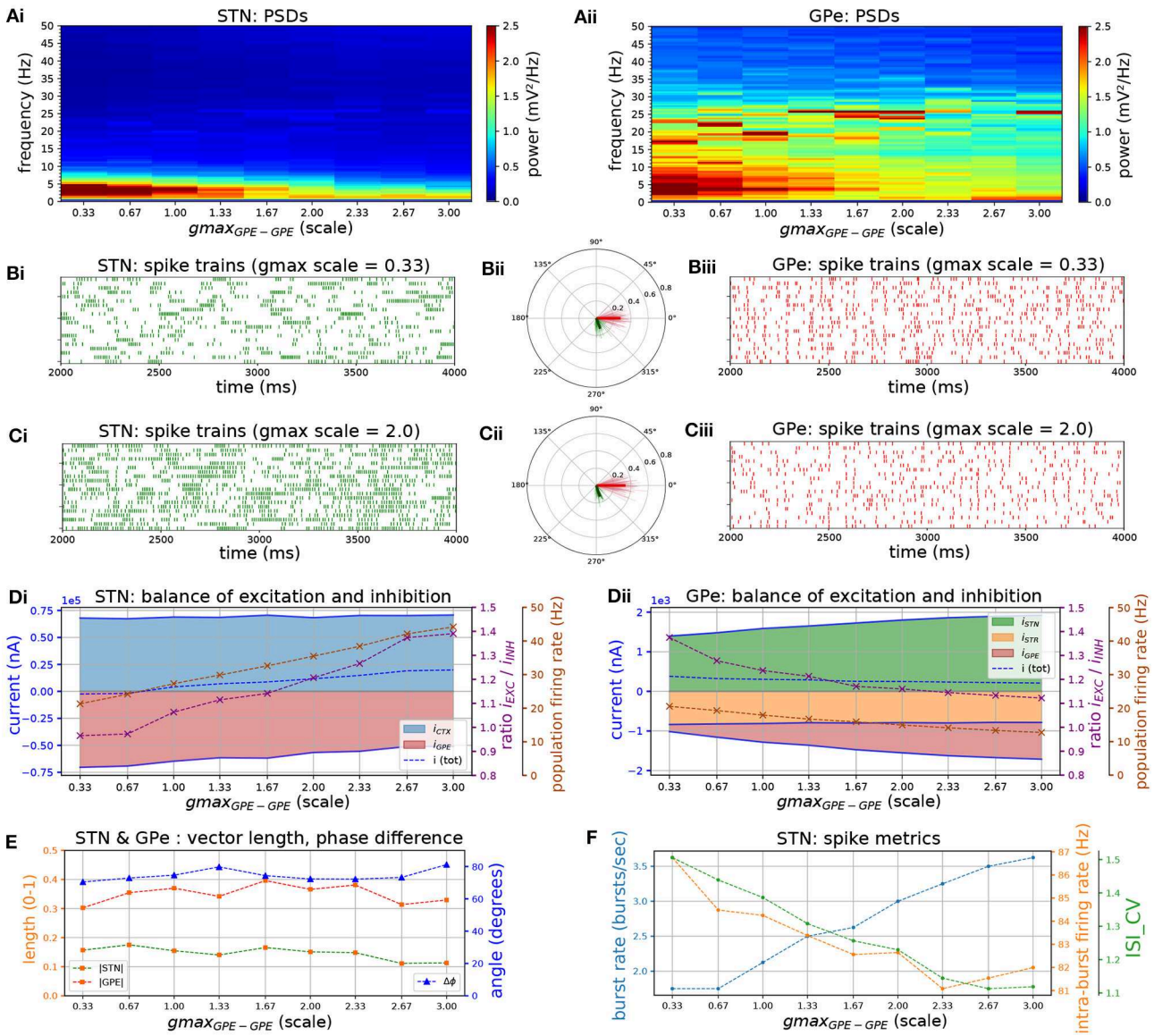


**FIGURE 3 |** Increasing the level of collateral GPe-GPe inhibition shifts the excitation-inhibition balance in STN and GPe in opposite directions. Behavior of the autonomous STN-GPe network for increasing values of GPe-GPe synaptic conductance. **(A)** Mean PSD of the somatic membrane voltages of STN **(Ai)** and GPe **(Aii)** neurons. **(B,C)** Representative spike trains and phase vectors for STN (column i, green) and GPe population (column iii, red) for two values of the GPe to GPe conductance [scale 0.33; 2.0 in rows **(B,C)**, respectively]. Column ii shows phase vectors of the STN and GPe populations (in green; red, respectively, mean population vectors plotted as thick solid lines and cell vectors as thin transparent lines) reflecting phase locking to the instantaneous GPe phase. **(D)**: Balance of excitation and inhibition in the STN **(Di)** and GPe **(Dii)** based on synaptic currents recorded in three neurons. Mean population firing rate (brown), E/I ratio (purple), and net synaptic current (blue). Shaded areas represent estimated total synaptic current from one pre-synaptic population during a simulation. **(E)** Population vector length and angle of STN and GPe population (green; red, respectively). **(F)** Metrics that characterize bursting in STN neurons: median burst rate, intra-burst firing rate, and coefficient of variation of ISIs across all STN cells.

low frequencies (∼ 2–5 Hz) but was not synchronized between cells (**Figure 4Ci**). Increasing the conductance also shifted the firing mode of STN neurons toward longer bursts with higher intra-burst firing rate against a lower background firing rate, characterized by a high coefficient of variation of ISIs (**Figures 4D,F**). Bursting with high intra-burst firing rates is mediated by a shift toward net inhibition in STN neurons (**Figure 4D**), leading to increased availability of voltage-sensitive $Na^+$ and $Ca^{2+}$ channels through de-inactivation at hyperpolarized membrane voltages (Baufreton et al., 2005; Gillies and Willshaw, 2005; Hallworth and Bevan, 2005). The GPe neuron model does not possess the same high density of $Ca^{2+}$ channels that underlies plateau potentials and strong bursting, and therefore has

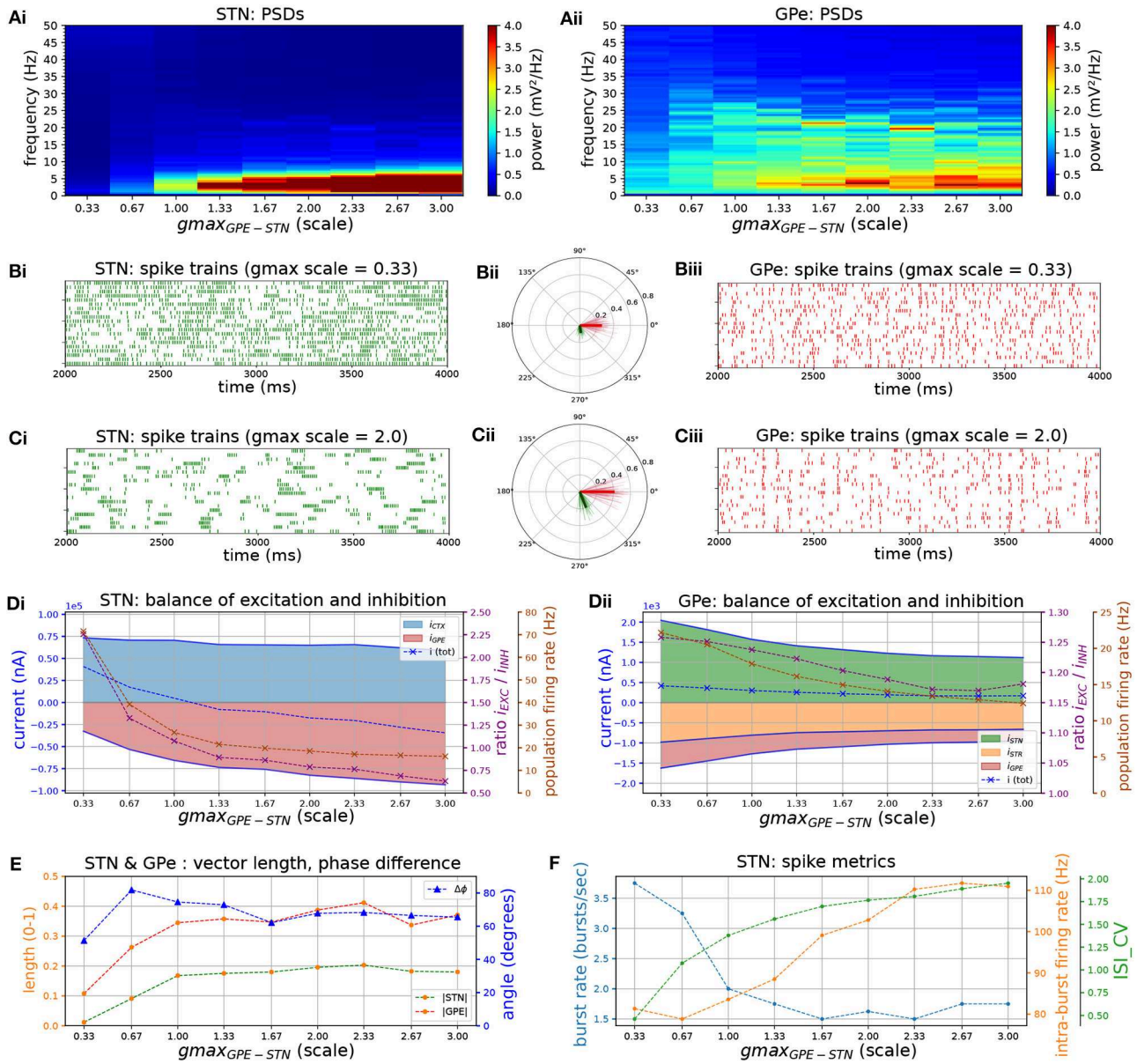


**FIGURE 4 |** Increasing the level of GPe-STN inhibition shifts STN to a low-frequency burst firing mode. Behavior of the autonomous STN-GPe network for increasing values of the GPe to STN synaptic conductance. **(A)** Mean PSD of the somatic membrane voltages of STN **(Ai)** and GPe **(Aii)** neurons. **(B,C)** Representative spike trains and phase vectors for STN (column i, green) and GPe population (column iii, red) for two values of the GPe to STN conductance [scale 0.33; 2.0 in rows **(C,D)**, respectively]. Column ii shows phase vectors of the STN and GPe populations (in green; red, respectively, mean population vectors plotted as thick solid lines and cell vectors as thin transparent lines) reflecting phase locking to the instantaneous GPe phase. **(D)** Balance of excitation and inhibition in the STN **(Di)** and GPe **(Dii)** based on synaptic currents recorded in three neurons. Population firing rate (brown), E/I ratio (purple), and net synaptic current (blue). Shaded areas represent estimated total synaptic current from one pre-synaptic population during a simulation. **(E)** Population vector length and angle of STN (green) and GPe (red) population. **(F)** Metrics that characterize bursting in STN neurons: median burst rate, intra-burst firing rate, and coefficient of variation of ISIs across all STN cells.

a lower tendency toward burst firing. While STN neurons were more weakly entrained to the beta oscillation they preferentially fired in an interval leading the GPe by 65 degrees (**Figures 4B,C,E**). The shift toward low-frequency, fast bursting coincided with an increase in synchronization in the network, as measured by the population vector length of the STN and GPe.

To investigate the effect of IPSC kinetics on the generation of beta oscillations within the network, the relative strength of the $GABA_A$ and $GABA_B$-mediated current was changed by

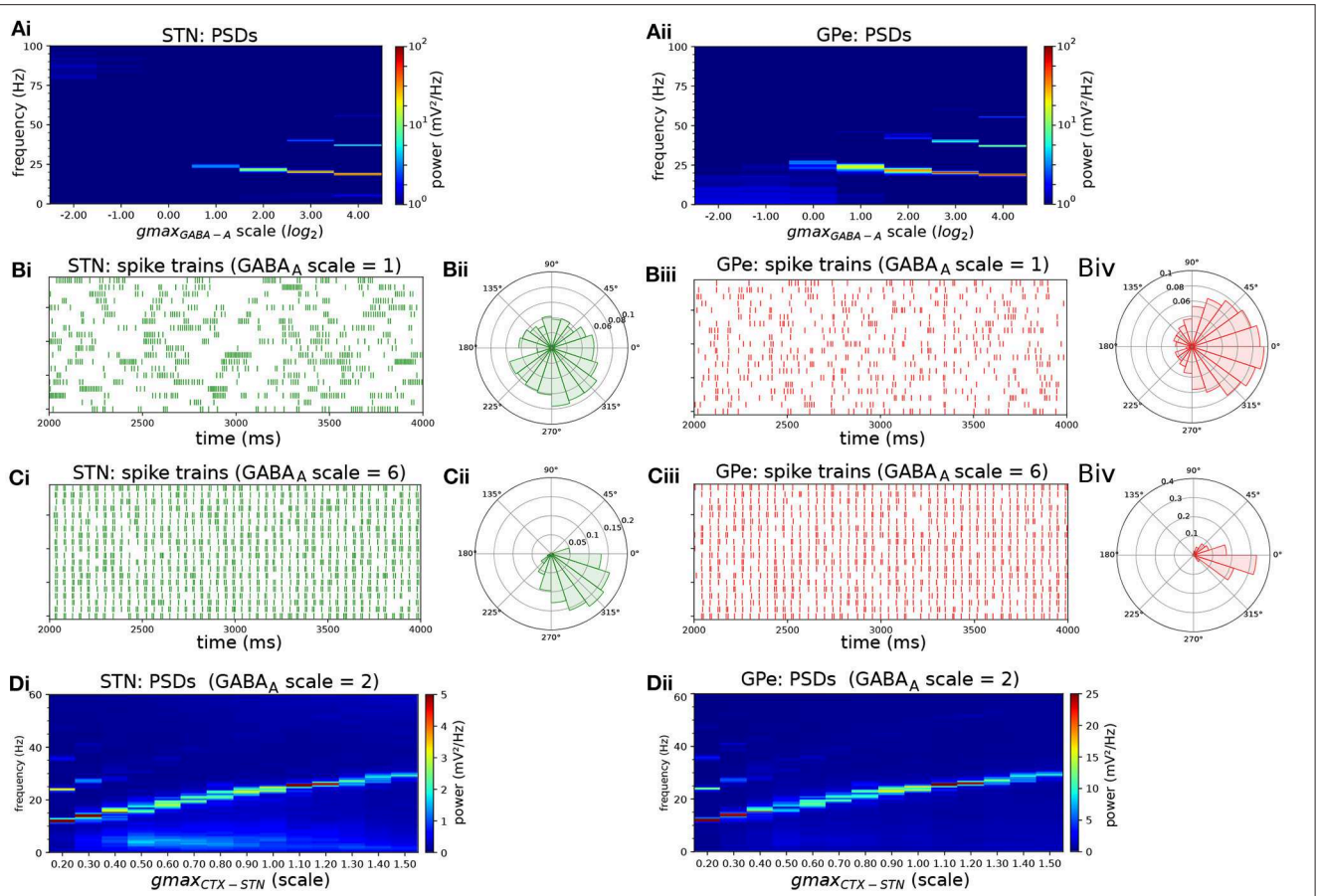


**FIGURE 5 |** Endogenous oscillations in the STN-GPe network are strengthened by shifting the GPe-STN synaptic current from slow $GABA_B$ receptors to fast $GABA_A$ receptors. Behavior or the STN-GPe network for increasing values of the $GABA_A$ to $GABA_B$ conductance ratio. The $GABA_B$ conductance of the GPe to STN projection was decreased by 50% and the $GABA_A$ conductance was increased progressively. **(A)** Mean PSD of the somatic membrane voltages of STN **(Ai)** and GPe **(Aii)** neurons. **(B,C)** Representative spike trains and phase histograms of STN (green) and GPe neurons (red) in baseline model without scaling of conductances **(B)** and model where $GABA_A$ conductance was scaled by a factor 6 and $GABA_B$ conductance was scaled by factor 0.2, chosen so that the E/I ratio was close to that in the baseline model (**B**: baseline model, E/I ratio was 0.89; 0.97 in STN, GPe respectively; **C**: scaled conductances, ratio was 0.89; 0.93). The presence of stronger $GABA_B$ currents results in higher phase dispersion **(B)** compared to the case with weaker $GABA_B$ currents and stronger, fast $GABA_A$ currents **(C)**. **(D)** Mean PSD of somatic membrane voltages of STN **(Di)** and GPe **(Dii)** neurons for increasing CTX-STN conductance and adjusted $GABA_A$ to $GABA_B$ ratio. The $GABA_B$ conductance of GPe to STN synapses was halved, and the $GABA_A$ conductance was doubled.

decreasing the $GABA_B$ conductance by 50% and increasing the $GABA_A$ conductance progressively (**Figure 5**). As this increased the level of inhibition in STN neurons, it resulted in a small shift in the oscillation frequency across the parameter sweep (**Figure 5A**). The simulations results showed that the slow nature of the $GABA_B$-mediated current prevented GPe neurons from patterning their targets with short duration IPSC required for strong entrainment in the 20–30 Hz range. When the $GABA_A$ conductance was increased, and the $GABA_B$ conductance decreased accordingly, both STN and GPe neurons entrained strongly to the beta rhythm as evident in phase histograms and spike trains (**Figures 5B,C**). When the experiment of **Figure 2** was repeated in the adjusted network with a higher $GABA_A$ to $GABA_B$ ratio, the oscillation frequency in both STN and GPe also showed a clear sensitivity to the strength of the Poisson distributed cortical excitatory input (**Figure 5D**).

## STN-GPe Network Shows Resonant Properties and Phase Locks to Cortical Beta Inputs

The degree of phase locking of the STN-GPe network to synchronous cortical rhythms and its sensitivity to intrinsic network parameters was then examined. The network was simulated with cortical inputs modeled as spike trains exhibiting sparse, synchronous bursts. The frequency of the synchronous cortical inputs was first increased from 3 to 60 Hz and the frequency response and phase locking strength of the STN-GPe loop was estimated (**Figures 6A–C**). Spectral power and phase locking, measured by the population vector length, were strongest when the cortical oscillation frequency was close to the network's endogenous oscillation frequency (**Figures 6B,C**), indicating a resonance effect. Spectral power at the oscillation frequency was increased considerably above that observed for

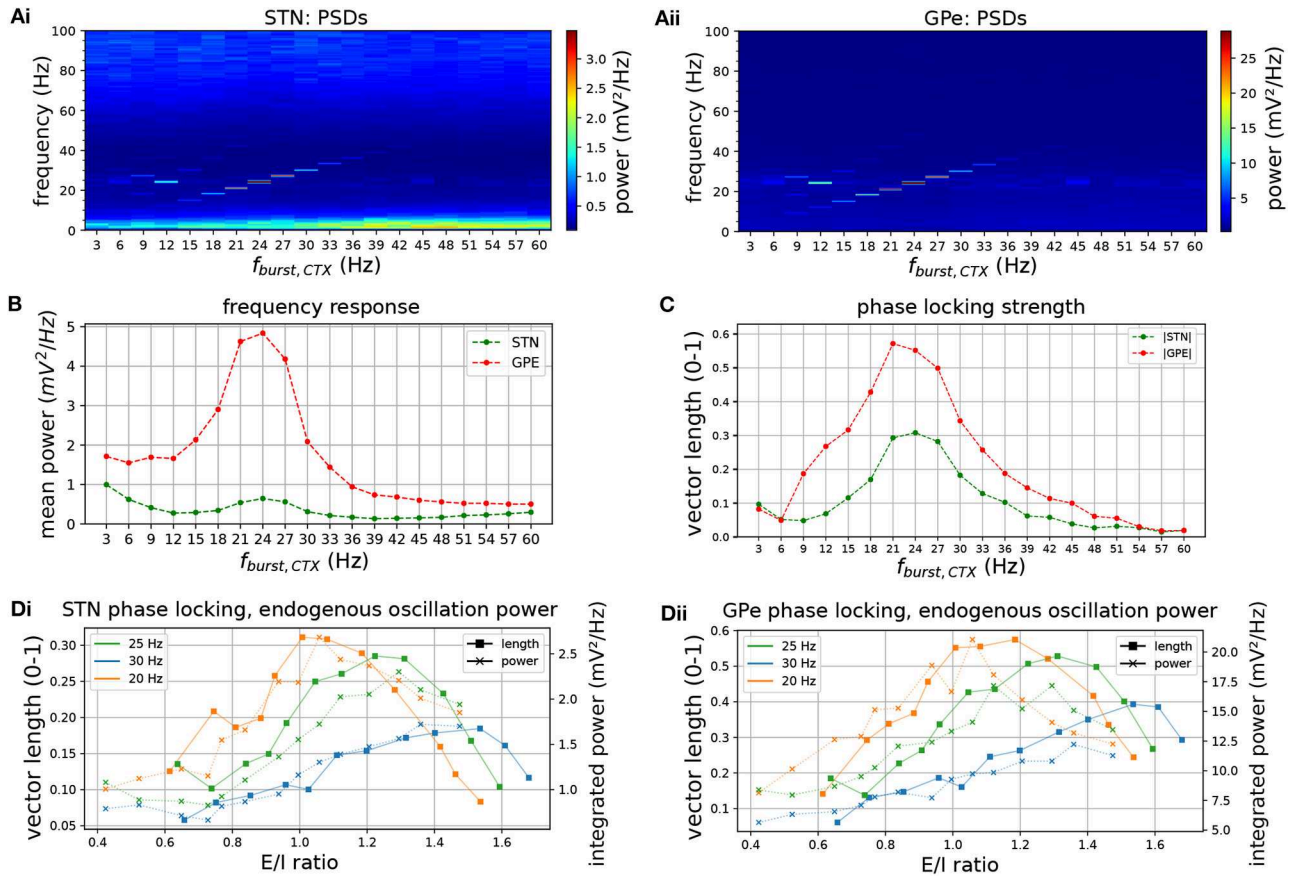


**FIGURE 6 |** Frequency response and phase locking of the STN-GPe network to cortical oscillatory bursting inputs. Sweep of cortical oscillation frequency (top row) and phase locking to cortical oscillations for increasing CTX-STN synapse strength (bottom row). **(A)** Mean PSD of somatic membrane voltages in STN **(Ai)** and GPe **(Aii)** for increasing oscillatory bursting frequency. **(B)** Mean PSD of the somatic membrane voltages of STN (green) and GPe (red) neurons, averaged within a 5 Hz wide frequency band centered on the cortical oscillation frequency. **(C)** Population vector length, indicating strength of phase locking to the cortical oscillation of STN (green) and GPe (red) neurons. **(D)** Change in population vector length (solid lines) for a fixed cortical oscillation frequency (20 Hz, 25 hz, 30 Hz in green, blue, orange, respectively) and increasing CTX-STN input synaptic conductance, reflected in an increased ratio of excitation to inhibition (E/I ratio). Endogenous oscillation power in simulations without oscillatory cortical input is plotted for comparison (dotted lines, power integrated in 5 Hz band centered on cortical frequency in equivalent simulation with cortical inputs). An increased E/I ratio results in maximum phase locking at a higher oscillation frequency, and power of endogenous oscillations follows trend of phase locking strength.

Poisson distributed cortical inputs (compare **Figures 6Ai,ii** to **Figures 2Ai,ii**). Moreover, the frequencies that were amplified by the STN-GPe network corresponded well to the beta-band, i.e., 13–30 Hz (**Figure 2B**). To study the dependence of the resonance peak on the excitation-inhibition balance in the STN, the cortical input strength was then varied while the oscillation frequency remained fixed (**Figure 6D**). The range of synaptic conductances was chosen so that the STN population firing rate traversed the experimentally reported range of 17–37 Hz (Mallet et al., 2008b; Kita and Kita, 2011) in the dopamine depleted state during cortical activation (**Figure 7C**). Maximum phase locking coincided with frequency of maximum endogenous oscillation power observed in the absence of oscillatory inputs (**Figures 2Di,ii**). The results demonstrate how the resonant frequency of the network can be shifted by changing the excitation-inhibition balance, biasing the network toward a slower or faster oscillation. GPe neurons synchronized stronger to the oscillatory input compared to STN neurons (**Figures 6B,C**, **7Bi,ii**), which showed a tendency to burst, mirroring the results for spontaneous synchronization in the autonomous STN-GPe network. Analogous to the autonomous loop, when the slow bursting behavior was reduced by shifting the GPe to STN synaptic current from $GABA_B$ to faster $GABA_A$ receptors, synchronization and phase locking of both STN and GPe neurons was greatly increased.

## Influence of Phase Relationship Between Cortical and Striatal Beta Inputs

Striatal microcircuits exhibit beta-band oscillations in healthy primates (Feingold et al., 2015) and parkinsonian rodent models (McCarthy et al., 2011; Sharott et al., 2017) and have been hypothesized to be part of the pacemaking circuit that generates

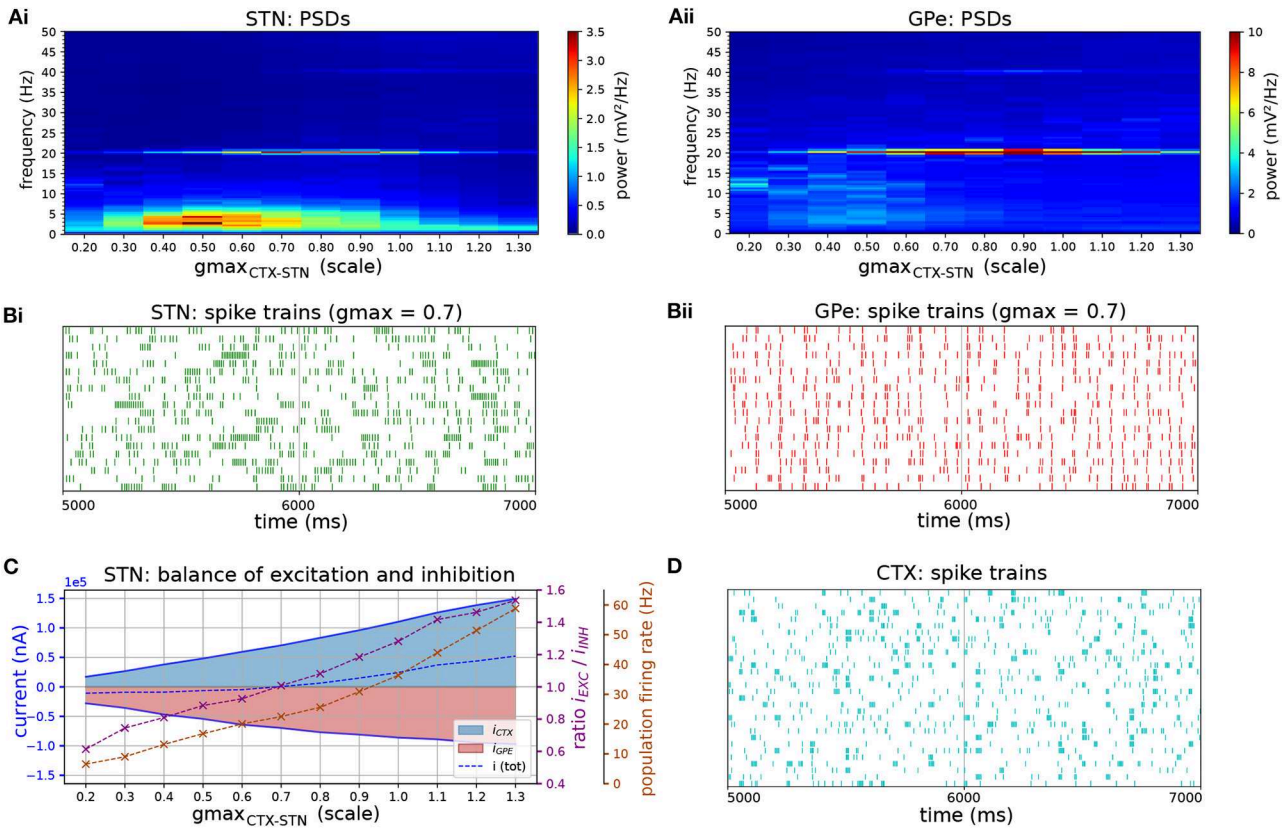


**FIGURE 7 |** Response of STN-GPe network to cortical oscillatory bursting at 20 Hz. **(A)** Mean PSD of somatic membrane voltages in STN **(Ai)** and GPe **(Aii)** as a function of the synaptic conductance (scale factor) of CTX-STN inputs. The peak at 20 Hz reaches a maximum when synapses are at 70% of their baseline strength, whereas the peak in low-frequency power (2-5 Hz) occurs at 50%. **(B)** Representative spike trains of STN (**Bi**, green) and GPe neurons (**Bii**, red) in simulation with synaptic conductances scaled to 70%, corresponding to maximum phase locking and 20 Hz power. **(C)** Balance of excitation and inhibition in the STN **(Ci)** and GPe **(Cii)** based on synaptic currents recorded in three neurons. Population firing rate (brown), E/I ratio (purple), and net synaptic current (blue). **(D)** Cortical oscillatory bursting pattern illustrated using representative spike trains. In each cycle of the oscillation 10% of cells were selected at random to fire a burst in phase with the oscillation, with a variation of 1 ms on the onset and spike timings.

them. In the previous section, the STN-GPe network was shown to generate weak beta-band oscillations in the absence of exogenous beta inputs (**Figures 2**, **3**, **4**), and to phase lock to cortical beta-band inputs which amplified oscillatory activity (**Figure 6**). A potential role of the pallido-striatal loop could be to amplify beta-band oscillations in the STN-GPe network to a more pathological level, as part of a double resonant loop converging on the GPe. A suggested mechanism is that altered striatal activity in PD could shift the phase of firing of the GPe relative to the STN to one that supports STN phase locking through increasing the availability of $Na^+$ and $Ca^{2+}$ channels post-inhibition and pre-excitation (Baufreton et al., 2005; Mallet et al., 2008a, 2012). Alternatively, oscillations that originate in striatal circuits could be transmitted via the striato-pallidal projection and thus introduced into the STN-GPe network (McCarthy et al., 2011; Corbit et al., 2016). Of the two loops converging on GPe neurons, inhibitory striatal afferents would be better suited to interrupt ongoing activity and influence the phase compared to excitatory STN afferents. Hence, the iMSN to GPe projection could play an important role in patterning neural activity in the STN-GPe network.

Phase vector plots in the previous section show that STN and GPe neurons settle into a particular phase relationship where STN leads GPe by 60–90 degrees which contributed to sustaining beta-band oscillations. We hypothesized that inhibitory inputs from the striatum would either disrupt this phase relationship, thereby suppressing beta-band oscillations, or reinforce them depending on where in the phase of the beta oscillation they arrive. To investigate this hypothesis, surrogate striatal spike trains exhibiting beta frequency bursts were generated and the phase with respect to the incoming cortical oscillation was increased in increments of 45 degrees by varying the onset time of bursts. As iMSN-GPe synapses exhibit short-term facilitation, bursts administered through this projection led to an increase in inhibition to the GPe that was greater than the relative increase in spike rate. To compensate for this effect and maintain a physiological firing rate range of the GPe neurons, the peak conductance of iMSN-GPe synapses was reduced by 60%.

Varying the phase of striatal relative to cortical bursts revealed that populations connected by an inhibitory projection, i.e., iMSN, GPe, and STN maintained a rigid phase relationship with respect to the cortical oscillation (**Figure 8**: population vectors in green, red, purple formed a rigid frame that rotated relative to the cyan-colored cortical population vector). The local maximum in phase locking occurred when excitatory CTX and inhibitory GPe afferents to STN fired in anti-phase, occurring when the CTX-iMSN phase difference was set to 225 degrees (**Figures 8B,D,Ei**). This supports the feedback inhibition hypothesis where cortical patterning is promoted when GPe-STN inhibition is offset in phase relative to cortical excitation in PD (Baufreton et al., 2005; Mallet et al., 2008a, 2012). The changing phase relationship of cortical spiking relative to the three other populations also shifted the balance of excitatory and inhibitory currents in the STN (**Figure 8Ai**). Maximum phase locking occurred where the STN was maximally inhibited (E/I ratio $\approx$ 1.1, population firing rate $\approx$ 21 Hz), whereas minimum phase locking coincided with maximum excitation (E/I ratio $\approx$ 1.3, population firing rate $\approx$ 40 Hz). In the GPe this relationship between phase locking strength and firing rate was reversed (**Figure 8Ai**) whereas the relationship with E/I ratio showed no clear trend. The optimal phase relationship of 225 degrees further strengthened phase locking to the applied beta rhythm compared to the situation with only cortical oscillatory inputs. Maximum vector length was increased by a factor of two, confirming increased synchronization, in both populations when compared to the case where only cortical beta frequency inputs were simulated. Maximum power at the oscillation frequency was also increased by a factor of 2.7 in STN and 5.2 in GPe.

## Mechanism of Phase Locking

To further illustrate the interaction between synaptically coupled STN and GPe neurons in the model under conditions of synchronous oscillatory beta-band activity, the mechanism of phase locking of STN cells is presented in **Figure 9**. Pooled cortical spike trains (**Figures 9A,B**, green) illustrate how sparse cortical beta bursts (**Figure 7B**) result in distributed synaptic inputs to individual STN neurons that are not tightly phase locked, but have a combined firing rate that is modulated at the beta frequency. While these exogenous cortical inputs had high spike timing variability, STN and GPe spikes became highly structured and tightly locked to the beta oscillation through the feedback inhibition mechanism. The cortical beta modulation is transmitted to the STN and then to the GPe through their excitatory projections (see phase vectors in **Figure 8Dii**). When the inhibitory feedback arrives back in STN this shuts down spiking (**Figure 9A**) and simultaneously primes the cell for the next period of increased cortical excitation by de-inactivating $Ca^{2+}$ channels (**Figure 9C**) and $Na^{+}$ channels. As the cortical firing rate rises again, synaptic currents (**Figure 9B**) combine with dendritic $Ca^{2+}$ currents to overcome any lingering inhibition and cause the next wave of phase-locked STN spikes. The striatal beta inputs further decreased spiking variability of GPe neurons by narrowing their time window of firing through phasic inhibition (purple phase vector in **Figure 8Dii**).

# DISCUSSION

A new model of the STN-GPe network is presented that incorporates biophysically detailed multi-compartment cell models. The individual STN and GPE cell models capture the interaction of intrinsic and synaptic membrane currents with non-uniform subcellular distributions across the dendritic structure, which can not be captured in single compartment models. The model illustrates how phase locking of STN and GPe neurons, and increased bursting of STN neurons, can arise from the interaction of these currents when their relative strengths and temporal relationships are altered. The STN-GPe model network showed an intrinsic susceptibility to beta-band synchrony that manifest as weak, autonomously-generated endogenous oscillations and selective amplification of exogenous beta-band synaptic inputs at the network's preferred oscillation frequency. The frequency at which endogenous beta oscillatory activity occurred varied with the ratio of excitatory to inhibitory currents to the STN. Varying the phase relationships between external beta-frequency inputs to the network through cortical and striatal pathways further increased or suppressed the level of amplification of cortical beta inputs by modulating the temporal dispersion of action potentials in STN neurons and thereby influencing the precision of phase locking. Varying synaptic strengths within the network affected the balance of excitation and inhibition in both STN and GPe neurons and produced a rich set of behaviors, not only modulating firing rates but also affecting synchronization and bursting properties of neurons. Homeostatic mechanisms mediated by feedback connections and short-term synaptic plasticity dynamics served to stabilize the excitation-inhibition balance in the GPe and reduced the sensitivity of its population firing rate to variations in pre-synaptic rates.

## Oscillatory Properties of the Multi-compartmental STN-GPe Network

In the autonomous STN-GPe network, under conditions of Poisson distributed external synaptic inputs, STN neurons exhibited weak synchronization to the endogenous beta rhythm but retained a weak phase preference with respect to the stronger oscillation in the GPe population (**Figures 2**–**4**). The synchronization strength of STN neurons was found to depend on the relative strength of $GABA_A$ and $GABA_B$ receptors in STN dendrites (**Figure 5**), with an increase in the proportion of fast-acting $GABA_A$ receptors resulting in an increase in the strength of oscillation. The endogenous oscillation frequency of the STN-GPe network was further influenced by the balance of excitatory and inhibitory currents in the STN. This balance affected the net level of excitatory drive in the network, shifting the oscillation frequency toward the higher beta range for increased levels of excitatory drive (**Figures 2A**, **5D**). Besides affecting population firing rates and the frequency of synchronous oscillations, the excitation-inhibition balance also strongly influenced the firing pattern of STN neurons: for a low ratio of excitation to inhibition and sufficiently strong inhibitory currents, STN neurons transitioned to a firing mode characterized by low-frequency tight bursts (high intra-burst

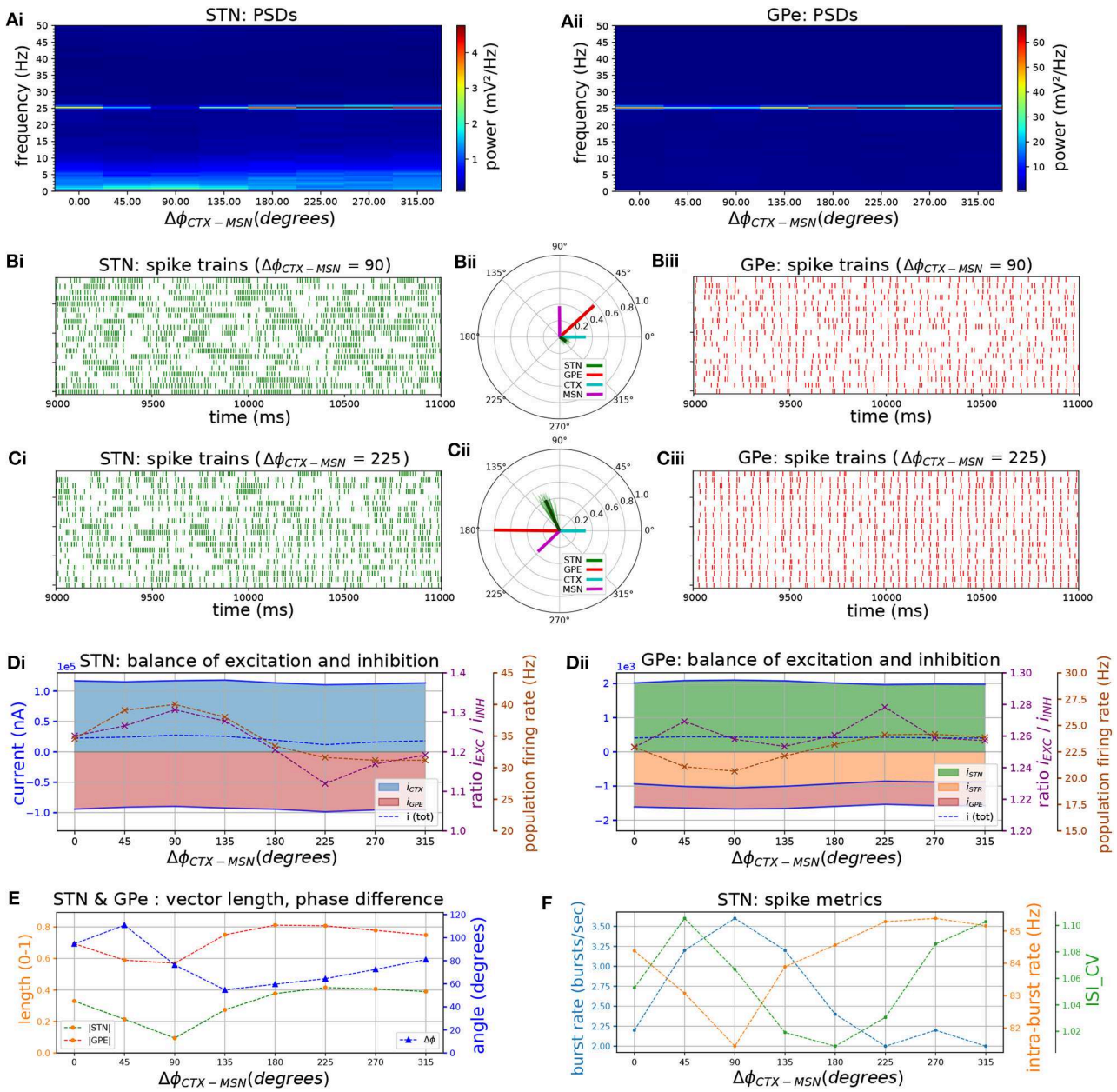


**FIGURE 8 |** The phase relationship between cortical and striatal beta-band inputs to the STN-GPe network affects the strength of phase-locking by setting the relative timing of excitatory and inhibitory STN afferents. Response of the STN-GPe network to oscillatory bursting inputs applied via both cortico-subthalamic (CTX-STN) and striato-pallidal (iMSN-GPe) afferents. The phase difference between cortical and striatal oscillatory bursts was increased in steps of 45°. All phase vectors were measured with respect to the instantaneous phase of the cortical oscillation. **(A)** Mean PSD of the somatic membrane voltages of STN **(Ai)** and GPe **(Aii)** neurons, showing weakening and strengthening of oscillations as relative phases of inputs are rotated. **(B,C)** Representative spike trains and phase vectors of STN (column i, green) and GPe population (column iii, red) for CTX-iMSN phase difference of 90° **(C)** and 225° **(D)**. Column ii shows phase vectors of the STN, GPe, CTX, iMSN populations (in green; red; blue; purple, respectively; mean population vectors plotted as thick solid lines and cell vectors as thin transparent lines). **(D)** Balance of excitation and inhibition in the STN **(Di)** and GPe **(Dii)** based on synaptic currents recorded in three neurons. Population firing rate (brown), E/I ratio (purple), and net synaptic current (blue). Shaded areas represent estimated total synaptic current from one pre-synaptic population during a simulation. **(E)** Population vector length and angle of STN (green) and GPe (red) population. **(F)** Metrics that characterize bursting in STN neurons: median burst rate, intra-burst firing rate, and coefficient of variation of ISIs across all STN cells.

firing rate, **Figures 2–4**). Low-frequency bursting was periodic at 2–5 Hz but was not synchronized between cells. This shift in firing pattern toward sparse, tight bursting is in correspondence with changes in burst-related measures such as intra-burst firing rate and sub-beta band power that are most predictive of akinetic-bradykinetic symptoms in humans (Sharott et al.,

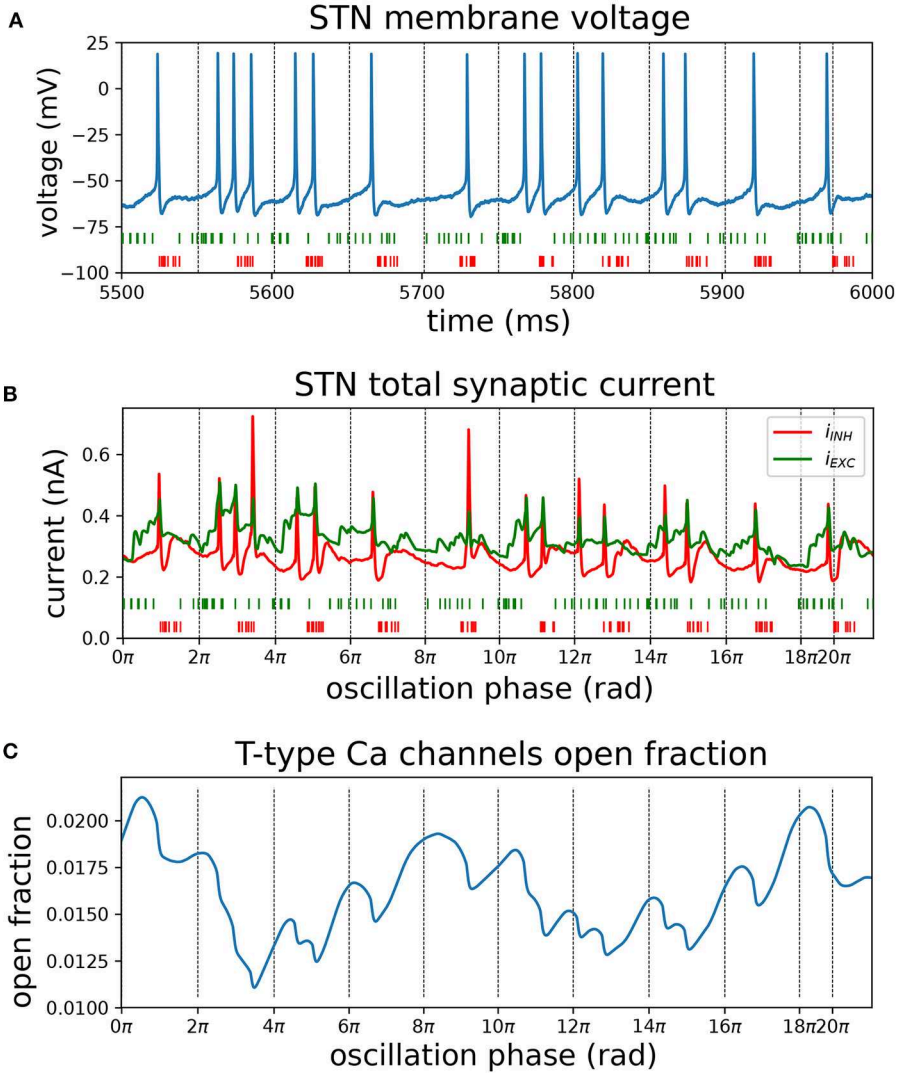


**FIGURE 9 |** Mechanisms contributing to phase locking of STN cells to cortical beta oscillations. Recordings of synaptic currents and T-type calcium (CaT) channel inactivation from an identified phase-locked STN cell during a simulation with high phase locking (analogous to **Figure 8D**, cortical and striatal beta bursts at 20 Hz with phase difference of 225 degrees). Inactivation variables were recorded from each compartment with CaT ion channels and averaged over all compartments in the cell. Zero-crossings of the instantaneous beta phase are indicated using vertical dotted lines. **(A)** Somatic membrane voltage during phase-locked interval (blue). Spike trains from excitatory (green) and inhibitory (red) afferents to the cell were pooled. **(B)** Total excitatory and inhibitory synaptic current (in green; red, respectively) and pooled spike trains underneath. **(C)** Mean CaT channel inactivation across the cell's dendritic tree. High values correspond to de-inactivation. Transient de-inactivation approximately one half period after an inhibitory barrage engages depolarizing T-type $Ca^{2+}$ current and contributes to phase-locked spiking.

2014) and monkeys (Sanders et al., 2013). The firing rate and pattern of GPe neurons was less sensitive than that of STN neurons to variations in its excitatory or inhibitory drive due to the contribution of negative feedback control by homeostatic mechanisms that operated in synergy to stabilize its E/I ratio. However, GPe neurons did synchronize more strongly under conditions of low excitatory drive from the STN enabling them to act more autonomously and synchronize through inhibitory collaterals within the GPe network.

When beta-band spiking inputs were applied to the STN-GPe network via cortico-STN afferents, the STN-GPe network phase locked to the beta rhythm. Frequencies near the autonomous oscillation frequency for a given E/I ratio were preferentially amplified, reflected in increased phase locking and power of the somatic membrane voltage at that frequency (**Figure 6**). This is supportive of experimental observations that oscillatory activity in STN is contingent on cortical oscillations (Magill et al., 2001), likely transmitted though the hyperdirect pathway (Tachibana et al., 2011). Phase locking and beta frequency power were further strengthened by the addition of striatal oscillatory inputs with a particular phase relationship to cortical oscillatory inputs (**Figure 8**). Maximum phase-locking occurred when GPe spiking was aligned in anti-phase with cortical inputs to the STN (**Figures 8C,E**). When excitation and inhibition occurred

in anti-phase, inhibition was likely more effective at transiently hyperpolarizing the membranes of STN neurons, suggested by the local minimum in their E/I ratio (**Figure 8Di**). Strong hyperpolarization can evoke low-latency, temporally precise responses to an excitatory stimulus by de-inactivating $Ca^{2+}$ and $Na^{+}$ channels, and thereby priming them to respond to excitatory cortical inputs (Bevan et al., 2007). This mechanism may be responsible for the increase in phase locking under this phase relationship. In contrast, phase alignment of cortical and GPe neurons, corresponding to coincident firing, desynchronized STN neurons (**Figure 8C**). These findings are in agreement with recent experimental observations which demonstrate that co-stimulation of GABAergic and glutamergic STN afferents disperses STN spiking and has a desynchronizing effect on the population (Amadeus Steiner et al., 2019). Overall, the simulation results are consistent with the hypothesis of cortical patterning and resonance of beta activity within the STN-GPe network through feedback inhibition, whereby GPe inhibition arriving in anti-phase to cortical excitation promotes phase locking of STN neurons to beta-band cortical inputs (Baufreton et al., 2005).

## Relation of Mechanism of Oscillations to Other Models of Oscillatory Activity in the STN-GPe Network

The mechanism by which oscillatory neural activity can be generated in the STN-GPe network, by alternating phases of excitation and inhibition in a delayed negative feedback loop, has been described in previous models (Terman et al., 2002; Holgado et al., 2010; Kumar et al., 2011). The mechanism of oscillation in the model presented here is consistent with this, and the model additionally illustrates the dual role of precisely timed GPe inhibition in transiently reducing STN neuron excitability and hyperpolarizing them such that they are primed to respond with bursting to excitatory cortical inputs (**Figure 9**). Furthermore, it highlights the sensitivity of the network oscillation to the excitation-inhibition balance in each population and synaptic current properties.

In the multicompartment model, endogenously generated beta frequency oscillations were generated within the STN-GPe network when the strength of short duration $GABA_A$-mediated currents was increased. Since the slow timescale, signaling cascade-mediated $GABA_B$ currents are typically not modeled, this result can be easily reconciled with results from single-compartment and firing rate models where high gain within the closed-loop is a necessary condition for strong endogenously-generated oscillations in the STN-GPe network (Holgado et al., 2010; Park et al., 2011; Pavlides et al., 2012; Wei et al., 2015). The strength of the endogenous oscillations in our model was relatively weak, except when inhibitory GPe-STN currents were strongly dominated by fast-acting $GABA_A$-mediated currents and $GABA_B$-mediated slow currents were weak. The oscillation frequency of the network could be modulated by varying the ratio of excitation to inhibition in STN and GPe, and increased as this ratio increased (**Figure 6**).

The oscillation frequency of the network has been shown to be sensitive to model parameters in previous computational models of the BGTC network. Specifically, in mean field models of the STN-GPe loop the oscillation frequency showed a strong sensitivity to transmission delays and neuronal membrane time constants (Holgado et al., 2010; Liénard et al., 2017), and a weaker sensitivity to coupling strengths (Holgado et al., 2010; Pavlides et al., 2015; Liu et al., 2017), also demonstrated in a spiking model (Wei et al., 2015). In the multicompartment model presented here, where active ion channels on the dendrites contribute to synaptic integration, synaptic strength and effective membrane time constant are interdependent since the membrane charging speed is affected by transient activation of ion channels as a response to synaptic inputs. In biological neurons the balance of excitation and inhibition is tightly regulated through multiple adaptive processes (Turrigiano, 2011), and likely maintains the range of possible oscillation frequencies within a narrow range.

Other than the condition where GPe-STN currents were dominated by fast-acting $GABA_A$ currents, strongly synchronized beta-band oscillations appeared only when exogenous beta-band inputs were introduced to the network (**Figures 6**, **8**). These results, therefore, support a role for resonance with oscillations throughout other basal ganglia loops in the generation of increased STN-GPe beta activity in Parkinson's disease. Such an oscillatory drive can be provided either by an extrinsic oscillator, assumed to originate within the cortex in the present model, or by reverberation of oscillations in connected feedback loops such as the pallido-striatal loop (Corbit et al., 2016), intra-striatal loops (McCarthy et al., 2011), or the larger thalamocortical loop (Dovzhenok and Rubchinsky, 2012; Kang and Lowery, 2013; Pavlides et al., 2015; Reis et al., 2019). The model exhibited clear resonance in response to excitatory synaptic inputs to the STN within the beta frequency range (**Figure 6**). The frequency at which the maximum resonance occurred increased with increasing ratio of excitation to inhibition, similar to the increase in frequency observed in the case of endogenously generated oscillations. Resonance phenomena in the beta-band have previously been reported in computational models of basal ganglia networks, consistent with our modeling results: Pavlides et al. (2015) fitted mean field rate models to experimental data from non-human primates and found that the models that best explained the data relied on a strong cortical oscillation to sustain beta-band oscillations ($\sim$ 15 Hz) in the network. In a comparable mean-field model, Liu et al. (2017) found that upper beta-band (21–35 Hz) oscillations in the STN-GPe loop originated from cortical oscillatory inputs and supported a lower beta-band (12–20 Hz) oscillation that was endogenously generated. Ahn et al. (2016) using 10 single compartment STN and GPe neurons observed multiple resonances in the beta-band when varying the strength of striato-pallidal and pallida-subthalamic inhibition, with resonant peaks occurring consistently between 18 and 21 Hz. Similarly, Fountas and Shanahan (2017) found that STN neurons in their model exhibited high spontaneous beta-band power (18–30 Hz) and synchronized selectively with cortical input in this frequency range.

## Model Complexity and Limitations

One of the main advantages of the biophysically detailed model presented here is that the model can capture the non-uniform distribution of afferent inputs from different pre-synaptic

populations across the dendritic tree (**Tables 3**, **5**). This targeting of specific regions of the dendrites by different populations can lead to variations in synaptic integration properties within the structure. This feature is potentially of particular importance in the generation of pathological oscillations given that neuronal phase response curves, used to quantify the tendency of neurons to synchronize to their inputs, differ when stimuli are applied to different subcellular regions in STN and GPe neurons (Schultheiss et al., 2010; Farries and Wilson, 2012). Hence, a model that incorporates a full complement of ion channel and the synapse groups that interact with them may be expected to yield a more realistic representation of how synchronization arises in the network. In future studies, this could also contribute to a better understanding of neuronal currents contributing to the local field potential in synchronized and asynchronous states, as synaptic and ionic transmembrane currents combine to form the extracellular currents that underpin this signal (Buzsáki et al., 2012).

A second advantage of such detailed multicompartment models is that parameters have a clear relationship to the underlying biophysical system and are more meaningful in terms of physiological processes compared to models where parameters are lumped, as in single-compartment conductance-based models, or abstracted as in mean-field or generalized integrate-and-fire models. This allows for a more direct translation of experimental findings to parameter variations in the model. On the other hand, detailed cell models are more sensitive to correct estimation of these parameters which is limited by measurements performed for the purpose of model fitting as well as the fitting procedures themselves. Biophysically detailed models offer new ways to study factors contributing to the development of synchrony. Such models provide a means to investigate the relative contributions of physiological mechanisms to the development of synchrony while controlling other factors in a manner that is not possible *in vivo*. Though the model presented incorporates a higher level of physiological detail than previous models of the STN-GPe network, several simplifications were necessary due to the model complexity, which should be considered.

Downregulation of HCN channel currents with dopamine depletion was modeled as a decrease in its peak conductance. However, dopamine is known to interact with several more ion channels that are involved in linearizing the current-firing rate curve and regularizing autonomous pacemaking of STN neurons (Loucif et al., 2008; Ramanathan et al., 2008; Yang et al., 2016) which are not included in the STN cell model used here (Gillies and Willshaw, 2005). Recent evidence suggests that the loss of autonomous spiking is a necessary condition for the exaggerated cortical patterning of STN related to motor dysfunction (McIver et al., 2018). Better characterization of the ion channels involved in pacemaking and their response to dopamine depletion will enable the systematic exploration of their contribution to STN response properties and pathological firing patterns.

In our network model the main sources of firing rate variability were randomness in the input spiking patterns, the presence of surrogate Poisson spike sources in STN and GPe, membrane noise, and randomness in connection patterns and the position of synapses. However these factors do not capture the full biological variability in morpho-electric cell types, synaptic strength distributions, and resulting firing patterns in each population. In the GPe, two distinct populations have been identified based on their molecular profile and axonal connectivity (Mallet et al., 2012). Only the prototypic sub-population projecting mainly to STN and preferentially firing in anti-phase to it was modeled here, with the arkypallidal sub-populations projecting back to striatum omitted. Moreover, the GPe cell model used was only one representative candidate out of a large set of models with varying ion channel expression and morphology that matched a corresponding database of electrophysiological recordings (Gunay et al., 2008). Similarly, the STN model represents a stereotypic characterization rather than a reconstruction of a specific STN cell and does not capture variability in firing properties and receptor expression. In particular, STN neurons *in vivo* are known to have variable expression of $GABA_B$ receptors (Galvan et al., 2004) which cause strong hyperpolarization responses and longer pauses in some but not all STN neurons (Hallworth and Bevan, 2005) and a strong rebound burst response (Galvan et al., 2004) in a subset of these. A model that accounts for the biological variability in $GABA_B$ expression and that of channels underlying the rebound response may reveal a wider range of responses to increased inhibition among STN neurons. In such a model, beta rhythms could be transmitted to a subset of STN neurons whereas others would show longer pauses with stronger rebound bursts. Moreover, the $GABA_B$ synapse model used does not fully account for activation of extrasynaptic $GABA_BR$ due to GABA spillover (Galvan et al., 2004) which is mediated by tonic high-frequency *and* coincident firing of afferents (Bevan et al., 2006). A model where multiple GABAergic synapses act on a shared pool of extrasynaptic $GABA_BR$ might increase the importance of synchronized pre-synaptic activity in switching STN neurons to a burst-firing mode.

The effect of the correlation between cortical and striatal inputs to the network was explored by varying the relative phases of both populations when firing in a synchronous oscillatory pattern (**Figure 8**). Uncorrelated firing between both populations was also explored (**Figures 2**–**7**). In reality, beta activity in both populations is likely to be correlated as the striatum receives topographic inputs from the same cortical areas projecting to the STN. Such correlation could lead to transient synchronization effects not explored here, that could promote or counteract additional oscillatory synchronization depending on the exact phase relationships. The effect of varying connectivity patterns between neuronal populations was not directly explored here. The development of neural synchronization and oscillatory activity are known to be dependent on network topology (Zhao et al., 2011), and this effect has previously been studied in a single compartment model of the STN-GPe network (Terman et al., 2002). The network topology used in the present study is closest to the random, sparsely-connected topology in Terman et al. (2002) which was shown to develop synchronized bursting patterns at lower frequencies. Choosing different randomly-generated connection matrices did not qualitatively change our results, however altering the connection topology would likely

lead to different synchronization properties. Moreover, it is known that connection patterns within the basal ganglia are altered with dopamine depletion, particularly within the striatum (Cho et al., 2002), leading to a loss of input specificity in neuronal responses (Bronfeld and Bar-Gad, 2011). These alterations in connection patterns and resulting effects on spike correlations were not taken into account as we did not consider cortico-striatal connectivity in our model. As arkypallidal GPe neurons were not modeled, the pallido-striatal feedback loop was not captured. This additional feedback loop has also been suggested as a candidate pacemaker circuit for beta-band oscillations (Corbit et al., 2016), however, blocking of striatal inputs was not found to reduce the power of beta oscillations in rat GPe (Tachibana et al., 2011).

Finally, while there is consistent evidence of increased beta-band oscillatory activity in Parkinsons disease (Sharott et al., 2005; Mallet et al., 2008b) and a reduction of pathological beta band activity with interventions that improve symptoms in patients and animal models of the disease (Kühn et al., 2006; Weinberger et al., 2006; Ray et al., 2008; Eusebio et al., 2011), strong evidence in support of a causal role for pathological beta activity in the symptoms of Parkinsons disease has yet to be established. Indeed, recent studies failed to find evidence of any causal link between artificially induced beta band activity and motor impairment in parkinsonian rats (Swan et al., 2019), nor between the reduction of beta band activity and alleviation of motor symptoms (Pan et al., 2016). A lack of causality, however, may not necessarily be incompatible with the use of beta-band oscillations as a clinical biomarker, particularly for akinetic-bradykinetic forms of Parkinson's Disease at advanced stages of disease progression. Initial trials of adaptive or closed-loop DBS strategies targeted at suppression of beta-band activity have been successful in demonstrating simultaneous reductions in patient symptoms (Little et al., 2013; Velisar et al., 2019). Beta-band power may thus still be a suitable biomarker to indirectly gauge underlying physiological changes that are more directly related to network dysfunction such as alterations in synaptic strengths and functional connectivity within the network.

Sharott et al. (2005), Mallet et al. (2008b), and Kuhn et al. (2008), and are reduced by DBS and pharmacological interventions that alleviate parkinsonian motor symptoms (Kühn et al., 2006; Weinberger et al., 2006; Ray et al., 2008; Eusebio et al., 2011).

### Conclusion

In summary, a biophysically detailed model of the parkinsonian STN-GPe network is presented which captures non-uniform distribution of ion channels and synapses in neuronal dendrites. The network model exhibited an intrinsic susceptibility to synchronous neural oscillations within the frequency range of pathological beta-band activity observed in Parkinson's disease. Oscillations in the autonomous STN-GPe network, however, were too weak to support a pacemaker role as the sole origin of beta-band oscillations in the wider BGTC network in Parkinson's disease. In particular in the STN, autonomous beta-band oscillations and phase locking of individual cells were weak unless slower $GABA_B$-mediated currents were substantially reduced. Beta-band oscillations were considerably amplified by a relatively sparse cortical beta input, with clear resonance occurring within the beta frequency range. The frequency at which the resonant peak occurred increased with increasing ratio of excitatory to inhibitory STN inputs. beta-band oscillations were further amplified by striatal beta inputs that promoted anti-phase firing of cortex and GPe. These results support the cortical patterning and network resonance hypothesis for the generation of pathological beta-band oscillatory activity in Parkinson's disease in a multi-compartment model of the STN-GPe network. They also illustrate the potential of the pallido-striatal feedback loop in further amplifying beta oscillations within the network.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

## AUTHOR CONTRIBUTIONS

All experiments were performed in the Neuromuscular Systems Laboratory in University College Dublin, Ireland. LK and ML: conceived and designed the experiments, interpreted results of experiments, prepared the figures, edited and revised the manuscript, and approved the final version of manuscript. LK: performed experiments and analyzed data and drafted the manuscript.

## FUNDING

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant ERC-2014-CoG-646923-DBSModel).

## ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at BioRxiv 611103 (Koelman and Lowery, 2019).

## REFERENCES

Abdi, A., Mallet, N., Mohamed, F. Y., Sharott, A., Dodson, P. D., Nakamura, K. C., et al. (2015). Prototypic and arkypallidal neurons in the dopamine-intact external globus pallidus. *J. Neurosci.* 35, 6667–6688. doi: 10.1523/JNEUROSCI.4662-14.2015

Ahn, S., Zauber, S. E., Worth, R. M., and Rubchinsky, L. L. (2016). Synchronized beta-band oscillations in a model of the globus pallidus-subthalamic nucleus network under external input. *Front. Comput. Neurosci.* 10:134. doi: 10.3389/fncom.2016.00134

Amadeus Steiner, L., Barreda Tomás, F. J., Planert, H., Alle, H., Vida, I., and Geiger, J. R. P. (2019). Connectivity and dynamics underlying

synaptic control of the subthalamic Nucleus. *J. Neurosci.* 39, 2470–2481. doi: 10.1523/JNEUROSCI.1642-18.2019

Atherton, J. F., Menard, A., Urbain, N., and Bevan, M. D. (2013). Short-term depression of external globus pallidus-subthalamic nucleus synaptic transmission and implications for patterning subthalamic activity. *J. Neurosci.* 33, 7130–7144. doi: 10.1523/JNEUROSCI.3576-12.2013

Baufreton, J., Atherton, J. F., Surmeier, D. J., and Bevan, M. D. (2005). Enhancement of excitatory synaptic integration by GABAergic inhibition in the subthalamic nucleus. *J. Neurosci.* 25, 8505–8517. doi: 10.1523/JNEUROSCI.1163-05.2005

Baufreton, J., and Bevan, M. D. (2008). D2-like dopamine receptor-mediated modulation of activity-dependent plasticity at GABAergic synapses in the subthalamic nucleus: dopaminergic modulation of synaptic plasticity in the subthalamus. *J. Physiol.* 586, 2121–2142. doi: 10.1113/jphysiol.2008.151118

Baufreton, J., Kirkham, E., Atherton, J. F., Menard, A., Magill, P. J., Bolam, J. P., et al. (2009). Sparse but selective and potent synaptic transmission from the globus pallidus to the subthalamic nucleus. *J. Neurophysiol.* 102, 532–545. doi: 10.1152/jn.00305.2009

Benabid, A. L., Chabardes, S., Mitrofanis, J., and Pollak, P. (2009). Deep brain stimulation of the subthalamic nucleus for the treatment of Parkinson's disease. *Lancet Neurol.* 8, 67–81. doi: 10.1016/S1474-4422(08)70291-6

Bevan, M. D., Atherton, J. F., and Baufreton, J. (2006). Cellular principles underlying normal and pathological activity in the subthalamic nucleus. *Curr. Opin. Neurobiol.* 16, 621–628. doi: 10.1016/j.conb.2006.10.003

Bevan, M. D., Francis, C. M., and Bolam, J. P. (1995). The glutamate-enriched cortical and thalamic input to neurons in the subthalamic nucleus of the rat: convergence with GABA-positive terminals. *J. Comp. Neurol.* 361, 491–511. doi: 10.1002/cne.903610312

Bevan, M. D., Hallworth, N. E., and Baufreton, J. (2007). "GABAergic control of the subthalamic nucleus," in *Progress in Brain Research*, Vol. 160, eds J. M. Tepper, E. D. Abercrombie, and J. P. Bolam (Amsterdam: Elsevier), 173–188. doi: 10.1016/S0079-6123(06)60010-1

Bronfeld, M., and Bar-Gad, I. (2011). Loss of specificity in basal ganglia related movement disorders. *Front. Syst. Neurosci.* 5:38. doi: 10.3389/fnsys.2011.00038

Bronte-Stewart, H., Barberini, C., Koop, M. M., Hill, B. C., Henderson, J. M., and Wingeier, B. (2009). The STN beta-band profile in Parkinson's disease is stationary and shows prolonged attenuation after deep brain stimulation. *Exp. Neurol.* 215, 20–28. doi: 10.1016/j.expneurol.2008.09.008

Buzsáki, G., Anastassiou, C. A., and Koch, C. (2012). The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* 13, 407–420. doi: 10.1038/nrn3241

Chan, C. S. (2004). HCN2 and HCN1 channels govern the regularity of autonomous pacemaking and synaptic resetting in globus pallidus neurons. *J. Neurosci.* 24, 9921–9932. doi: 10.1523/JNEUROSCI.2162-04.2004

Chan, C. S., Glajch, K. E., Gertler, T. S., Guzman, J. N., Mercer, J. N., Lewis, A. S., et al. (2011). HCN channelopathy in external globus pallidus neurons in models of Parkinson's disease. *Nat. Neurosci.* 14, 85–92. doi: 10.1038/nn.2692

Cho, J., Duke, D., Manzino, L., Sonsalla, P. K., and West, M. O. (2002). Dopamine depletion causes fragmented clustering of neurons in the sensorimotor striatum: evidence of lasting reorganization of corticostriatal input. *J. Comp. Neurol.* 452, 24–37. doi: 10.1002/cne.10349

Chu, H.-Y., Atherton, J. F., Wokosin, D., Surmeier, D. J., and Bevan, M. D. (2015). Heterosynaptic regulation of external globus pallidus inputs to the subthalamic nucleus by the motor cortex. *Neuron* 85, 364–376. doi: 10.1016/j.neuron.2014.12.022

Chu, H.-Y., McIver, E. L., Kovaleski, R. F., Atherton, J. F., and Bevan, M. D. (2017). Loss of hyperdirect pathway cortico-subthalamic inputs following degeneration of midbrain dopamine neurons. *Neuron* 95, 1306.e5–1318.e5. doi: 10.1016/j.neuron.2017.08.038

Cooper, A., and Stanford, I. (2001). Dopamine D2 receptor mediated presynaptic inhibition of striatopallidal GABAA IPSCs *in vitro*. *Neuropharmacology* 41, 62–71. doi: 10.1016/S0028-3908(01)00038-7

Corbit, V. L., Whalen, T. C., Zitelli, K. T., Crilly, S. Y., Rubin, J. E., and Gittis, A. H. (2016). Pallidostriatal projections promote oscillations in a dopamine-depleted biophysical network model. *J. Neurosci.* 36, 5556–5571. doi: 10.1523/JNEUROSCI.0339-16.2016

Cragg, S. J., Baufreton, J., Xue, Y., Bolam, J. P., and Bevan, M. D. (2004). Synaptic release of dopamine in the subthalamic nucleus. *Eur. J. Neurosci.* 20, 1788–1802. doi: 10.1111/j.1460-9568.2004.03629.x

Destexhe, A., and Sejnowski, T. J. (1995). G protein activation kinetics and spillover of gamma-aminobutyric acid may account for differences between inhibitory responses in the hippocampus and thalamus. *Proc. Natl. Acad. Sci. U.S.A.* 92, 9515–9519. doi: 10.1073/pnas.92.21.9515

Dovzhenok, A., and Rubchinsky, L. L. (2012). On the origin of tremor in Parkinson's disease. *PLoS ONE* 7:e41598. doi: 10.1371/journal.pone.0041598

Drouot, X., Oshino, S., Jarraya, B., Besret, L., Kishima, H., Remy, P., et al. (2004). Functional recovery in a primate model of Parkinson's disease following motor cortex stimulation. *Neuron* 44, 769–778. doi: 10.1016/j.neuron.2004.11.023

Eusebio, A., Thevathasan, W., Doyle Gaynor, L., Pogosyan, A., Bye, E., Foltynie, T., et al. (2011). Deep brain stimulation can suppress pathological synchronisation in parkinsonian patients. *J. Neurol. Neurosurg. Psychiatry* 82, 569–573. doi: 10.1136/jnnp.2010.217489

Fan, K. Y., Baufreton, J., Surmeier, D. J., Chan, C. S., and Bevan, M. D. (2012). Proliferation of external globus pallidus-subthalamic nucleus synapses following degeneration of midbrain dopamine neurons. *J. Neurosci.* 32, 13718–13728. doi: 10.1523/JNEUROSCI.5750-11.2012

Farries, M. A., and Wilson, C. J. (2012). Phase response curves of subthalamic neurons measured with synaptic input and current injection. *J. Neurophysiol.* 108, 1822–1837. doi: 10.1152/jn.00053.2012

Feingold, J., Gibson, D. J., DePasquale, B., and Graybiel, A. M. (2015). Bursts of beta oscillation differentiate postperformance activity in the striatum and motor cortex of monkeys performing movement tasks. *Proc. Natl. Acad. Sci. U.S.A.* 112, 13687–13692. doi: 10.1073/pnas.1517629112

Fieblinger, T., Graves, S. M., Sebel, L. E., Alcacer, C., Plotkin, J. L., Gertler, T. S., et al. (2014). Cell type-specific plasticity of striatal projection neurons in Parkinsonism and L-DOPA-induced dyskinesia. *Nat. Commun.* 5:5316. doi: 10.1038/ncomms6316

Flint, A. C., Maisch, U. S., Weishaupt, J. H., Kriegstein, A. R., and Monyer, H. (1997). NR2A subunit expression shortens NMDA receptor synaptic currents in developing neocortex. *J. Neurosci.* 17, 2469–2476. doi: 10.1523/JNEUROSCI.17-07-02469.1997

Fountas, Z., and Shanahan, M. (2017). The role of cortical oscillations in a spiking neural network model of the basal ganglia. *PLoS ONE* 12:e0189109. doi: 10.1371/journal.pone.0189109

Froux, L., Le Bon-Jego, M., Miguelez, C., Normand, E., Morin, S., Fioramonti, S., et al. (2018). D5 dopamine receptors control glutamatergic AMPA transmission between the motor cortex and subthalamic nucleus. *Sci. Rep.* 8:8858. doi: 10.1038/s41598-018-27195-6

Fujimoto, K., and Kita, H. (1993). Response characteristics of subthalamic neurons to the stimulation of the sensorimotor cortex in the rat. *Brain Res.* 609, 185–192. doi: 10.1016/0006-8993(93)90872-K

Galvan, A., Charara, A., Pare, J.-F., Levey, A., and Smith, Y. (2004). Differential subcellular and subsynaptic distribution of GABAA and GABAB receptors in the monkey subthalamic nucleus. *Neuroscience* 127, 709–721. doi: 10.1016/j.neuroscience.2004.05.014

Gillies, A., and Willshaw, D. (2005). Membrane channel interactions underlying rat subthalamic projection neuron rhythmic and bursting activity. *J. Neurophysiol.* 95, 2352–2365. doi: 10.1152/jn.00525.2005

Gillies, A., and Willshaw, D. (2007). Neuroinformatics and modeling of the basal ganglia: bridging pharmacology and physiology. *Expert Rev. Med. Devices* 4, 663–672. doi: 10.1586/17434440.4.5.663

Gillies, A., Willshaw, D., and Li, Z. (2002). Subthalamic-pallidal interactions are critical in determining normal and abnormal functioning of the basal ganglia. *Proc. R. Soc. B Biol. Sci.* 269, 545–551. doi: 10.1098/rspb.2001.1817

Gradinaru, V., Mogri, M., Thompson, K. R., Henderson, J. M., and Deisseroth, K. (2009). Optical deconstruction of parkinsonian neural circuitry. *Science* 324, 354–359. doi: 10.1126/science.1167093

Gunay, C., Edgerton, J. R., and Jaeger, D. (2008). Channel density distributions explain spiking variability in the globus pallidus: a combined physiology and computer simulation database approach. *J. Neurosci.* 28, 7476–7491. doi: 10.1523/JNEUROSCI.4198-07.2008

Hallworth, N. E., and Bevan, M. D. (2005). Globus pallidus neurons dynamically regulate the activity pattern of subthalamic nucleus neurons through the frequency-dependent activation of postsynaptic GABAA and GABAB receptors. *J. Neurosci.* 25, 6304–6315. doi: 10.1523/JNEUROSCI.0450-05.2005

Hanson, J. E., and Jaeger, D. (2002). Short-term plasticity shapes the response to simulated normal and Parkinsonian input patterns in the globus

pallidus. *J. Neurosci.* 22, 5164–5172. doi: 10.1523/JNEUROSCI.22-12-05164.2002

Hernández, A., Ibáñez-Sandoval, O., Sierra, A., Valdiosera, R., Tapia, D., Anaya, V., et al. (2006). Control of the subthalamic innervation of the rat globus pallidus by $D_{2/3}$ and $D_4$ dopamine receptors. *J. Neurophysiol.* 96, 2877–2888. doi: 10.1152/jn.00664.2006

Hines, M. L., and Carnevale, N. T. (1997). The NEURON simulation environment. *Neural Comput.* 9, 1179–1209. doi: 10.1162/neco.1997.9.6.1179

Holgado, A. J. N., Terry, J. R., and Bogacz, R. (2010). Conditions for the generation of beta oscillations in the subthalamic nucleus-globus pallidus network. *J. Neurosci.* 30, 12340–12352. doi: 10.1523/JNEUROSCI.0817-10.2010

Jahr, C., and Stevens, C. (1990). Voltage dependence of NMDA-activated macroscopic conductances predicted by single-channel kinetics. *J. Neurosci.* 10(9):3178–3182. doi: 10.1523/JNEUROSCI.10-09-03178.1990

Johnson, P. I., and Napier, T. C. (1997). GABA- and glutamate-evoked responses in the rat ventral pallidum are modulated by dopamine. *Eur. J. Neurosci.* 9, 1397–1406. doi: 10.1111/j.1460-9568.1997.tb01494.x

Jones, E., Oliphant, T., and Peterson, P. (2001). *SciPy: Open Source Scientific Tools for Python*.

Kang, G., and Lowery, M. M. (2013). Interaction of oscillations, and their suppression via deep brain stimulation, in a model of the cortico-basal ganglia network. *IEEE Trans. Neural Syst. Rehab. Eng.* 21, 244–253. doi: 10.1109/TNSRE.2013.2241791

Kita, H. (2007). "Globus pallidus external segment," in *Progress in Brain Research*, Vol. 160, eds J. M. Tepper, E. D. Abercrombie, and J. P. Bolam (Amsterdam: Elsevier), 111–133. doi: 10.1016/S0079-6123(06)60007-1

Kita, H., and Jaeger, D. (2016). "Organization of the globus pallidus," in *Handbook of Behavioral Neuroscience*, Vol. 24, eds H. Steiner and K. Y. Tseng (Amsterdam: Elsevier), 259–276. doi: 10.1016/B978-0-12-802206-1.00013-1

Kita, H., and Kita, T. (2011). Cortical stimulation evokes abnormal responses in the dopamine-depleted rat basal ganglia. *J. Neurosci.* 31, 10311–10322. doi: 10.1523/JNEUROSCI.0915-11.2011

Kita, H., and Kitai, S. (1991). Intracellular study of rat globus pallidus neurons: Membrane properties and responses to neostriatal, subthalamic and nigral stimulation. *Brain Res.* 564, 296–305. doi: 10.1016/0006-8993(91)91466-E

Koelman, L. A., and Lowery, M. M. (2019). Autonomous oscillations and phase-locking in a biophysically detailed model of the STN-GPe network. *bioRxiv*. doi: 10.1101/611103

Kuhn, A. A., Kempf, F., Brucke, C., Gaynor Doyle, L., Martinez-Torres, I., Pogosyan, A., et al. (2008). High-frequency stimulation of the subthalamic nucleus suppresses oscillatory activity in patients with Parkinson's disease in parallel with improvement in motor performance. *J. Neurosci.* 28, 6165–6173. doi: 10.1523/JNEUROSCI.0282-08.2008

Kühn, A. A., Kupsch, A., Schneider, G.-H., and Brown, P. (2006). Reduction in subthalamic 8-35 Hz oscillatory activity correlates with clinical improvement in Parkinson's disease: STN activity and motor improvement. *Eur. J. Neurosci.* 23, 1956–1960. doi: 10.1111/j.1460-9568.2006.04717.x

Kumar, A., Cardanobile, S., Rotter, S., and Aertsen, A. (2011). The role of inhibition in generating and controlling Parkinson's disease oscillations in the basal ganglia. *Front. Syst. Neurosci* 5:86. doi: 10.3389/fnsys.2011.00086

Kumaravelu, K., Brocker, D. T., and Grill, W. M. (2016). A biophysical model of the cortex-basal ganglia-thalamus network in the 6-OHDA lesioned rat model of Parkinson's disease. *J. Comput. Neurosci.* 40, 207–229. doi: 10.1007/s10827-016-0593-9

Leblois, A. (2006). Competition between feedback loops underlies normal and pathological dynamics in the basal ganglia. *J. Neurosci.* 26, 3567–3583. doi: 10.1523/JNEUROSCI.5050-05.2006

Leblois, A., Meissner, W., Bioulac, B., Gross, C. E., Hansel, D., and Boraud, T. (2007). Late emergence of synchronized oscillatory activity in the pallidum during progressive parkinsonism: pallidal activity during progressive parkinsonism. *Eur. J. Neurosci.* 26, 1701–1713. doi: 10.1111/j.1460-9568.2007.05777.x

Li, Q., Ke, Y., Chan, D. C., Qian, Z.-M., Yung, K. K., Ko, H., et aL. (2012). Therapeutic deep brain stimulation in Parkinsonian rats directly influences motor cortex. *Neuron* 76, 1030–1041. doi: 10.1016/j.neuron.2012.09.032

Liénard, J. F., Cos, I., and Girard, B. (2017). Beta-band oscillations without pathways: the opposing roles of D2 and D5 receptors. *bioRxiv*. doi: 10.1101/161661

Little, S., Pogosyan, A., Neal, S., Zavala, B., Zrinzo, L., Hariz, M., et al. (2013). Adaptive deep brain stimulation in advanced Parkinson disease: adaptive DBS in PD. *Ann. Neurol.* 74, 449–457. doi: 10.1002/ana.23951

Litvak, V., Jha, A., Eusebio, A., Oostenveld, R., Foltynie, T., Limousin, P., et al. (2011). Resting oscillatory cortico-subthalamic connectivity in patients with Parkinson's disease. *Brain* 134, 359–374. doi: 10.1093/brain/awq332

Liu, C., Zhu, Y., Liu, F., Wang, J., Li, H., Deng, B., et al. (2017). Neural mass models describing possible origin of the excessive beta oscillations correlated with Parkinsonian state. *Neural Netw.* 88, 65–73. doi: 10.1016/j.neunet.2017.01.011

Loucif, A. J., Woodhall, G. L., Sehirli, U. S., and Stanford, I. M. (2008). Depolarisation and suppression of burst firing activity in the mouse subthalamic nucleus by dopamine D1/D5 receptor activation of a cyclic-nucleotide gated non-specific cation conductance. *Neuropharmacology* 55, 94–105. doi: 10.1016/j.neuropharm.2008.04.025

Magill, P., Bolam, J., and Bevan, M. (2001). Dopamine regulates the impact of the cerebral cortex on the subthalamic nucleus–globus pallidus network. *Neuroscience* 106, 313–330. doi: 10.1016/S0306-4522(01)00281-0

Magill, P. J., Sharott, A., Bolam, J. P., and Brown, P. (2004). Brain State–dependency of coherent oscillatory activity in the cerebral cortex and basal ganglia of the rat. *J. Neurophysiol.* 92, 2122–2136. doi: 10.1152/jn.00333.2004

Mallet, N., Micklem, B. R., Henny, P., Brown, M. T., Williams, C., Bolam, J. P., et al. (2012). Dichotomous organization of the external globus pallidus. *Neuron* 74, 1075–1086. doi: 10.1016/j.neuron.2012.04.027

Mallet, N., Pogosyan, A., Marton, L. F., Bolam, J. P., Brown, P., and Magill, P. J. (2008a). Parkinsonian beta oscillations in the external globus pallidus and their relationship with subthalamic nucleus activity. *J. Neurosci.* 28, 14245–14258. doi: 10.1523/JNEUROSCI.4199-08.2008

Mallet, N., Pogosyan, A., Sharott, A., Csicsvari, J., Bolam, J. P., Brown, P., et al. (2008b). Disrupted dopamine transmission and the emergence of exaggerated beta oscillations in subthalamic nucleus and cerebral cortex. *J. Neurosci.* 28, 4795–4806. doi: 10.1523/JNEUROSCI.0123-08.2008

Mathai, A., Ma, Y., Paré, J.-F., Villalba, R. M., Wichmann, T., and Smith, Y. (2015). Reduced cortical innervation of the subthalamic nucleus in MPTP-treated parkinsonian monkeys. *Brain* 138, 946–962. doi: 10.1093/brain/awv018

McCarthy, M. M., Moore-Kochlacs, C., Gu, X., Boyden, E. S., Han, X., and Kopell, N. (2011). Striatal origin of the pathologic beta oscillations in Parkinson's disease. *Proc. Natl. Acad. Sci. U.S.A.* 108, 11620–11625. doi: 10.1073/pnas.1107748108

McIver, E. L., Chu, H.-Y., Atherton, J. F., Cosgrove, K. E., Kondapalli, J., Wokosin, D., et al. (2018). Chemogenetic restoration of autonomous subthalamic nucleus activity ameliorates Parkinsonian motor dysfunction. *bioRxiv*. doi: 10.1101/385443

Miguelez, C., Morin, S., Martinez, A., Goillandeau, M., Bezard, E., Bioulac, B., and Baufreton, J. (2012). Altered pallido-pallidal synaptic transmission leads to aberrant firing of globus pallidus neurons in a rat model of Parkinson's disease: increase in pallidal recurrent inhibition in experimental Parkinsonism. *J. Physiol.* 590, 5861–5875. doi: 10.1113/jphysiol.2012.241331

Moran, R. J., Mallet, N., Litvak, V., Dolan, R. J., Magill, P. J., Friston, K. J., et al. (2011). Alterations in brain connectivity underlying beta oscillations in Parkinsonism. *PLoS Comput. Biol.* 7:e1002124. doi: 10.1371/journal.pcbi.1002124

Nevado-Holgado, A. J., Mallet, N., Magill, P. J., and Bogacz, R. (2014). Effective connectivity of the subthalamic nucleus-globus pallidus network during Parkinsonian oscillations: effective connectivity of subthalamic nucleus-globus pallidus network. *J. Physiol.* 592, 1429–1455. doi: 10.1113/jphysiol.2013.259721

Oorschot, D. E. (1996). Total number of neurons in the neostriatal, pallidal, subthalamic, and substantia nigral nuclei of the rat basal ganglia: a stereological study using the cavalieri and optical disector methods. *J. Comp. Neurol.* 366, 580–599. doi: 10.1002/(SICI)1096-9861(19960318)366:4<580::AID-CNE3>3.0.CO;2-0

Oorschot, D. E., Zhang, R., and Wickens, J. R. (1999). "Absolute number and three-dimensional spatial distribution of rat neostriatal large interneurons: a first and second order stereological study," in *Proceeding of 10th International Congress Stereological*, Vol. 87 (Washington, DC).

Pan, M.-K., Kuo, S.-H., Tai, C.-H., Liou, J.-Y., Pei, J.-C., Chang, C.-Y., et al. (2016). Neuronal firing patterns outweigh circuitry oscillations in parkinsonian motor control. *J. Clin. Invest.* 126, 4516–4526. doi: 10.1172/JCI88170

Park, C., Worth, R. M., and Rubchinsky, L. L. (2011). Neural dynamics in Parkinsonian brain: the boundary between synchronized and nonsynchronized dynamics. *Phys. Rev. E* 83:042901. doi: 10.1103/PhysRevE.83.042901

Pavlides, A., Hogan, S. J., and Bogacz, R. (2015). Computational models describing possible mechanisms for generation of excessive beta oscillations in Parkinson's disease. *PLoS Comput. Biol.* 11:e1004609. doi: 10.1371/journal.pcbi.1004609

Pavlides, A., John Hogan, S., and Bogacz, R. (2012). Improved conditions for the generation of beta oscillations in the subthalamic nucleus-globus pallidus network: generation of beta oscillations. *Eur. J. Neurosci.* 36, 2229–2239. doi: 10.1111/j.1460-9568.2012.08105.x

Paz, J. T. (2005). Rhythmic bursting in the cortico-subthalamo-pallidal network during spontaneous genetically determined spike and wave discharges. *J. Neurosci.* 25, 2092–2101. doi: 10.1523/JNEUROSCI.4689-04.2005

Plenz, D., and Kital, S. T. (1999). A basal ganglia pacemaker formed by the subthalamic nucleus and external globus pallidus. *Nature* 400, 677–682. doi: 10.1038/23281

Ramanathan, S., Tkatch, T., Atherton, J. F., Wilson, C. J., and Bevan, M. D. (2008). D2-like dopamine receptors modulate SKCa channel function in subthalamic nucleus neurons through inhibition of Cav2.2 channels. *J. Neurophysiol.* 99, 442–459. doi: 10.1152/jn.00998.2007

Ray, N., Jenkinson, N., Wang, S., Holland, P., Brittain, J., Joint, C., et al. (2008). Local field potential beta activity in the subthalamic nucleus of patients with Parkinson's disease is associated with improvements in bradykinesia after dopamine and deep brain stimulation. *Exp. Neurol.* 213, 108–113. doi: 10.1016/j.expneurol.2008.05.008

Reis, C., Sharott, A., Magill, P. J., van Wijk, B. C., Parr, T., Zeidman, P., et al. (2019). Thalamocortical dynamics underlying spontaneous transitions in beta power in Parkinsonism. *Neuroimage* 193, 103–114. doi: 10.1016/j.neuroimage.2019.03.009

Sadek, A. R., Magill, P. J., and Bolam, J. P. (2007). A single-cell analysis of intrinsic connectivity in the rat globus pallidus. *J. Neurosci.* 27, 6352–6362. doi: 10.1523/JNEUROSCI.0953-07.2007

Sanders, T. H., Clements, M. A., and Wichmann, T. (2013). Parkinsonism-related features of neuronal discharge in primates. *J. Neurophysiol.* 110, 720–731. doi: 10.1152/jn.00672.2012

Sanders, T. H. and Jaeger, D. (2016). Optogenetic stimulation of cortico-subthalamic projections is sufficient to ameliorate bradykinesia in 6-ohda lesioned mice. *Neurobiol. Dis.* 95, 225–237. doi: 10.1016/j.nbd.2016.07.021

Schultheiss, N. W., Edgerton, J. R., and Jaeger, D. (2010). Phase response curve analysis of a full morphological globus pallidus neuron model reveals distinct perisomatic and dendritic modes of synaptic integration. *J. Neurosci.* 30, 2767–2782. doi: 10.1523/JNEUROSCI.3959-09.2010

Sharott, A., Gulberti, A., Zittel, S., Tudor Jones, A. A., Fickel, U., Munchau, A., et al. (2014). Activity parameters of subthalamic nucleus neurons selectively predict motor symptom severity in Parkinson's disease. *J. Neurosci.* 34, 6273–6285. doi: 10.1523/JNEUROSCI.1803-13.2014

Sharott, A., Magill, P. J., Harnack, D., Kupsch, A., Meissner, W., and Brown, P. (2005). Dopamine depletion increases the power and coherence of $\beta$-oscillations in the cerebral cortex and subthalamic nucleus of the awake rat. *Eur. J. Neurosci.* 21, 1413–1422. doi: 10.1111/j.1460-9568.2005.03973.x

Sharott, A., Vinciati, F., Nakamura, K. C., and Magill, P. J. (2017). A population of indirect pathway striatal projection neurons is selectively entrained to Parkinsonian beta oscillations. *J. Neurosci.* 37, 9977–9998. doi: 10.1523/JNEUROSCI.0658-17.2017

Shen, K.-Z., and Johnson, S. W. (2005). Dopamine depletion alters responses to glutamate and GABA in the rat subthalamic nucleus. *NeuroReport* 16, 171–174. doi: 10.1097/00001756-200502080-00021

Sherman, M. A., Lee, S., Law, R., Haegens, S., Thorn, C. A., Hämäläinen, M. S., et al. (2016). Neural mechanisms of transient neocortical beta rhythms: Converging evidence from humans, computational modeling, monkeys, and mice. *Proc. Natl. Acad. Sci. U.S.A.* 113, E4885–E4894. doi: 10.1073/pnas.1604135113

Shin, R.-M., Masuda, M., Miura, M., Sano, H., Shirasawa, T., Song, W.-J., et al. (2003). Dopamine D4 receptor-induced postsynaptic inhibition of GABAergic currents in mouse globus pallidus neurons. *J. Neurosci.* 23, 11662–11672. doi: 10.1523/JNEUROSCI.23-37-11662.2003

Shink, E., and Smith, Y. (1995). Differential synaptic innervation of neurons in the internal and external segments of the globus pallidus by the GABA- and glutamate-containing terminals in the squirrel monkey. *J. Comp. Neurol.* 358, 119–141. doi: 10.1002/cne.903580108

Smith, Y., Bolam, J. P., and Krosigk, M. (1990). Topographical and synaptic organization of the GABA-containing pallidosubthalamic projection in the rat. *Eur. J. Neurosci.* 2, 500–511. doi: 10.1111/j.1460-9568.1990.tb00441.x

Swan, C. B., Schulte, D. J., Brocker, D. T., and Grill, W. M. (2019). Beta frequency oscillations in the subthalamic nucleus are not sufficient for the development of symptoms of parkinsonian bradykinesia/akinesia in rats. *eneuro*. 6:ENEURO.0089-19.2019. doi: 10.1523/ENEURO.0089-19.2019

Tachibana, Y., Iwamuro, H., Kita, H., Takada, M., and Nambu, A. (2011). Subthalamo-pallidal interactions underlying parkinsonian neuronal oscillations in the primate basal ganglia: BG oscillations in Parkinson's disease. *Eur. J. Neurosci.* 34, 1470–1484. doi: 10.1111/j.1460-9568.2011.07865.x

Terman, D., Rubin, J. E., Yew, A. C., and Wilson, C. J. (2002). Activity patterns in a model for the subthalamopallidal network of the basal ganglia. *J. Neurosci.* 22, 2963–2976. doi: 10.1523/JNEUROSCI.22-07-02963.2002

Tsodyks, M., Pawelzik, K., and Markram, H. (1998). Neural networks with dynamic synapses. *Neural Comput.* 10, 821–835. doi: 10.1162/089976698300017502

Turrigiano, G. (2011). Too many cooks? Intrinsic and synaptic homeostatic mechanisms in cortical circuit refinement. *Annu. Rev. Neurosci.* 34, 89–103. doi: 10.1146/annurev-neuro-060909-153238

Velisar, A., Syrkin-Nikolau, J., Blumenfeld, Z., Trager, M., Afzal, M., Prabhakar, V., et al. (2019). Dual threshold neural closed loop deep brain stimulation in Parkinson disease patients. *Brain Stimul.* 12, 868–876. doi: 10.1016/j.brs.2019.02.020

Vitek, J. L., Zhang, J., Hashimoto, T., Russo, G. S., and Baker, K. B. (2012). External pallidal stimulation improves parkinsonian motor signs and modulates neuronal activity throughout the basal ganglia thalamic network. *Exp. Neurol.* 233, 581–586. doi: 10.1016/j.expneurol.2011.09.031

Wang, Y.-Y., Wang, Y., Jiang, H.-F., Liu, J.-H., Jia, J., Wang, K., et al. (2018). Impaired glutamatergic projection from the motor cortex to the subthalamic nucleus in 6-hydroxydopamine-lesioned hemi-parkinsonian rats. *Exp. Neurol.* 300, 135–148. doi: 10.1016/j.expneurol.2017.11.006

Wei, W., Rubin, J. E., and Wang, X.-J. (2015). Role of the indirect pathway of the basal ganglia in perceptual decision making. *J. Neurosci.* 35, 4052–4064. doi: 10.1523/JNEUROSCI.3611-14.2015

Weinberger, M., Mahant, N., Hutchison, W. D., Lozano, A. M., Moro, E., Hodaie, M., et al. (2006). Beta oscillatory activity in the subthalamic nucleus and its relation to dopaminergic response in Parkinson's disease. *J. Neurophysiol.* 96, 3248–3256. doi: 10.1152/jn.00697.2006

Yang, C., Yan, Z., Zhao, B., Wang, J., Gao, G., Zhu, J., and Wang, W. (2016). D2 dopamine receptors modulate neuronal resonance in subthalamic nucleus and cortical high-voltage spindles through HCN channels. *Neuropharmacology* 105, 258–269. doi: 10.1016/j.neuropharm.2016.01.026

Zhao, L., Beverlin, B., Netoff, T., and Nykamp, D. Q. (2011). Synchronization from second order network connectivity statistics. *Front. Comput. Neurosci.* 5:28. doi: 10.3389/fncom.2011.00028

Zhu, Z.-T., Shen, K.-Z., and Johnson, S. W. (2002). Pharmacological identification of inward current evoked by dopamine in rat subthalamic neurons *in vitro*. *Neuropharmacology* 42, 772–781. doi: 10.1016/S0028-3908(02)00035-7

**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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 09 January 2020
doi: 10.3389/fncom.2019.00086

# Principles of Mutual Information Maximization and Energy Minimization Affect the Activation Patterns of Large Scale Networks in the Brain

*Kosuke Takagi**

*Independent Researcher, Saitama, Japan*

Successive patterns of activation and deactivation in local areas of the brain indicate the mechanisms of information processing in the brain. It is possible that this process can be optimized by principles, such as the maximization of mutual information and the minimization of energy consumption. In the present paper, I showed evidence for this argument by demonstrating the correlation among mutual information, the energy of the activation, and the activation patterns. Modeling the information processing based on the functional connectome datasets of the human brain, I simulated information transfer in this network structure. Evaluating the statistical quantities of the different network states, I clarified the correlation between them. First, I showed that mutual information and network energy have a close relationship, and that the values are maximized and minimized around a same network state. This implies that there is an optimal network state in the brain that is organized according to the principles regarding mutual information and energy. On the other hand, the evaluation of the network structure revealed that the characteristic network structure known as the criticality also emerges around this state. These results imply that the characteristic features of the functional network are also affected strongly by these principles. To assess the functional aspects of this state, I investigated the output activation patterns in response to random input stimuli. Measuring the redundancy of the responses in terms of the number of overlapping activation patterns, the results indicate that there is a negative correlation between mutual information and the redundancy in the patterns, suggesting that there is a trade-off between communication efficiency and robustness due to redundancy, and the principles of mutual information and network energy are important to network formation and its function in the human brain.

**Keywords: functional connectome, information processing, mutual information, network energy, activation pattern, large scale brain network**

OPEN ACCESS

***Edited by:***
*Yu-Guo Yu,*
*Fudan University, China*

***Reviewed by:***
*Lianchun Yu,*
*Lanzhou University, China*
*Tuo Zhang,*
*Northwestern Polytechnical University, China*
*Rubin Wang,*
*East China University of Science and Technology, China*

***Correspondence:**
*Kosuke Takagi*
*coutakagi@mes.biglobe.ne.jp*

***Received:*** *23 August 2019*
***Accepted:*** *12 December 2019*
***Published:*** *09 January 2020*

***Citation:***
*Takagi K (2020) Principles of Mutual Information Maximization and Energy Minimization Affect the Activation Patterns of Large Scale Networks in the Brain. Front. Comput. Neurosci. 13:86. doi: 10.3389/fncom.2019.00086*

## 1. INTRODUCTION

Interactions of ∼100 billion neurons, which are a part of the human brain, maintain its functions within a hierarchical and modular network structure (Azevedo et al., 2009; Meunier et al., 2010; Park and Friston, 2013). Empirical evidence demonstrate that a stimulus for local excitatory neurons at a cellular level can be etiologically associated with large-scale brain activity, which may propagate through numerous neuronal interconnections (Beggs and Plenz, 2003; Beggs, 2008; Lee et al., 2010; Fenno et al., 2011; Tagliazucchi et al., 2012). Over the years, studying task evoked brain activity via whole-brain imaging has been successful in mapping specific cognitive functions onto distinct regions of the human brain (e.g., Kanwisher et al., 1997).

Furthermore, several studies that have examined the brain's responses to more complex tasks, reported that various cognitive functions arise from interactions between regions of the brain rather than independent single activities in distinct regions of the brain (Ghazanfar and Schroeder, 2006; Bressler and Menon, 2010). In the large-scale networks of the human brain, activation signals from segregated and specialized regions are integrated in information processing (Tononi et al., 1994; Hilgetag and Grant, 2000; Sporns, 2013). Thus, the brain can be conceptualized as an information processing system, hereby successive patterns of activation and deactivation in multiple distributed regions constitute integrated information processing. Furthermore, the brain must adapt to changing environments, so these processes might be optimized to ensure rapid and flexible response (Bassett et al., 2006; Kitzbichler et al., 2009; Clark, 2013; Park and Friston, 2013; Mnih et al., 2015). On the other hand, the brain is limited by its energy requirements and by other biological realities (Bullmore and Sporns, 2012). Thus, the need to maximize efficiency of information processing and minimize total energy consumption may regulate the mechanisms underlying the structure and the function of the brain (Linsker, 1990; Friston, 2010; Bullmore and Sporns, 2012).

This argument is known as the energy efficiency hypothesis, which covers a wide range of activities from the cellular level of neurons to the global level observed at the scale of the whole brain (Bullmore and Sporns, 2012; Yu and Yu, 2017). Evidence for this hypothesis has shown that the energy constraints and limitations may affect multiple aspects of the brain neurons by inducing efficient activities (e.g., Niven and Laughlin, 2008; Tomasi et al., 2013; Yu and Yu, 2017). The energy consumption models of neurons have especially been studied in detail, and they have revealed the requirements from energy efficiency effects on neuronal activities or on those at the cortical level follow the energy efficient principle (Wang et al., 2008, 2015, 2018; Wang and Wang, 2014).

In the present paper, I present evidence that this pattern is especially the case in the information integrating processes in a large scale network, demonstrating that maximization and minimization principles guide the network structure and activation patterns of the human brain. Based on functional connectome data acquired using resting-state functional MRI (fMRI) (Sporns, 2002; Fox and Raichle, 2007; van den Heuvel et al., 2008; Greicius et al., 2009; Biswal et al., 2010; Van Dijk et al., 2010; Brown et al., 2012), I simulated information transfer by applying randomly activated signals to a network represented by brain connectivity matrices (Takagi, 2018). I measured mutual information (Linsker, 1990) between random stimulus signals and their responses and also quantified the network energy associated with these activities (Hopfield, 1984; Hinton and Salakhutdinov, 2006). By varying the functional connectivity network between noisy and sparse states, I showed an explicit correlation between these quantities. The results suggest that there is an optimal intermediate between these states, whereby mutual information is maximized and the network energy is minimized.

On the other hand, evaluation of the network structure around this optimal intermediate state revealed some features that are characteristic of the functional connectome, such as small-world and criticality (Watts and Strogatz, 1998; Achard et al., 2006; Bassett and Bullmore, 2006; Hagmann et al., 2008; van den Heuvel and Sporns, 2011; Takagi, 2017, 2018). These characteristic attributes are thought to explain the brain's rapid adaptive responses to external stimuli and the robustness of its internal communication (Kitzbichler et al., 2009; Chialvo, 2010; Tagliazucchi et al., 2012). Experiments at a cellular level demonstrated that neuronal firing successively propagated similar to neuronal avalanches; however, their size has no characteristic scale (Beggs and Plenz, 2003; Beggs, 2008). However, analyzing the fMRI dynamics revealed that the dynamic and statistical properties which regulate activation events on a scale of the whole brain were identical (Tagliazucchi et al., 2012). This feature of the dynamics appeared across multiscale from the cellular level to the brain macro scale is explained by the feature of the criticality, the absence of the characteristic scale (Beggs and Plenz, 2003; Beggs, 2008; Tagliazucchi et al., 2012). Additionally, optogenetic methods combined with fMRI facilitate direct visualization of the global level activity caused by local neuronal excitation (Lee et al., 2010; Fenno et al., 2011). Besides the absence of a characteristic scale for these dynamical activation events, an identical feature that is predicted from the criticality can be confirmed in the functional network structure, which was constructed using the spatio-temporal correlations between brain regions. To illustrate, network node degree statistics exhibit the distribution characteristic similar to the critical phenomenon (Achard et al., 2006; Bassett and Bullmore, 2006; Hagmann et al., 2008; Takagi, 2017). Moreover, within these networks, strongly connected pathways compose core structures with highly connected hub regions that modulate information processing in the brain (Hagmann et al., 2008; van den Heuvel and Sporns, 2011). Processing in these regions may control multiple brain functions (Rubinov and Sporns, 2011). The results show that, to ensure optimal efficiency and energy use, the network structure converges on this characteristic state exhibiting small-world and criticality.

Further analyses of the simulation results of the information transfer model revealed direct evidence that this characteristic state regulates activation patterns (Takagi, 2018). In the simulation, response patterns exhibited redundancy in that they contained repeatedly co-activated regions with different

stimulation signals. In the cerebral cortex, activation patterns exhibit overlapping that can be measured as the proportion of regions activated equally in the different patterns. This can in turn be related to cognitive processes, such as memory retrieval (Haxby et al., 2001; Kumaran et al., 2016). While functionally overlapped regions may offer robustness in communication and facilitate adaptation (Whitacre, 2010; Bassett et al., 2018), excess overlapping causes interference. This can result in decoding difficulties that can be costly in terms of metabolic consumption (Kumaran et al., 2016). In the present study, the average of the overlapping numbers depended on the mutual information and the network energy. It showed the negative correlation to the mutual information and the correlation to the network energy. The results imply that the principles of mutual information and network energy strongly affect the activation patterns and the underlying structure of the functional network in the brain.

On the other hand, it is known that the functional connectivity is flexible within certain dynamics; for example, alterations in diseased brains, or the break-down from criticality in the unconsciousness, have been reported (Tagliazucchi et al., 2016; Song et al., 2019). As such, the robustness of the simulation results in this paper are validated, in comparison to different datasets, such as those with different sized matrices with different sets of nodes and those that were constructed from the structural connections based on diffusion tensor imaging (DTI) (Sporns et al., 2005; Brown et al., 2012). They also indicate that the relationships between the mutual information, the energy of the activation, and the activation patterns that emerge are stable within these networks as well.

# 2. MATERIALS AND METHODS

## 2.1. Connectome Datasets and Information Transfer Model

### 2.1.1. Functional Connectome Datasets

I modeled information transfer in the large scale network of the human brain using a functional connectivity matrix constructed from fMRI observation (Takagi, 2018). As explained in the introduction, a stimulus at the cellular level can trigger avalanche events at a whole-brain scale due to the characteristic features of the critical phenomenon. Whole-brain scale observation through fMRI revealed that neighboring voxels overlapping in their dynamics show similarities in time series data, because of successive appearances of these events (Calhoun et al., 2009; Smith et al., 2011; Smith, 2012; Tagliazucchi et al., 2012). Therefore, it is possible that the information relevant to the underlying brain activity is compressed (Tagliazucchi et al., 2012). Furthermore, a relevant network model is constructed by extracting nodes, through independent component analysis or clustering voxels on the basis of the similarity (Calhoun et al., 2009; Smith et al., 2011; Smith, 2012; Tagliazucchi et al., 2012).

To accurately analyze the network in the whole-brain scale, hundreds of nodes are typically utilized to construct a network from fMRI time series data (Smith, 2012; Finn et al., 2015). The validity of the network construction is then indicated by the robustness for different individual subjects (Smith, 2012; Finn et al., 2015). Here, the validation of the pre-processed network datasets was demonstrated by the results of my previous study using the same dataset, which reported a stable statistical significance regarding the network structure (Takagi, 2018). Additionally, the robustness of the current study and the consistency with other studies will be discussed in the final section.

For each combination of single regions in the brain, the connectivity matrix was described as a matrix ($w_{ij}$), whereby $(i, j)$ represented the connection weight between regions denoted as $i$ or $j$. For the time series data of the fMRI image, the connectivity was calculated as the Pearson correlation coefficient between voxels corresponding to these regions. In the present study, I used the preprocessed connectivity matrices, which are available from http://umcd.humanconnectomeproject.org/: the website of the USC Multimodal Connectivity Database (Brown et al., 2012), which contains matrices constructed from the functional connectome datasets of the "1,000 connectome project" (Biswal et al., 2010). The original datasets in this project were obtained using resting-state fMRI (R-fMRI), which records activation patterns in brain regions during the resting state and is thought to describe the common architecture of the human brain (van den Heuvel et al., 2008; Greicius et al., 2009; Biswal et al., 2010; Van Dijk et al., 2010; Brown et al., 2012). The matrices comprised $N \times N$ elements with $N = 177$ brain regions and were assumed to cover the entire brain. The details of the processing sequence to construct these matrices are shown in the above website and, in this analysis, I use 986 matrices for different individuals, which are available from the same site (Brown et al., 2012).

Brain activity naturally fluctuates and the connectivity matrix contains noise and artifacts (Eguiluz et al., 2005; Fox and Raichle, 2007; Brown et al., 2012). To construct the network structure with significant elements, threshold was applied to the matrix ($w_{ij}$) (Eguiluz et al., 2005; Brown et al., 2012; Zuo et al., 2012). Because strongly connected pathways form core structures that are relevant to the network structure of the brain (Eguiluz et al., 2005; Brown et al., 2012), I removed connections with small connectivity weights using a threshold and constructed the network with the residual connections. After introducing the threshold $w_t$ for the connectivity weight $w_{ij}$, I obtained a network description consisting of connections corresponding to the $|w_{ij}| > w_t$ elements. In this analysis, considering the differences between individuals, I defined the threshold value of each individual connectivity matrix $w_t$ based on the average connectivity $< |w| >$ and the standard deviation $\sigma_{|w|}$. I calculated $< |w| >$ and $\sigma_{|w|}$, and defined the cut-off threshold in terms of the following equation:

$$w_t =< |w| > +n \cdot \sigma_{|w|}, \tag{1}$$

with a parameter of $n$.

### 2.1.2. Structural Connectome Dataset

The simulation results based on the above functional connectome datasets were compared to the structural connectome, and the other connectome datasets describing the physical connection between brain regions. The structural connectome datasets are

constructed by the diffusion tensor imaging (DTI) method, which traces the fiber tracts between brain regions and forms another network at the whole brain scale, known as the structural connectome (Sporns et al., 2005). The dataset is available from the above website (http://umcd.humanconnectomeproject.org/) with the pre-processed matrix of the connectivity strength being the same to the fMRI cases (Brown et al., 2012). The DTI dataset is taken from a subsect of the "1,000 connectome project," tagged as "*NKI_Rockland*" for the "Study Name" item, from the Nathan Kline Institute (NKI)/Rockland sample in the web site. It contains the matrices of 196 individuals, and each matrix has $N = 188$ matrix elements ($188 \times 188$).

#### 2.1.3. Information Transfer

Information in the brain is transferred by successive signal propagation; this can be represented by the activated state of each site (Tononi et al., 1994; Hilgetag and Grant, 2000; Beggs and Plenz, 2003; Ghazanfar and Schroeder, 2006; Beggs, 2008; Bressler and Menon, 2010; Sporns, 2013; Takagi, 2018). For each node in the functional network, three states $\{1, -1, 0\}$ were assigned because the responses of neuronal activity can be categorized as positive and negative (Fox et al., 2005; Shmuel et al., 2006). In this representation, the inactivated regions were assigned the 0 state, while the two states at $\pm 1$ represented positive and negative activation states, respectively.

When considering information transfer, I represented a whole state of the brain as $S = (s_1, \ldots, s_N)$ for a network size $N$, whereby the $i$-th node was assigned as $s_i \in \{1, -1, 0\}$. I could then calculate the responses $R = (r_1, \ldots, r_N)$ $r_i \in \{1, -1, 0\}$ for a given connectivity matrix and threshold. For the given set of $S$ and connectivity matrix ($w_{ij}$), the response state was evaluated using the following equation:

$$r_j = \sigma(\sum_{i \in N} w_{ij} s_i). \tag{2}$$

I denoted $\sum_{i \in N} w_{ij} s_i$ as $\hat{r}_j$, so a threshold of $w_t$, $\sigma(\hat{r}_j)$ was defined as $r_j = 1, -1, 0$ for cases $\hat{r}_j > w_t$, $\hat{r}_j < -w_t$, and $|\hat{r}_j| \leq w_t$. In this simulation, I calculated the information transfer of stimuli $S$. The input signals were taken randomly, although I did use the same probability for positive and negative activation. I then assigned 1 and $-1$ to each input signal $s_i$, with the probability $p$ being set to 0 in the other cases with the probability $1 - 2p$. Each condition in this simulation was repeated 100 times with each input signal.

### 2.2. Statistical Quantities of Information Transfer Model

#### 2.2.1. Mutual Information

To measure information transfer from the imposed stimuli to the responses, I evaluated the mutual information for the set of stimulus signals $S$ and the corresponding responses $R$. It is defined as $H(R) - H(R|S)$ with $H(R)$, the information of the response $R$, and $H(R|S)$, the conditional entropy. This quantity was used to assess the efficiency of information transfer in the neural network models and in real biological data (Beggs and Plenz, 2003; Beggs, 2008).

In the analysis, the mutual information of the transfer between $i$ and $j$ nodes was estimated using the following equation:

$$m(i, j) = H(s_i) + H(r_j) - H(s_i, r_j), \tag{3}$$

where the entropy $H(s_i)$ and $H(r_j)$, as well as the joint entropy $H(s_i, r_j)$, were calculated using the probabilities of each state: $s_i, r_j \in \{\pm 1, 0\}$. Next, this quantity was estimated for the whole network as follows: $m = \sum_j < m(j)) > /N$, with averaging as $< m(j) >= (\sum_i m(i, j))/(N - 1)$ for all possible connections of each node $j$.

#### 2.2.2. Network Energy

The energy of the brain network is described in different ways, which are mainly categorized into wiring costs for organizing the network structure and those related to their activity. The total number of connections determine the wiring cost to organize the network structure (Achard and Bullmore, 2006; Bullmore and Sporns, 2012). Thus, the wiring cost based on the topological structure basically describes the energy demands of the brain functional network. It is assumed that many characteristic attributes of the brain network can be explained by minimizing the wiring cost (Bullmore and Sporns, 2012).

Hence, I defined this energy, the wiring cost denoted as $E_W$, using the following equation:

$$E_W = \sum_{i,j} a_{i,j}, \tag{4}$$

where $a_{i,j}$ denotes the element of the adjacency matrix. For an undirected topological graph of a given matrix, the connection for each pair of $i$ and $j$ was represented using the adjacency matrix element, which is connected as $a_{i,j} = 1$ for $|w_{ij}| > w_t$, with threshold $w_t$, and disconnected as $a_{i,j} = 0$ in other cases.

On the other hand, the Hopfield energy gives a definition related to the dynamics and the associated information of the neural networks. For a given network state of activation, the Hopfield energy provides one definition of the network energy. It models the network state of the neurons and can also be applied to artificial neuronal networks (Hopfield, 1984; Hinton and Salakhutdinov, 2006). Hopfield networks and similar types of energy representation have been introduced to describe the energy state of neural networks, modeling the spin glass network (Hopfield, 1984). One example of the artificial learning models that use this type of function is the restricted Boltzmann machine, which evolves by adjusting the network variables according to rules learned from the energy function (Hinton and Salakhutdinov, 2006). It is defined as

$$E_H = -(\sum_{i,j} r_i w_{i,j} r_j) \tag{5}$$

whereby I took a bias-free case in accordance with the transfer model Equation (2).

In the original definition of the Hopfield energy (Hopfield, 1984) bias terms are present, such as those expressed as $\sum_i r_i b_i$ with constant bias $b_i$ assigned the value for each node. In

this simulation, they are excluded as the constants under the assumption of homogeneity of nodes. According to this simplification, the simulation is given in the bias-free form, and takes 0 for the cases, such as the random state as well as negative values, especially in the low energy states.

However, as the indicator of total activity cost, the Hopfield energy would be estimated as small as the positive and negative terms were not included in the definition. To avoid this cancelation and estimate the total energy cost for the activity, I introduced a definition of the activity cost using the absolute values of each term and compared them with the values mentioned above Equation (5). The definition given was as follows:

$$E_A = \sum_{i,j} |r_i w_{i,j} r_j| \tag{6}$$

represents the total energy cost for the activation dynamics. It assesses the contributions from the positive and the negative signal equally and then evaluates the total amount of signal activations with their weights. In the discussion, I assessed the energy of the functional brain network based on these definitions.

### 2.2.3. Overlapping Number in the Activation Patterns

I analyzed the pattern of the response signals $R = (r_i)$ using the overlapping numbers of the different signals. I evaluated it in terms of the number of regions activated or deactivated equally with the different patterns. For the set of response patterns $R^j = (r_i^j)$, whereby $j$ is an index of the input state, I counted the number of the same responses $r_i^j = r_i^{j'}$ for the pair $j$ and $j'$. I normalized this overlapping number by the total number of regions $N$. I then wrote it down as

$$h(j, j') = \sum_i \pi(r_i^j, r_i^{j'})/N \tag{7}$$

where $\pi(r_i^j, r_i^{j'})$ is 1 for $r_i^j = r_i^{j'}$ and 0 was taken in the other cases. I then took the averages of all the pairs of $R^j$ and $R^{j'}$.

The definition of the overlapping number (Equation 7) is the same as that of the Hamming distance of the information theory. It is used, for example, to measure the error in the signal transfer. In the analysis, it was used to analyze the relationship between the activation patterns and efficiency of the information transfer. As excess overlapped states indicated that the variation in the response $S$ is lost, they resulted in the decrease in the mutual information entropy.

The program for this network model is available at https://github.com/coutakagi/fcn2019.git.

## 2.3. Network Structure and Statistical Evaluation

The functional connectome is often described in topological or weighted terms. Different measures are required to assess the topological network structure, especially in terms of criticality. To specify the criticality in the activation dynamics, the characterization is given by the statistics of the avalanche events. One measure is the mutual information entropy, such as defined above, which is maximized in this state in comparison to the super-critical state (in which excess activation is saturated) and the sub-critical state (in which activations die out due to poor sensitivity to the stimulus) (Beggs and Plenz, 2003; Beggs, 2008). This is contrasted to the criticality of the topological structure, which is usually characterized by appearances of the giant connected component or other states, such as the small-world topology (Watts and Strogatz, 1998), which are evaluated by quantities, such as degree or the clustering coefficients.

Besides the total number of connections, topological structures were measured in terms of the largest connected component to provide a basic measure of the topological network. With using the adjacency matrix, the size of each connected component was then measured in terms of the number of nodes in each connected subgraph, and these values determined the largest connected component of each network. In the present paper, I measured this quantity using R-package igraph (Barrat et al., 2004).

On the other hand, to account for connectivity strength $w_{ij}$, I took the absolute node strength value $ns_i = \sum_j |w_{ij}|$ in each node and evaluated its statistical features using a distribution model (Takagi, 2017, 2018). Due to the criticality of the brain (Achard et al., 2006; Bassett and Bullmore, 2006; Hagmann et al., 2008; van den Heuvel and Sporns, 2011; Takagi, 2017, 2018), the distributions of network variables, such as degree, exhibit a characteristic shape similar to the power law. However, when I adapted the power law to the distributions, this straightforward application was prohibited because the energy constraints on brain activity constitute an upper limits (Takagi, 2017, 2018). In the present study, the same assumption was applied, and I introduced an upper strength limit of $ns_{max}$. Following this assumption, I obtained an expression for the normalized variable $\tilde{s} = (ns_{max} - ns)$ as

$$p(ns) \propto (\tilde{ns})^\gamma = (ns_{max} - ns)^\gamma, \tag{8}$$

with a constant $\gamma$ (Takagi, 2017, 2018).

Next, I assessed the strength distribution $ns$ in terms of deviations from this model using the Kolmogorov-Smirnov (KS) distance (Clauset et al., 2009; Klaus et al., 2011). For the cumulative distribution $p_e(ns)$, which was experimentally given, and that of the model $p_c(ns)$, which was fitted to the data, the KS distance $D$ was defined using the following equation:

$$D = \max_w |p_e(ns) - p_c(ns)| \tag{9}$$

which measures the maximum distance of the model from the experimental data. If this value was sufficiently small, the network probably exhibits the feature characterized by this distribution model.

Finally, I measured the clustering coefficient $C$, also known as transitivity, for each adjacency matrix. This is another important topological quantity which is often used as an indicator of the small-world network (Watts and Strogatz, 1998). It is defined as the probability that the adjacent vertices of a vertex are connected (Watts and Strogatz, 1998). Here, it is measured for each adjacency matrix, using the R-package igraph (Barrat et al., 2004).

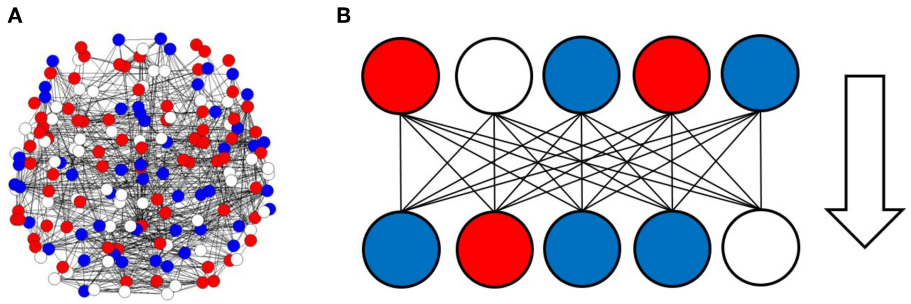


**FIGURE 1 |** Activation and deactivation patterns in the local regions, and the information transfer associated with these patterns. **(A)** Activation and deactivation in the brain is illustrated. In this figure, each circle represents the states of local regions in the brain, with solid lines corresponding to the connection between regions. Activated regions are represented by red circles, while negatively activated ones are colored in blue. The residual white circles correspond to other inactivated states. **(B)** The information transfer associated with these patterns is illustrated. The pattern state in the upper side, which is shown on the line, consists of signals transferred to the lower sides, where each region state is changed according to the upper input patterns and the connection strengths between the regions.

# 3. RESULTS

## 3.1. Information Transfer Model

To analyze the information processing in the large scale network of the human brain, I simulated information transfer using successive activation patterns. Because activity in the brain can be observed as activation and deactivation in local regions, signal transmission associated with information processing can be described in terms of successive changing at each site, with positive or negative activation (Beggs and Plenz, 2003; Fox et al., 2005; Ghazanfar and Schroeder, 2006; Shmuel et al., 2006; Beggs, 2008; Bressler and Menon, 2010; Takagi, 2018). In the model, the given brain state sites, as illustrated in **Figure 1A**, were transferred to successive states, which were determined by the correlation among the sites given by the matrix $(w_{ij})$ as **Figure 1B**.

In the simulation, I calculated the response state $R = (r_i)$ of the randomly stimulated signals $S = (s_j)$, as represented by Equation (2), using the connectivity matrices $(w_{ij})$ of the human connectome. Next, as shown in **Figure 2**, I evaluated the efficiency of transfer of the mutual information, defined as the average of Equation (3). As part of the preliminary evaluation, I used randomly selected 100 individual matrices for calculation. I compared this quantity among the different states, which were varied in terms of the connectivity strength threshold value $w_t$ and the activation probability of the input stimuli. As shown in this figure, information transfer depended on these parameters, while the activation probability $p = 0.05$ gave the maximum values for these different conditions. Starting from the flatten values for lower thresholds due to its negative threshold value on the left end, the measurements of the mutual information entropy increased to their maximum values in the intermediate states. Moreover, the standard deviations for the thresholds $n = 1.0, 0, -1.0$ were evaluated for $p = 0.05$ as $9.11 \times 10^{-2}, 7.45 \times 10^{-2}, 1.34 \times 10^{-1}$. These values were smaller than their mean values, and these results were stable. Because I were interested in the state with maximum mutual information, I used this value, $p = 0.05$, in the following simulation.

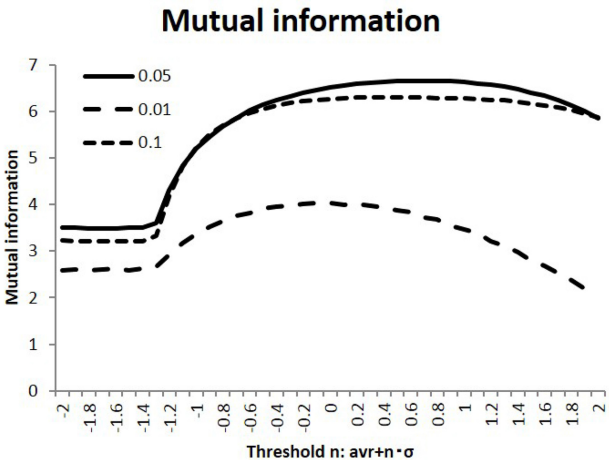


**FIGURE 2 |** Mutual information with different activation probabilities. Three different values of the activation probability are shown: $p = 0.1, 0.05, 0.01$. These conditions are shown with the dotted, the dashed, and solid lines, respectively. The different states are measured, with the threshold values on the x-axis being varied. I took the threshold as $< |w| > + n \cdot \sigma_{|w|}$ with the connectivity average $< |w| >$ and its standard deviation $\sigma_{|w|}$ for each connectivity matrix. Following this, the mutual information is evaluated by taking the average of Equation (3).

## 3.2. Network Energy and Efficiency of the Information Transfer

Constraints regarding energy would be a major factor regulating network structure and activity in the brain (Bullmore and Sporns, 2012). Hence, I evaluated the network energy of each brain state, which is a basic parameter to analyze brain activity. Then, I showed the results of the measured energies using three different definitions in **Figure 3**; in each graph, the connectivity strength threshold differed. Further, I compared these values with those of the random networks, which were considered as the null model. The random networks with the same network size were determined together with the randomly taken weights

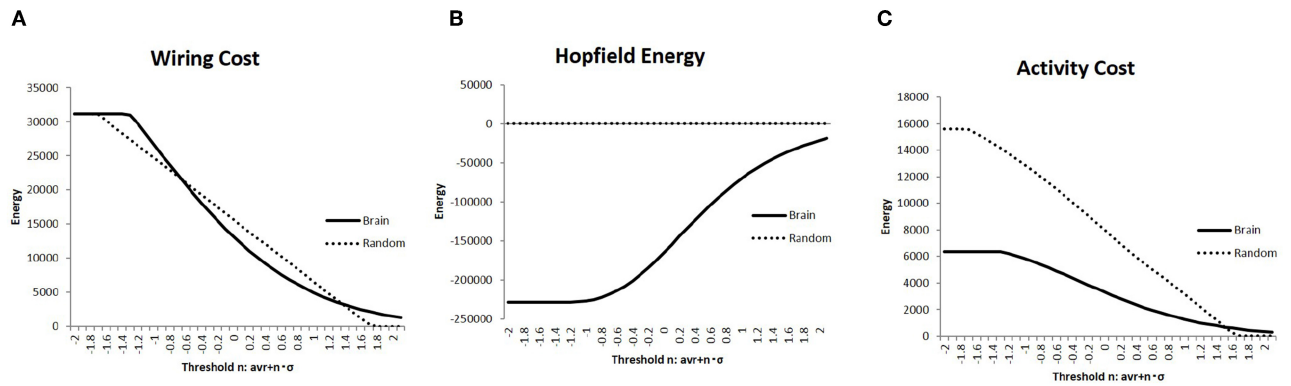


**FIGURE 3 |** The three different types of energies, the wiring cost (Equation 4), the Hopfield energy (Equation 5), and the activity cost (Equation 6) are shown in **(A–C)**, respectively, with the solid lines on each graph. They are compared with the results of the random network, which are shown in the dashed lines. I applied the same threshold set of values to **Figure 2**.

$w_{i,j} \in [-1, 1]$, and 1,000 random matrices were obtained. On the contrary, the results for the brain network were measured using the whole datasets, which contained 986 matrices of different individuals.

The wiring cost defined in Equation (4) is shown in **Figure 4A**. It was compared with the wiring cost of the random network, which appeared as a straight line proportional to $n$ of the threshold value defined in Equation (1). In comparison with these networks, it was found that the plotted curve of the brain network has a relatively long tail for higher values of $w_t$, which indicated the well-known attributes of the brain network, such as the scale-free and small-world network, as will be discussed.

The difference between the brain network and the random one was enhanced in comparison with the values of the Hopfield energy. **Figure 4B** shows the relatively large values for the brain, while it took almost 0 for the random network due to the cancellation of the positive and negative terms. To avoid this cancellation and evaluate the total amount of the activity cost, I calculated the energy with another definition given in Equation (6) and plotted it in **Figure 3C**. The energy for each range except for 0 states had higher $w_t$ values; the activity cost for the random model was higher than those of the brain network as expected.

## 3.3. Normalized Energy and the Mutual Information Entropy

Due to energy constraints, it was assumed that the activities for the information transfer is required to be efficient (Bullmore and Sporns, 2012). One description of the network efficiency for a given cost was based on the energy consumed during the activity, which was normalized by the wiring cost to organize the network structure (Takagi, 2017). Then, at first, the Hopfield energy was normalized with the wiring cost as $E_H/E_W$ and shown in **Figure 4A**. With regards to mutual information, the correlations are depicted in **Figure 4B**, which indicates a negative correlation, whereby decreasing the network energy resulted in increases in mutual information. The same figure shows that there was a peak around the maximum point of mutual information, where mutual information was maximized and the network energy associated with activity was minimized.

To clarify the cost performance of the activity in the brain, I took another quantity, $E_W/E_A$, the wiring cost (Equation 4) normalized by the activity cost (Equation 6). This normalized quantity represents the wiring cost required to maintain a unit amount of activity. The measurement is then shown in **Figure 4C**, and its correlation with the mutual information entropy is presented in **Figure 4D**. It shows the clear correlation with a sharp peak, around which the mutual information is maximized and the normalized wiring cost is minimized. These results (**Figures 4B,D**) for different definitions of the normalized energy exhibit the similar behavior and the clear dependency of the mutual information entropy on the network energy. Thus, these peaks on correlations define the optimal state of the brain functional network, in which the efficiency of information transfer for a given network energy cost was maximized.

## 3.4. Network Structure and the Optimal State

I analyzed the network structure around this peak state. At first, the topological network structure of each state around this point was characterized in terms of the largest component size: a basic quantity of the network topology. This result is shown in **Figure 5A**, wherein the component size is shown normalized to the network size. In the same graph, the largest component size of the connected subgraph decreases with increasing threshold, with the normalized size being 1, which corresponds to the fully connected graph. Next, I took the correlation between mutual information and this quantity in **Figure 5B**. The sharp peak on this graph indicates that maximum information was realized in the fully connected graph with minimum connections.

As shown in **Figure 6**, the topological network graph contains excess connections in the lower threshold. In this state, signals with information transfer also contain noise due to these excess elements. At the higher threshold value, the network loses this fully connected structure, and the graph is fragmented into

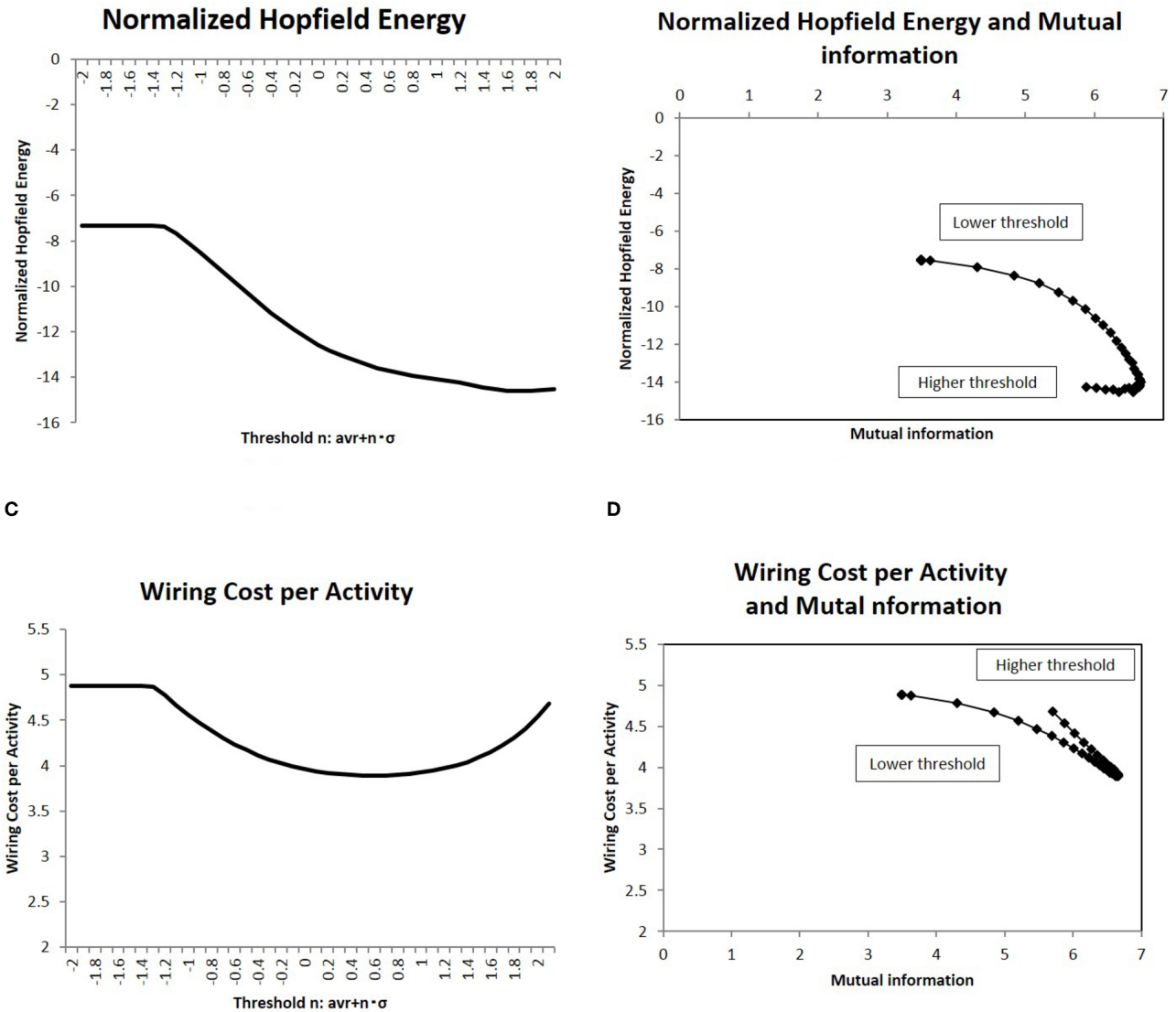


**FIGURE 4 |** Network energy and the correlation with mutual information. **(A)** The Hopfield energy $E_H$ defined in Equation (5) is normalized by wiring cost $E_W$, Equation (4), and this normalized energy $E_H/E_W$ is shown for the different threshold values. I applied the same threshold set of values to **Figure 2**. **(B)** showed a correlation between mutual information and the normalized Hopfield energy. On the x-axis, I took the mutual information calculated using the whole datasets of matrices by applying the method, which is denoted by a solid line ($p = 0.05$) (**Figure 2**). I plotted the corresponding points of each threshold value with the network energy on the y-axis, as given in **(A)**. **(C)** Another normalized energy, $E_W/E_A$ where the wiring cost $E_W$ is defined as Equation (4) and the activity cost $E_A$ is defined as Equation (6), is shown. The threshold is the same as that of the case of **(A)**. Its correlation with mutual information entropy on **(D)** is the same as the case of **(B)**.

multiple disconnected sub-components, as shown in **Figure 6A**. In this state, mutual communication between disconnected nodes is hindered, so the efficiency of the information transfer might be reduced. The sharp peak on **Figure 5B** corresponds to the boundary state between these two states, where the network preserves the fully connected structure with minimum connections. Combined with the correlation between mutual information and network energy (**Figures 4B,D**, **5B**), this result can be interpreted as showing that efficiency and energy consumption are optimized in this state, with a fully connected structure that eliminates transfer noise.

Because this optimal state resides in the boundary state between the fully connected and fragmented phases, it constitutes a critical state, whereby connectivity strength, another important variable of network structure, shows a characteristic distribution. To introduce a distribution model for this critical state (Equation 8), I measured the statistical deviation of the total connectivity strength of each node *ns* in **Figure 7** using the KS distance,

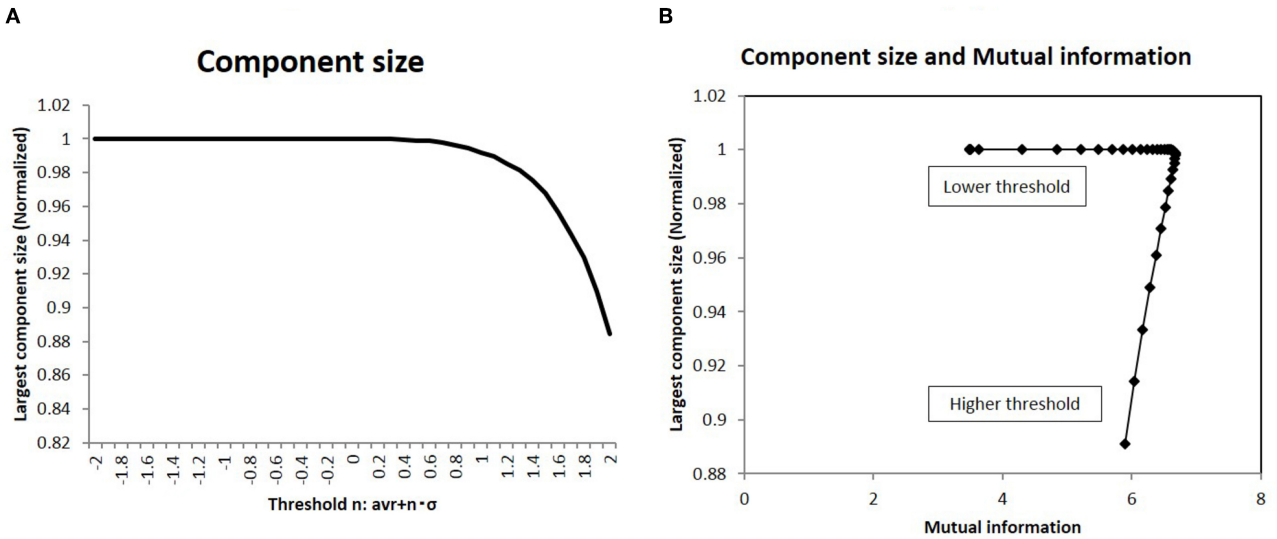


**FIGURE 5 |** The largest component size of the functional network and the mutual information **(A)**. I showed the largest component size of the topological representation for each threshold values. I took the average of the values of the largest component size, which are normalized by the total number of nodes in the topological graph. **(B)** I show the relation between the mutual information and the largest component size. On the x-axis, I took the mutual information calculated using the whole datasets of matrices by applying the method, which is denoted by a solid line ($p = 0.05$) (**Figure 2**). Then I plot the corresponding points for each threshold value with the largest component size on the y-axis, which are given in **(A)**.

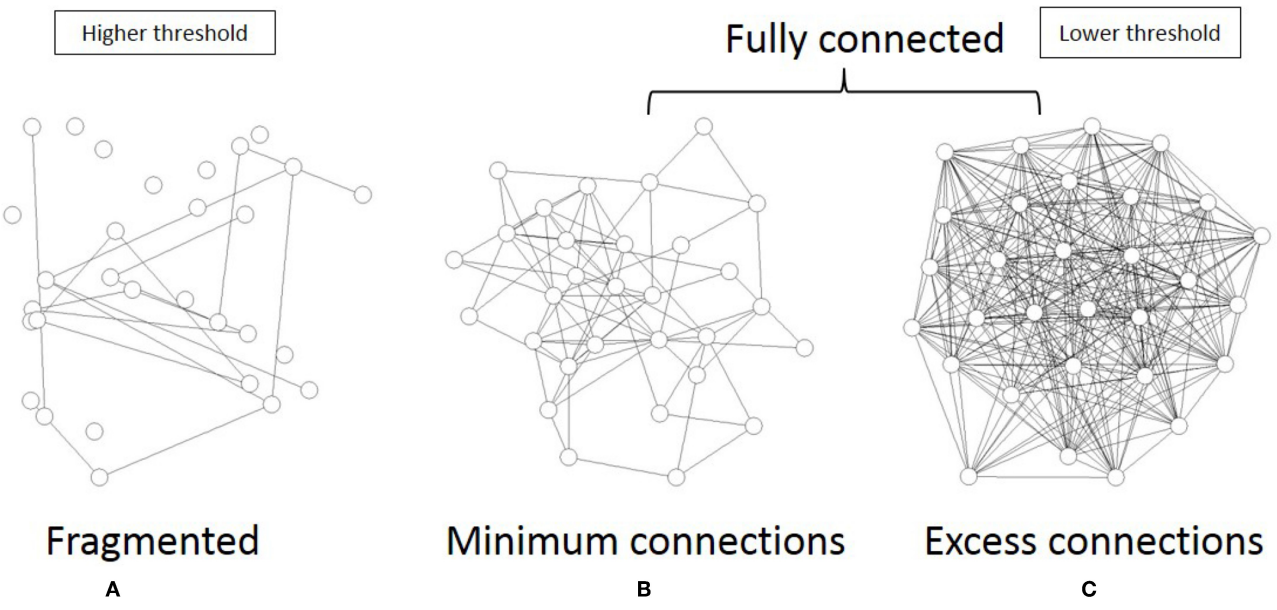


**FIGURE 6 |** I illustrate the three different states of the topological network structures. **(A)** I show an example of the fragmented sate which contains the disconnected components. **(B)** I show the critical state of the network topology, all components of which are fully connected with the minimum connections. **(C)** This state is also the fully connected network, but it contains excess connections.

defined as Equation (9). As shown in **Figure 7B**, the KS distance measured about 0.07 around its minimum value, which was a sufficiently small fitting. Moreover, the model fitting was validated by comparison to other distribution models (Takagi, 2017, 2018). In the case of node strength, the KS distance values of this model $6.6 \times 10^{-2}$ and of the normal distribution $8.5 \times 10^{-2}$ support this model, with its lower value. In addition, the correlation with mutual information is shown in **Figure 7C**, which indicates that the characteristic distribution of *ns* depends on this quantity, as is the case with larger component sizes (**Figure 5B**) and with energy (**Figures 4B,D**). Therefore, around the optimal state defined for efficiency and energy, the

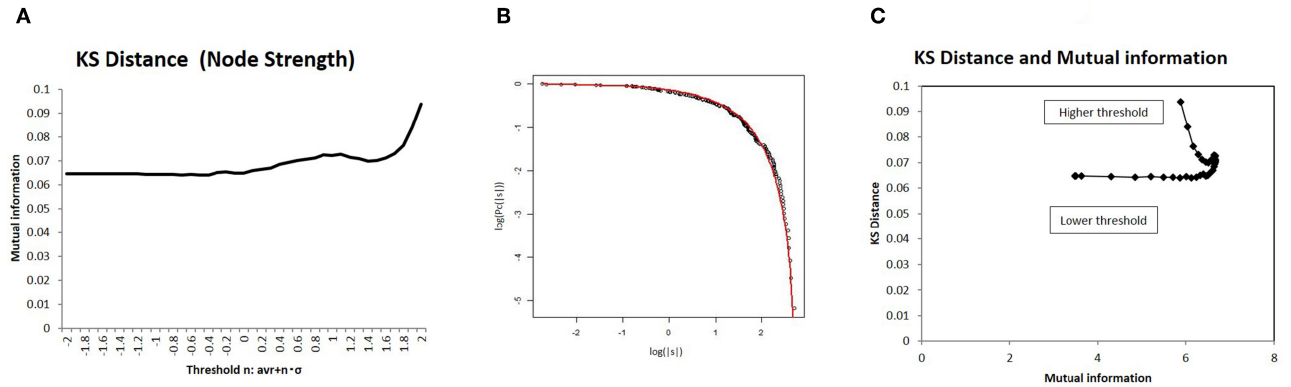


**FIGURE 7 |** I evaluated the distribution of node strength in terms of deviation from the distribution model using Kolmogorov-Smirnov (KS) distance. I showed a correlation with mutual information. **(A)** I also showed KS distance values from the distribution model for different threshold values. I assessed this value in terms of node strength, as well as the sum of the absolute values of connectivity strength for each node, using the distribution model. The threshold values were taken as the same as in the other cases in **Figure 2**. **(B)** Using an example, I compared empirical distribution and the distribution model. The distribution was taken as the threshold $< |w| >$ with $n = 0$, at which the average KS distance was evaluated as $6.64 \times 10^{-2}$. **(C)** I showed a correlation between mutual information and KS distance. On the x-axis, I took the mutual information calculated using whole datasets of matrices by applying the method, which is denoted by a solid line ($p = 0.05$) (**Figure 2**). I then plotted the corresponding points for each threshold value with KS distance on the y-axis, as given in **(A)**.

distribution of the node strength converges on this model, and the characteristic network structure emerges.

## 3.5. Activation Patterns and Overlapping

To analyze how this characteristic state regulates information transfer, I investigated the overlapping patterns of the response signals, applying the results of the information transfer model to Equation (2). As explained above, the repeatedly co-activated regions of different stimulation signals are related to cognitive processes (Haxby et al., 2001; Kumaran et al., 2016). For the set of response signals to random stimuli, the number of overlapping co-activated regions between different response signals was quantified in terms of Equation (7). The results for different network states are shown in **Figure 8A**, where the average of the overlapping numbers is taken for all combinations of responses. From lower thresholds to higher ones, the overlapping number decreased with decreasing excess connections. While it took large values at higher thresholds, it took the minimum value at the intermediate state.

The correlation to efficiency of information transfer is shown in the next panel (**Figure 8B**), in which the mutual information and the number of the overlapping sites has a strong correlation. As indicated by this graph, the overlapping number took the minimum value for maximum mutual information. On the other hand, the same number reduced network energy, as shown in **Figures 8C,D**, which show negative correlation. Therefore, the activation patterns evaluated in terms of the overlapping numbers are correlated strongly with the statistical quantities, network efficiency and energy.

The relation between the overlapping number and the network topological structures were also analyzed in **Figure 9**, which shows the direct relation to the small world topology. As described in the introduction, the small-world structure is considered as another relevant attribute of the brain network. The clustering coefficient was measured for each threshold value in **Figure 9A**. The correlation to the overlapping number was plotted on **Figure 9B**, in which a sharp peak around the minimum overlapping number indicated that the phase transition occurs around this point (with respect to the topological structure). The further evidence for the relation to small-world topology is given by the changes of this value. According to the observation in the Watts-Strogatz model (Watts and Strogatz, 1998), the clustering coefficient is stable near the state of the small-world topology, which is accompanied by the phase transition. The changes to the clustering coefficient $C$ were taken as the difference from the neighbor value, and were plotted in **Figure 9C** (Takagi, 2018). The correlation against the corresponding overlapping number is shown in **Figure 9D**. This result explicitly shows the dependency of the stability of the clustering coefficient and the phase transition of the topological structure. Thus, the minimization of the overlapping number can be correlated to the small-world topology.

## 3.6. Comparison to the Different Datasets

In order to verify the robustness of the above results, the simulation results based on other matrix datasets are presented. The first set is the sub-matrix, which is taken with randomly selected nodes from the original matrix of the functional connectome. The other set is the structural connectome, which is constructed using the physical connections of fiber tracts in the brain with the DTI method.

At first, the results with the sub-matrix were analyzed (**Figures 10A,B**). This simulation uses the connectivity matrices size in 100 nodes, which are selected randomly from the total 177 nodes in each original matrix. Comparison of **Figures 4D**, **10A** shows the relations between the wiring cost and the mutual information, and the minimization/maximization relations between these quantities are exhibited adequately in

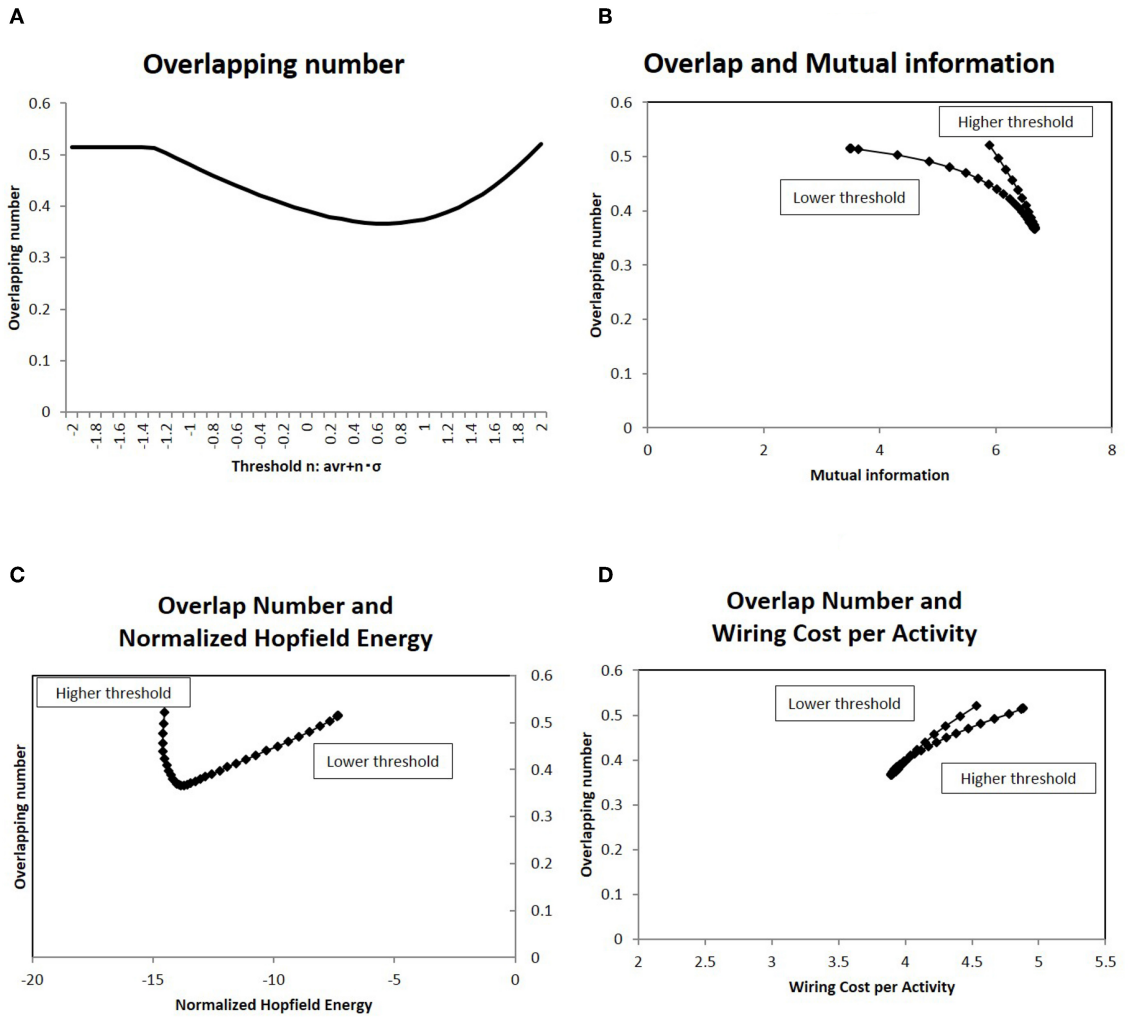


**FIGURE 8 |** I show the number of overlapping patterns in the response signals, as well as their correlation with the mutual information **(A)**. I showed the average of the number of overlapping co-activated regions defined as Equation (7). I took different network states by varying the threshold, as with the cases in the other figures. **(B)** I showed the correlation between the overlapping patterns given in **(A)** with the mutual information. On the x-axis, I took the mutual information calculated using the whole datasets of matrices by applying the method, which is denoted by a solid line ($p = 0.05$) (**Figure 2**). Then I plotted the corresponding points of each threshold value by overlapping numbers on the y-axis, which are given in **(A)**. **(C)** There was a correlation between the overlapping patterns given in **(A)** and the normalized Hopfield energy. On the x-axis, I measured the normalized Hopfield energy given in **Figure 5A**. I plotted the corresponding points of each threshold value with the network energy on the y-axis, as given in **(A)**. **(D)** A correlation was observed between the overlapping patterns given in **(A)** and the wiring cost performance given in **Figure 5C**, similar to **(C)**.

these panels. The other relation (**Figure 8D**) is also supported by **Figure 10B**, which displays the relation between the overlapping numbers and the wiring cost. These results with the sub-matrices show that important properties between the overlapping numbers and the wiring cost are stably obtained. The results suggest that these values are independent to other factors, such as the connectivity matrix size or the specific location of the brain regions taken as nodes.

The results with the structural connectome are displayed in **Figures 10C,D**. The results observed in the functional connectome can be confirmed with **Figures 10C,D**, where the simulation results exhibit similar properties to those given with the fMRI datasets (**Figures 4D**, **8D**), respectively. They also agree with the similarity between the functional and the structural connectome, in that the functional connectivity in the restring-state has close relation to the physical connections, such as the fiber tracts which organize the structural connectivity (Biswal et al., 2010). Thus, the robustness and the stability of the major properties obtained in this paper are given more strong evidence by the results of the structural network datasets. Because the structural network is comprised of the fiber tracts, the network structure is more stable compared to the functional connectivity based on the temporal dynamics correlations. In addition, the results obtained with the physical connections further clarify the

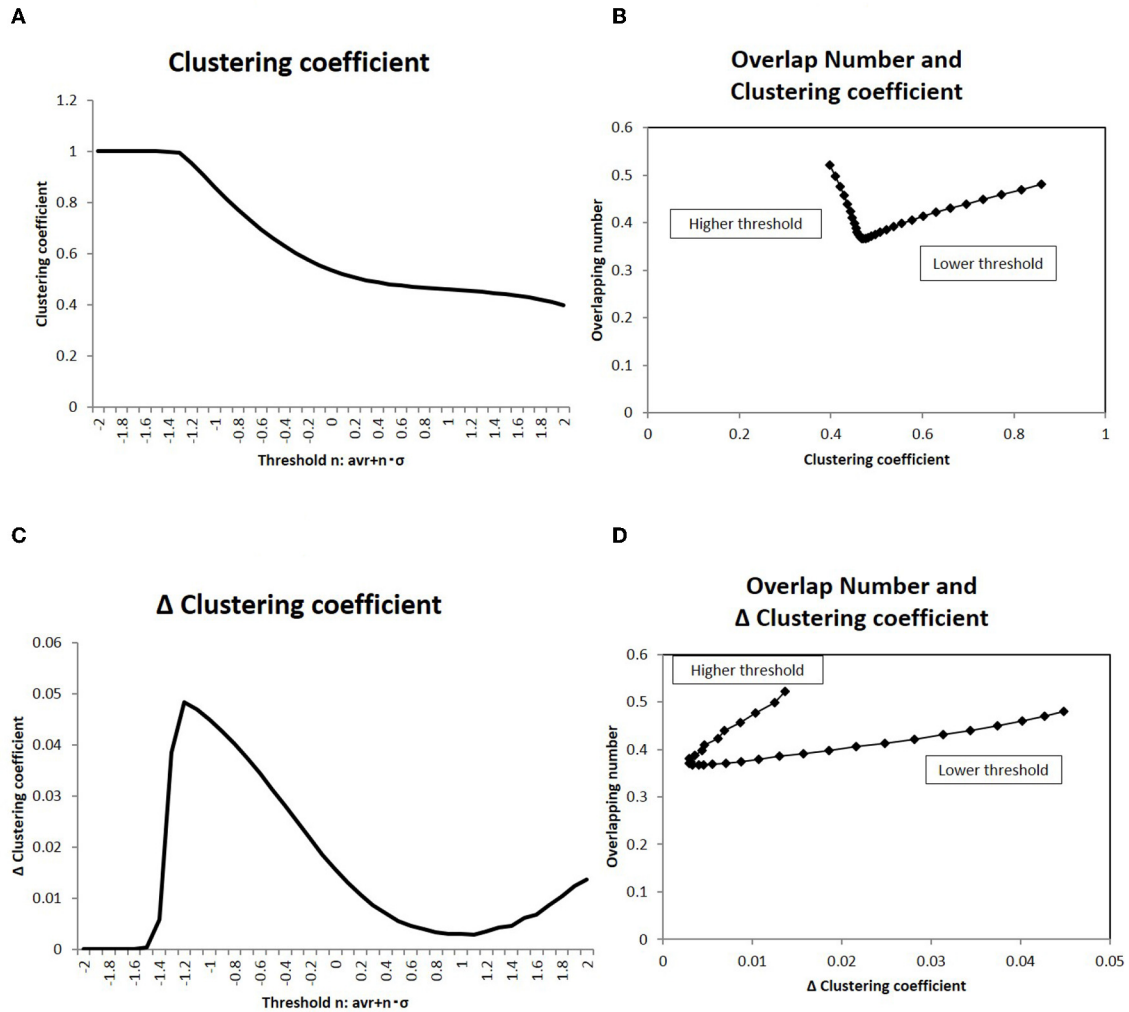


**FIGURE 9 |** Clustering coefficient and overlapping number. **(A)** The clustering coefficient $C$ for the topological description with the adjacency matrix is averaged and shown. The threshold values in the x-axis and the corresponding adjacency matrices are taken to be the same as those in **Figure 4**. The datasets of the matrix are also the same to those used in **Figure 4**. **(B)** The correlation of the clustering coefficient $C$ (shown in **A**) to the overlapping number is shown. The overlapping number is the same to those in **Figure 8A**. The threshold range in this panel is taken in $[-1.0, -2.0]$ so as to exclude the flat values in the lower thresholds. **(C)** The differences of the clustering coefficient in **(A)** is shown. The difference $\Delta C$ is calculated as $\Delta C = C(i) - C(i+1)$, where the difference is taken with the next value in the graph and $i$ is the number of the threshold position counted from the lower side. **(D)** The correlation of the clustering coefficient difference $\Delta C$ (shown in **C**) to the overlapping number is shown. The values of the overlapping numbers are the same as those in **Figure 8A**. The threshold range (which was the same as **C**) is taken.

meaning of the energy. In particular, wiring cost can be explicitly related to the real energy cost of the brain for network formation.

## 4. DISCUSSION

In the present paper, I modeled information transfer in the brain based on a dataset of the human functional connectome. As illustrated by **Figure 1A**, I represented brain activity using the activation patterns of multiple regions. That is, information processing was modeled in terms of the dynamics of successive patterns of activation. These dynamics were described in terms of the changing of activation states, as illustrated in **Figure 1B**, wherein positively or negatively activated states were transferred by activating or inactivating connected regions.

### 4.1. Information Transfer Model and Basic Statistical Quantities

In this simulation, I calculated the information transfer of randomly activated signals using Equation (2). Using this model, I evaluated the mutual information, defined as the average of Equation (3), and the network energy, defined as Equations (4–6). They are shown in **Figures 2**, **4A**, respectively. On the other hand, numerous empirical studies have suggested that information transfer in the brain is optimized, under constraints, such as the energy consumption, by maximizing

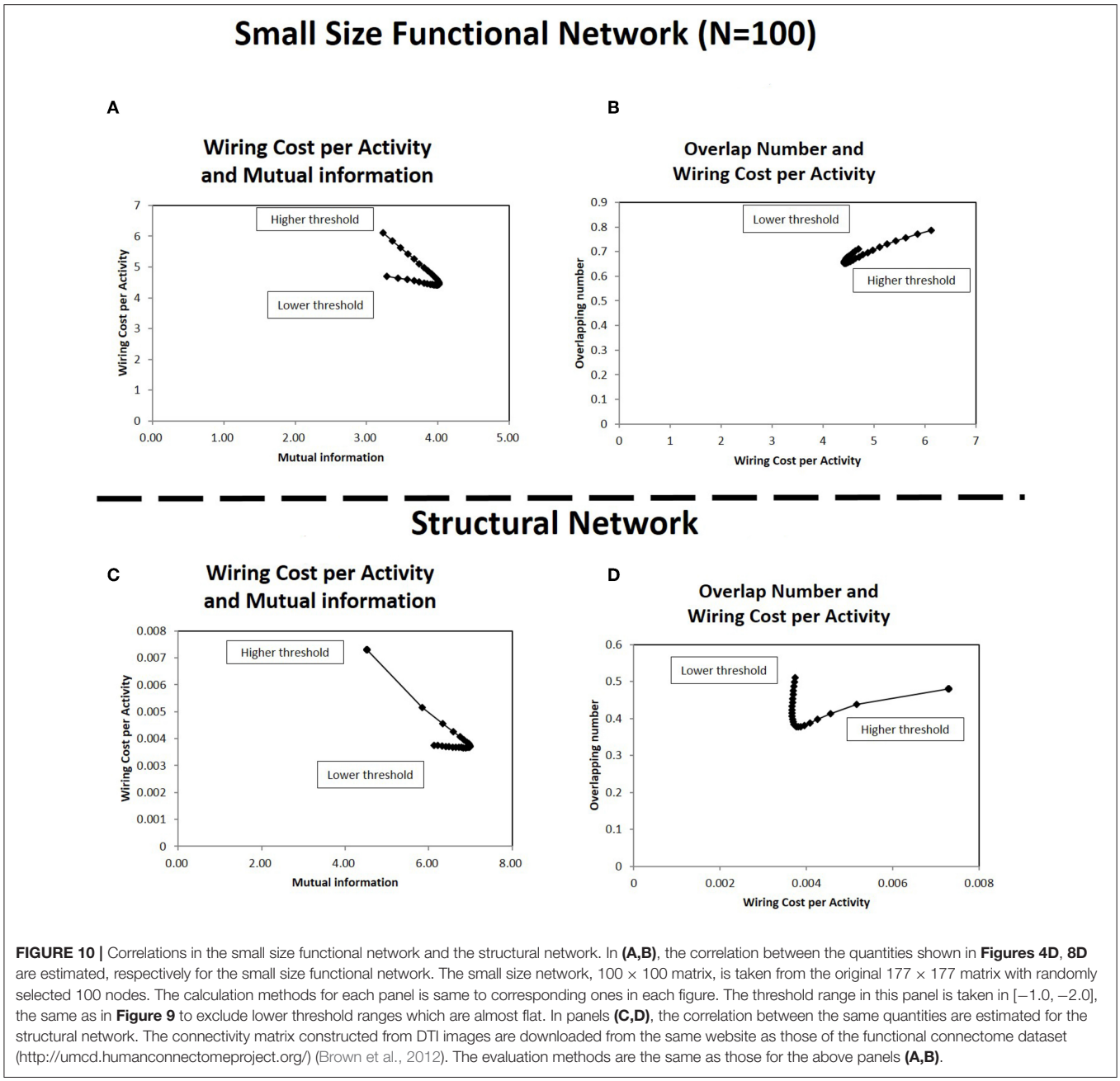


**FIGURE 10 |** Correlations in the small size functional network and the structural network. In **(A,B)**, the correlation between the quantities shown in **Figures 4D**, **8D** are estimated, respectively for the small size functional network. The small size network, 100 × 100 matrix, is taken from the original 177 × 177 matrix with randomly selected 100 nodes. The calculation methods for each panel is same to corresponding ones in each figure. The threshold range in this panel is taken in [−1.0, −2.0], the same as in **Figure 9** to exclude lower threshold ranges which are almost flat. In panels **(C,D)**, the correlation between the same quantities are estimated for the structural network. The connectivity matrix constructed from DTI images are downloaded from the same website as those of the functional connectome dataset (http://umcd.humanconnectomeproject.org/) (Brown et al., 2012). The evaluation methods are the same as those for the above panels **(A,B)**.

mutual information in the communication between brain regions (Linsker, 1990; Friston, 2010; Bullmore and Sporns, 2012). Therefore, I assessed the correlation between these two quantities. In these results given in **Figures 4B,D**, the energy is evaluated in terms of its cost performance, then the Hopfield energy normalized by the wiring cost and the wiring cost per total activity cost are shown, respectively and the decreasing of these quantities indicates the improvement of the cost performance. I showed these relationships in **Figures 4B,D**, in which I plotted the corresponding values of each network state. The figure indicated a negative correlation between the values, whereby increases in mutual information led to decreases in network energy, and vice versa. Thus, these two quantities must be correlated.

In particular, the peak around the maximum mutual information in **Figures 4B,D** shows that information transfer is optimized at this point by maximizing the quantity and minimizing the energy. According to the theory of the brain economy (Bullmore and Sporns, 2012), the efficiency of information processing in the brain is likely optimized by trading off with energy consumption. Although biological and empirical requirements regarding efficiency and energy are independent of each other, the result indicates that they are correlated, so there may be a mechanism that controls information transfer

while satisfying these two principles regarding the efficiency and the energy.

## 4.2. Network Structure and Information Transfer

Network analyses around this optimal state may explain the mechanism by which information transfer is organized in the brain. In **Figure 5A**, to allow topological representation, I estimated the largest component size of the network. The correlation with mutual information (**Figure 5B**) indicated that the efficiency of information transfer is maximized at the critical point between the fully connected network and the fragmented network state, which contains disconnected subcomponents. At this optimal state, the network maintains its fully connected structure with the minimum number of connections (**Figure 6**). This can be contrasted with the fragmented states, which inhibit efficient communication due to disconnections between regions. On the other hand, excess connections generate noise in the response. Therefore, in the intermediate phase at the optimal state, information transfer is cost effective, suppressing excess signals, and preserving fully connected structure.

As illustrated in **Figure 6**, this state can be described as the topological phase transition between the fully connected and fragmented phases. In this way, it constitutes a critical state. The distribution shape of node strength, another variable of the network, corroborates the notion that mutual information is maximized at the critical state. As shown in **Figure 7A**, the distribution of node strength converges in the model Equation (8), which assumes criticality and energy constraints (Takagi, 2017, 2018). This correlation shows that mutual information (**Figure 7C**) increases as the values converge upon the critical state. This result, along with the weighted network description (**Figure 7B**), also suggests that topological states are also correlated (**Figures 4B,D**). Both of these results indicate that the optimal state regarding the efficiency of the information transfer emerges in the critical state, suggesting that there is criticality in the brain, as has been confirmed empirically in various studies (Beggs and Plenz, 2003; Achard et al., 2006; Beggs, 2008; van den Heuvel and Sporns, 2011).

Although the state, which was specified as optimal, depends on the parameters, such as the threshold value, the criticality that supports its generality. Because the critical state was obtained without adjusting or fine-tuning multiple system parameters, it indicates that this state has the generality, which was obtained regardless of the details of the parameters. In fact, the stable results for the large samples about 1,000 individuals imply that these features around the optimal state are general ones, which emerge commonly and stably for different individuals.

This statistical features of node strength provide further information about the mechanism of the information transfer in this optimal state. The distribution of node strength exhibits a characteristic shape, as illustrated in **Figure 7B**. The cumulative distribution curve on the log-log plot indicates that the network contains a large number of higher strength nodes, which correspond to hubs in the functional network and comprise the core structure within networks (Hagmann et al., 2008; van den Heuvel and Sporns, 2011). Thanks to such core networks, whole networks can acquire the attributes of a small-world structure, allowing efficient communication with shortened distance between the nodes (Bassett and Bullmore, 2006) and improved robustness of information transfer.

## 4.3. Activation Patterns and Principles of Energy and Efficiency

The importance of these network states in regulating activity in the brain can be evaluated using activation patterns. According to the definition of the information transfer (Equation 2), the response signals for the random input stimuli might be determined, reflecting the network structure. For example, the response probabilities are determined by the combination of $w_{i,j} \neq 0$ elements for each $i$, and then the overlapping number would be given accordingly. Then, the overlapping number was an indicator, which reflects the network structure, activation patterns, and information transfer.

In **Figure 8A**, I evaluated the number of overlapping activated regions between different response signals. The correlation with efficiency of information transfer and energy are shown in **Figures 8B–D**, which show that network structure behaves in a similar way (**Figures 5B**, **7C**), indicating that these quantities depend strongly on the overlapping number. Increase in this quantity to the higher threshold was explained by the over-inactivated states with many 0 signals. The saturation of the activated signals, the higher density of the signals shown in **Figure 3C**, explains the same tendency, that is, increasing this quantity from the lower threshold. In each case, the overlapping number is increased at this state than during the intermediate states, at which activated and inactivated signals are balanced. Thus, the correlation between the mutual information entropy and the activation patterns can be explained by this quantity, the overlapping number.

As discussed above, increased overlapping may improve robustness in signal transfer and facilitate rapid response to the outer environment, with shortened communication distance between nodes. Despite these advantages, excess overlapping in the activation phase reduces the efficiency of the information transfer and causes the energy loss (**Figures 8B–D**). This implies that excess overlapping causes loss of efficiency and increases the energy consumption related to information transfer. Thus, these features have a trade-off relationship; that is, the robustness and the rapidity of responses are balanced with loss of efficiency and energy in information transfer.

In summary, the present results suggest that the principles of efficiency and energy consumption are important to information transfer. These principles affect multiple aspects of the functional network in the brain, and I have shown the connectivity strength (**Figure 7C**), activation patterns (**Figures 8B–D**), and topological network of such structures (**Figure 5B**). The same figures show the contribution of these principles to statistical quantities, in which sharp peaks indicate a strong tendency toward these quantities. Thus, these principles regarding

efficiency and of information transfer are important factors in regulating the characteristic attributes of the functional network in the human brain, such as network structure and activation patterns.

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://umcd.humanconnectomeproject.org/.

## AUTHOR CONTRIBUTIONS

KT designed the study, conducted the simulations and data analyses, and wrote the manuscript.

## ACKNOWLEDGMENTS

I would like to thank Editage (www.editage.jp) for English language editing.

## REFERENCES

Achard, S., and Bullmore, E. (2006). Efficiency and cost of economical brain functional networks. *PLoS Comp. Biol.* 3:e17. doi: 10.1371/journal.pcbi.0030017

Achard, S., Salvador, R., Whitcher, B., Suckling, J., and Bullmore, E. (2006). A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *J. Neurosci.* 26, 63–72. doi: 10.1523/JNEUROSCI.3874-05.2006

Azevedo, F. A., Carvalho, L. R., Grinberg, L. T., Farfel, J. M., Ferretti, R. E., Leite, R. E., et al. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* 513:532. doi: 10.1002/cne.21974

Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. (2004). The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U.S.A.* 101, 3747–3752. doi: 10.1073/pnas.0400087101

Bassett, D., and Bullmore, E. (2006). Small-world brain networks. *Neuroscientist* 12, 512–523. doi: 10.1177/1073858406293182

Bassett, D., Meyer-Lindenberg, A., Achard, S., Duke, T., and Bullmore, E. (2006). Adaptive reconfiguration of fractal small-world human brain functional networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 19518–19523. doi: 10.1073/pnas.0606005103

Bassett, D. S., Wymbs, N. F., Porter, M. A., Mucha, P. J., Carlson, J. M., and Grafton, S. T. (2018). Dynamic reconfiguration of human brain networks during learning. *Proc. Natl. Acad. Sci. U.S.A.* 118, 7641–7646. doi: 10.1073/pnas.1018985108

Beggs, J. (2008). The criticality hypothesis: how local cortical networks might optimize information processing. *Philos. Trans. A Math. Phys. Eng. Sci.* 366, 329–343. doi: 10.1098/rsta.2007.2092

Beggs, J., and Plenz, D. (2003). Neuronal avalanches in neocortical circuits. *J. Neurosci.* 23, 11167–11177. doi: 10.1523/JNEUROSCI.23-35-11167.2003

Biswal, B., Mennes, M., Zuo, X.-N., Gohel, S., Kelly, C., Smith, S. M., et al. (2010). Toward discovery science of human brain function. *Proc. Natl. Acad. Sci. U.S.A.* 107, 4734–4739. doi: 10.1073/pnas.0911855107

Bressler, S., and Menon, V. (2010). Large-scale brain networks in cognition: emerging methods and principles. *Trends Cogn. Sci.* 14, 277–290. doi: 10.1016/j.tics.2010.04.004

Brown, J., Rudie, J., Bandrowski, A., Van Horn, J., and Bookheimer, S. (2012). The UCLA multimodal connectivity database: a web-based platform for brain connectivity matrix sharing and analysis. *Front. Neuroinform.* 6:28. doi: 10.3389/fninf.2012.00028

Bullmore, E., and Sporns, O. (2012). The economy of brain network organization. *Nat. Rev. Neurosci.* 13, 336–349. doi: 10.1038/nrn3214

Calhoun, V., Liu, J., and Adali, T. (2009). A review of group ica for fMRI data and ica for joint inference of imaging, genetic, and ERP data. *Neuroimage* 45:S163. doi: 10.1016/j.neuroimage.2008.10.057

Chialvo, D. (2010). Emergent complex neural dynamics. *Nat. Phys.* 6, 744–750. doi: 10.1038/nphys1803

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* 36:181–204. doi: 10.1017/S0140525X12000477

Clauset, A., Shalizi, C., and Newman, M. (2009). Power-law distributions in empirical data. *SIAM Rev.* 51, 661–703. doi: 10.1137/070710111

Eguiluz, V., Chialvo, D., Cecchi, G., Baliki, M., and Apkarian, A. (2005). Scale-free brain functional networks. *Phys. Rev. Lett.* 94:018102. doi: 10.1103/PhysRevLett.94.018102

Fenno, L., Yizhar, O., and Deisseroth, K. (2011). The development and application of optogenetics. *Annu. Rev. Neurosci.* 34:389. doi: 10.1146/annurev-neuro-061010-113817

Finn, E., Shen, X., Scheinost, D., Rosenberg, M., Huang, J., Chun, M. M., et al. (2015). Functional connectome fingerprinting: identifying individuals based on patterns of brain connectivity. *Nat. Neurosci.* 18:1664. doi: 10.1038/nn.4135

Fox, M., and Raichle, M. (2007). Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* 8, 700–711. doi: 10.1038/nrn2201

Fox, M., Snyder, A., Vincent, J., Corbetta, M., Van Essen, D., and Raichle, M. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl. Acad. Sci. U.S.A.* 102, 9673–9678. doi: 10.1073/pnas.0504136102

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* 11, 127–138. doi: 10.1038/nrn2787

Ghazanfar, A., and Schroeder, C. (2006). Is neocortex essentially multisensory? *Trends Cogn. Sci.* 10, 278–285. doi: 10.1016/j.tics.2006.04.008

Greicius, M., Supekar, K., Menon, V., and Dougherty, R. (2009). Restingstate functional connectivity reflects structural connectivity in the default mode network. *Cereb. Cortex* 19, 72–78. doi: 10.1093/cercor/bhn059

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C., Wedeen, V. J., et al. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biol.* 6:e159. doi: 10.1371/journal.pbio.0060159

Haxby, J., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425–2430. doi: 10.1126/science.1063736

Hilgetag, C., and Grant, S. (2000). Uniformity, specificity and variability of corticocortical connectivity. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 355, 7–20. doi: 10.1098/rstb.2000.0546

Hinton, G. E., and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science* 313, 504–507. doi: 10.1126/science.1127647

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proc. Natl. Acad. Sci. U.S.A.* 81, 3088–3092. doi: 10.1073/pnas.81.10.3088

Kanwisher, N., McDermott, J., and Chun, M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17, 4302–4311. doi: 10.1523/JNEUROSCI.17-11-04302.1997

Kitzbichler, M. G. Smith, M., Christensen, S., and Bullmore, E. (2009). Broadband criticality of human brain network synchronization. *PLoS Comput. Biol.* 5:e1000314. doi: 10.1371/journal.pcbi.1000314

Klaus, A., Yu, S., and Plenz, D. (2011). Statistical analyses support power law distributions found in neuronal avalanches. *PLoS ONE* 6:e19779. doi: 10.1371/journal.pone.0019779

Kumaran, D., Hassabis, D., and McClelland, J. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends Cogn. Sci.* 20, 512–534. doi: 10.1016/j.tics.2016.05.004

Lee, J. H., Durand, R., Gradinaru, V., Zhang, F., Goshen, I., Kim, D.-S., et al. (2010). Global and local fmri signals driven by neurons defined optogenetically by type and wiring. *Nature* 465:788. doi: 10.1038/nature09108

Linsker, R. (1990). Perceptual neural organisation: some approaches based on network models and information theory. *Annu. Rev. Neurosci.* 13, 257–281. doi: 10.1146/annurev.ne.13.030190.001353

Meunier, D., Lambiotte, R., and Bullmore, E. (2010). Modular and hierarchically modular organization of brain networks. *Front. Neurosci.* 4:200. doi: 10.3389/fnins.2010.00200

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A., Veness, J., Bellemare, M., et al. (2015). Human-level control through deep reinforcement learning. *Nature* 518, 529–533. doi: 10.1038/nature14236

Niven, J., and Laughlin, S. (2008). Energy limitation as a selective pressure on the evolution of sensory systems. *J. Exp. Biol.* 211, 1792–1804. doi: 10.1242/jeb.017574

Park, H., and Friston, K. (2013). Structural and functional brain networks: from connections to cognition. *Science* 342:1238411. doi: 10.1126/science.1238411

Rubinov, M., and Sporns, O. (2011). Weight-conserving characterization of complex functional brain networks. *Neuroimage* 56, 2068–2079. doi: 10.1016/j.neuroimage.2011.03.069

Shmuel, A., Augath, M., Oeltermann, A., and Logothetis, N. (2006). Negative functional mri response correlates with decreases in neuronal activity in monkey visual area v1. *Nat. Neurosci.* 9, 569–577. doi: 10.1038/nn1675

Smith, S. (2012). The future of fMRI connectivity. *Neuroimage* 62:1257. doi: 10.1016/j.neuroimage.2012.01.022

Smith, S., Miller, K., Salimi-Khorshidi, G., and Webster, M. (2011). Network modelling methods for fMRI. *Neuroimage* 54:875. doi: 10.1016/j.neuroimage.2010.08.063

Song, B., Ma, N., Liu, G., Zhang, H., Yu, L., Liu, L., et al. (2019). Maximal flexibility in dynamic functional connectivity with critical dynamics revealed by fMRI data analysis and brain network modelling. *J. Neural Eng.* 16:056002. doi: 10.1088/1741-2552/ab20bc

Sporns, O. (2002). Network analysis, complexity, and brain function. *Complexity* 8, 56–60. doi: 10.1002/cplx.10047

Sporns, O. (2013). Network attributes for segregation and integration in the human brain. *Curr. Opin. Neurobiol.* 23, 162–171. doi: 10.1016/j.conb.2012.11.015

Sporns, O., Tononi, G., and Kotter, R. (2005). The human connectome: a structural description of the human brain. *PLoS Comput. Biol.* 1:e42. doi: 10.1371/journal.pcbi.0010042

Tagliazucchi, E., Balenzuela, P., Fraiman, D., and Chialvo, D. (2012). Criticality in large-scale brain fMRI dynamics unveiled by a novel point process analysis. *Front. Physiol.* 3:15. doi: 10.3389/fphys.2012.00015

Tagliazucchi, E., Chialvo, D., Siniatchkin, M., Amico, E., Brichant, J.-F., Bonhomme, V., et al. (2016). Large-scale signatures of unconsciousness are consistent with a departure from critical dynamics. *J. R. Soc. Interface* 13:20151027. doi: 10.1098/rsif.2015.1027

Takagi, K. (2017). A distribution model of functional connectome based on criticality and energy constraints. *PLoS ONE* 12:e0177446. doi: 10.1371/journal.pone.0177446

Takagi, K. (2018). Information-based principle induces small-world topology and self-organized criticality in a large scale brain network. *Front. Comp. Neurosci.* 12:65. doi: 10.3389/fncom.2018.00065

Tomasi, D., Wang, G.-J., and Volkow, N. (2013). Energetic cost of brain functional connectivity. *Proc. Natl. Acad. Sci. U.S.A.* 110, 13642–13647. doi: 10.1073/pnas.1303346110

Tononi, G., Sporns, O., and Edelman, G. (1994). A measure for brain complexity: relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. U.S.A.* 91:5033. doi: 10.1073/pnas.91.11.5033

van den Heuvel, M., and Sporns, O. (2011). Rich-club organization of the human connectome. *J. Neurosci.* 31, 15775–15786. doi: 10.1523/JNEUROSCI.3539-11.2011

van den Heuvel, M., Stam, C., Boersma, M., and HulshoffPol, H. (2008). Small world and scale-free organization of voxel based resting-state functional connectivity in the human brain. *Neuroimage* 43, 528–539. doi: 10.1016/j.neuroimage.2008.08.010

Van Dijk, K., Hedden, T., Venkataraman, A., Evans, K., and Lazar, S. (2010). Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. *J. Neurophysiol.* 103, 297–321. doi: 10.1152/jn.00783.2009

Wang, R., Tsuda, I., and Zhang, Z. (2015). A new work mechanism on neuronal activity. *Int. J. Neural Syst.* 25:1450037. doi: 10.1142/S0129065714500373

Wang, R., Zhang, Z., and Chen, G. (2008). Energy function and energy evolution on neural population. *IEEE Trans. Neural Netw.* 19, 535–538. doi: 10.1109/TNN.2007.914177

Wang, Y., Xu, X., and Wang, R. (2018). An energy model of place cell network in three dimensional space. *Front. Neurosci.* 12:264. doi: 10.3389/fnins.2018.00264

Wang, Z., and Wang, R. (2014). Energy distribution property and energy coding of a structural neural network. *Front. Comp. Neurosci.* 8:14. doi: 10.3389/fncom.2014.00014

Watts, D., and Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442. doi: 10.1038/30918

Whitacre, J. M. (2010). Degeneracy: a link between evolvability, robustness and complexity in biological systems. *Theor. Biol. Med. Model.* 7:6. doi: 10.1186/1742-4682-7-6

Yu, L., and Yu, Y. (2017). Energy–efficient neural information processing in individual neurons and neuronal networks. *J. Neurosci. Res.* 95, 2253–2266. doi: 10.1002/jnr.24131

Zuo, X.-N., Ehmke, R., Mennes, M., Imperati, D., Castellanos, F., Sporns, O., et al. (2012). Network centrality in the human functional connectome. *Cereb. Cortex* 22, 1862–1875. doi: 10.1093/cercor/bhr269

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

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 10 January 2020
doi: 10.3389/fncom.2019.00085

# Measuring the Non-linear Directed Information Flow in Schizophrenia by Multivariate Transfer Entropy

*Dennis Joe Harmah[1,2], Cunbo Li[1,2], Fali Li[1,2], Yuanyuan Liao[1,2], Jiuju Wang[3], Walid M. A. Ayedh[1,2], Joyce Chelangat Bore[1,2], Dezhong Yao[1,2], Wentian Dong[3]* and Peng Xu[1,2]**

[1] The Clinical Hospital of Chengdu Brain Science Institute, MOE Key Lab for Neuroinformation, University of Electronic Science and Technology of China, Chengdu, China, [2] School of Life Science and Technology, Center for Information in Medicine, University of Electronic Science and Technology of China, Chengdu, China, [3] Institute of Mental Health, Peking University Sixth Hospital, National Clinical Research Center for Mental Disorders & Key Laboratory of Mental Health, Ministry of Health, Peking University, Beijing, China

People living with schizophrenia (SCZ) experience severe brain network deterioration. The brain is constantly fizzling with non-linear causal activities measured by electroencephalogram (EEG) and despite the variety of effective connectivity methods, only few approaches can quantify the direct non-linear causal interactions. To circumvent this problem, we are motivated to quantitatively measure the effective connectivity by multivariate transfer entropy (MTE) which has been demonstrated to be able to capture both linear and non-linear causal relationships effectively. In this work, we propose to construct the EEG effective network by MTE and further compare its performance with the Granger causal analysis (GCA) and Bivariate transfer entropy (BVTE). The simulation results quantitatively show that MTE outperformed GCA and BVTE under varied signal-to-noise conditions, edges recovered, sensitivity, and specificity. Moreover, its applications to the P300 task EEG of healthy controls (HC) and SCZ patients further clearly show the deteriorated network interactions of SCZ, compared to that of the HC. The MTE provides a novel tool to potentially deepen our knowledge of the brain network deterioration of the SCZ.

**Keywords: network deterioration, schizophrenia, non-linear causal interaction, multivariate transfer entropy, granger causality, bivariate transfer entropy**

OPEN ACCESS

***Edited by:***
*Abdelmalik Moujahid, University of the Basque Country, Spain*

***Reviewed by:***
*Duan Li, University of Michigan, United States*
*Junfeng Sun, Shanghai Jiao Tong University, China*

***Correspondence:**
*Peng Xu*
*xupeng@uestc.edu.cn*
*Wentian Dong*
*dongwentian@bjmu.edu.cn*

***Received:*** *15 August 2019*
***Accepted:*** *04 December 2019*
***Published:*** *10 January 2020*

***Citation:***
*Harmah DJ, Li C, Li F, Liao Y, Wang J, Ayedh WMA, Bore JC, Yao D, Dong W and Xu P (2020) Measuring the Non-linear Directed Information Flow in Schizophrenia by Multivariate Transfer Entropy. Front. Comput. Neurosci. 13:85. doi: 10.3389/fncom.2019.00085*

## INTRODUCTION

The brain usually fizzles with the non-linear causal activity of electroencephalogram (EEG) at a microscopic level (Gourévitch et al., 2006; Sabesan et al., 2010; Mehta and Kliewer, 2018). The complex nature of the brain makes its non-linear causal dynamics unknown, and how the brain matches its rhythm as well as its metabolic processes and a causal relationship is still under investigation. The brain might be attacked with many psychosomatic diseases such as schizophrenia (SCZ), leading to deteriorated brain network, which eventually affects its cognitive functions (Shovon et al., 2017; Li et al., 2018). Researchers have explored the EEG non-linearity in multiple psychiatric disorders, for example, in epileptic patients probably due to low dimensional chaos during a seizure (Lee et al., 2001; Henderson et al., 2011; Liu et al., 2017). Thus, the behavioral and psychological attitudes of people with psychiatric disorders call for the need to effectively investigate the transient information exchange in the brain (Zhang et al., 2011; Mehta and Kliewer, 2016). Multiple techniques or measures for linear and non-linear brain connectivity such as structural,

functional, and effective connectivity are in use for this purpose (Selskii et al., 2017; Hristopulos et al., 2019). Exploring the linear and non-linear interactions, more importantly when the system structure is unknown, holds promise for deepening the knowledge of the causal mechanism in the brain for the SCZ (Pereda et al., 2005; Zhao et al., 2013).

SCZ is the most prevalent functional psychotic disorder, and people living with the disorder can present with a variety of symptoms and manifestations that can be seen in their behaviors. The disease is a chronic psychotic disorder that disrupts the patient's thoughts and affect their total well-being (Patel et al., 2014; Ure et al., 2018). Previous studies have demonstrated a coherent or uniform reduction in the brain regions of the SCZ patients, including the insula, superior temporal gyrus, amygdala, parahippocampus, inferior and medial frontal gyri, hippocampus, and anterior cingulate cortex (ACC) (Ehrlich et al., 2014; Alonso-Solís et al., 2015; Domínguez-Iturza et al., 2018). In neurophysiological research, it is more interesting to explore the specific performance of the SCZ under certain task like oddball paradigm involving the P300 (Alvarado-González et al., 2016), as the P300 serves as the reliable biomarker to identify the SCZ from healthy control (HC) (Somani and Shukla, 2012). For example, during working memory, the P300 amplitude decreases with increasing the load for HC but remains low in all conditions for the SCZ (Gaspar et al., 2011). Besides the P300 amplitude, the occurrence of the SCZ is also accompanied by the abnormal task brain network (Krusienski et al., 2006; Pérez-Vidal et al., 2018). For example, we have previously found a crucial role of the ACC in regulating the P300 (Li et al., 2018), especially a compensatory pathway from the dorsolateral prefrontal cortex to intraparietal sulcus for the SCZ.

Effective connectivity in the brain brings in the element of causal interactions or causation. Consequently, a signal activation in one area of the brain directly causes a change or signal, activation or depression, in another area (Mastrovito et al., 2018; Zhu et al., 2018). Effective connectivity in a domain of data-driven approaches such as Granger causality analysis (GCA) which performs poorly in non-linear context rely on its past to formulate linear causal interactions in the EEG signal (Venkatesh and Grover, 2016; Li et al., 2017). The GCA is initially formulated for linear models and later extended to non-linear systems by applying to local linear models. Despite its success in detecting the direction of interactions in the brain, it either makes assumptions about the structure of the interacting systems or the nature of their interactions and as such, it may suffer from the shortcomings of modeling systems/signals of unknown structure (Lainscsek et al., 2013; Sohrabpour et al., 2016; Bonmati, 2018). Even though much has been achieved with the GCA, a different data-driven approach which involves information theoretic measures like Transfer entropy (TE) may play a critical role in elucidating the effective connectivity of non-linear complex systems that the GCA may fail to unearth (Schreiber, 2006; Madulara et al., 2012; Dejman et al., 2017). Mathematically, the TE uses its entropy to quantitatively infer the coupling strength between two variables (Liu and Aviyente, 2012; Shovon et al., 2017) and has the potential for capturing both the linear and non-linear causal interactions effectively. Thus, TE works in bivariate fashion where information transfer is quantified between all source-target pairs but bivariate analysis has spurious, redundant and synergistic interaction problems (James et al., 2016; Wollstadt et al., 2019).

To quantify the effective connectivity and exploring the corresponding network aberration in the SCZ, the reliable estimation of the brain network seems to be of great urgency. In this work, we used the TE in a multivariate fashion (Lainscsek et al., 2013; Alonso-Solís et al., 2015; Bonmati, 2018), i.e., multiple TE (MTE) (Montalto et al., 2014; Novelli et al., 2019; Wollstadt et al., 2019). The MTE has great ability to handle problems that the GCA and the BVTE cannot, such as spurious or redundant interactions, where multiple sources provide the same information about the target, the MTE also cannot miss synergistic interactions between multiple relevant sources and the target, where these multiple sources jointly transfer more information into the target than what could be detected from examining source contributions individually. The MTE is designed to remove redundancies and capture synergistic interactions and account for all relevant sources of a target, unearth both the linear and non-linear dynamics in the brain; thus making it a powerful tool over GCA and BVTE (Stokes et al., 2018; Wollstadt et al., 2019). Herein, we first proposed to infer the linear and non-linear simulations of the GCA, BVTE, and MTE under various conditions, including varied signal-to-noise(SNR) conditions, edges recovered, sensitivity, and specificity, to explore their performances; thereafter, we also applied both methods to P300 task EEG of the SCZ and HC to investigate the brain network deterioration for the SCZ.

## TRANSFER ENTROPY

If a signal $X$ directly interacts with signal $Y$, then the past information of $X$ should possess ample information that can help predict $Y$ beyond the information possessed in the history of $Y$ only. That is, there is a Granger-causal interaction from $X$ to $Y$ (Sørensen and Causality, 2005). The GCA paves a way for the examination of the directed interaction between variables. In essence, GCA is designed to measure the linear

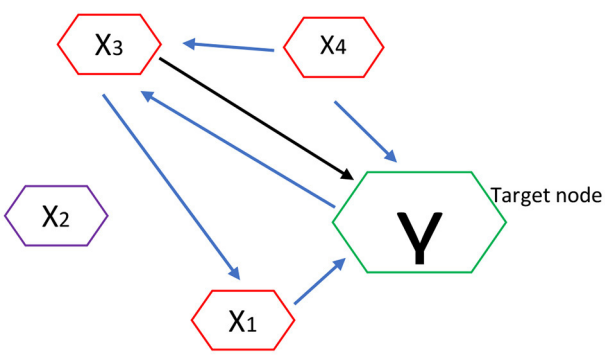


**FIGURE 1 |** Estimation of MTE into a target node Y. Blue arrows show the estimation of MTE into a target node.

coupling among time series, which determines that the GCA can only capture the linear causality well, and may not work for the non-linear cases (Bose et al., 2017). In addition, the neural coupling in the brain is far from the linearity, and the conventional GCA may not capture this hidden coupling in the brain.

To capture the non-linear interactions in the brain, we alternatively used the TE to measure the directed information exchange.

Let $X = \{x_1, x_2..., x_T\}$ and $Y = \{y_1, y_2, ..., y_T\}$ denote the time series of two brain areas with T observations, we define an entropy rate which is the amount of additional information required to represent the value of the next observation of $X$ as:

$$h_1 = -\sum_{x_n+1,x_n,y_n} p(x_n+1, x_n, y_n)\log_2 p(x_n+1|x_n, y_n) \quad (1)$$

Also, we define another entropy rate assuming that $x_n + 1$ as:

$$h_2 = -\sum_{x_n+1,x_n,y_n} p(x_n+1, x_n, y_n)\log_2 p(x_n+1|x_n) \quad (2)$$

Therefore, the TE from $Y$ to $X$ is given by $h_2 - h_1$, and this corresponds to information transfer from $Y$ to $X$:

$$TE_{Y\to X} = h_2 - h_1, = \sum_{x_n+1,x_n,y_n} p(x_n+1, x_n, y_n)\log_2\left(\frac{p(x_n+1|x_n, y_n)}{p(x_n+1|x_n)}\right) \quad (3)$$

Similarly, we can define the transfer entropy from $X$ to $Y$ as:

$$TE_{X\to Y} = \sum_{y_n+1,x_n,y_n} p(y_n+1, x_n, y_n)\log_2\left(\frac{p(y_n+1|x_n, y_n)}{p(y_n+1|y_n)}\right) \quad (4)$$

Then, we compute the TE by writing (3) and (4) using conditional probabilities as:

$$TE_{Y\to X} = \sum_{x_n+1,x_n,y_n} p(x_n+1, x_n, y_n)\log_2\left(\frac{p(x_n+1, x_n, y_n)p(x_n)}{p(x_n, y_n)p(x_n+1, x_n)}\right) \quad (5)$$

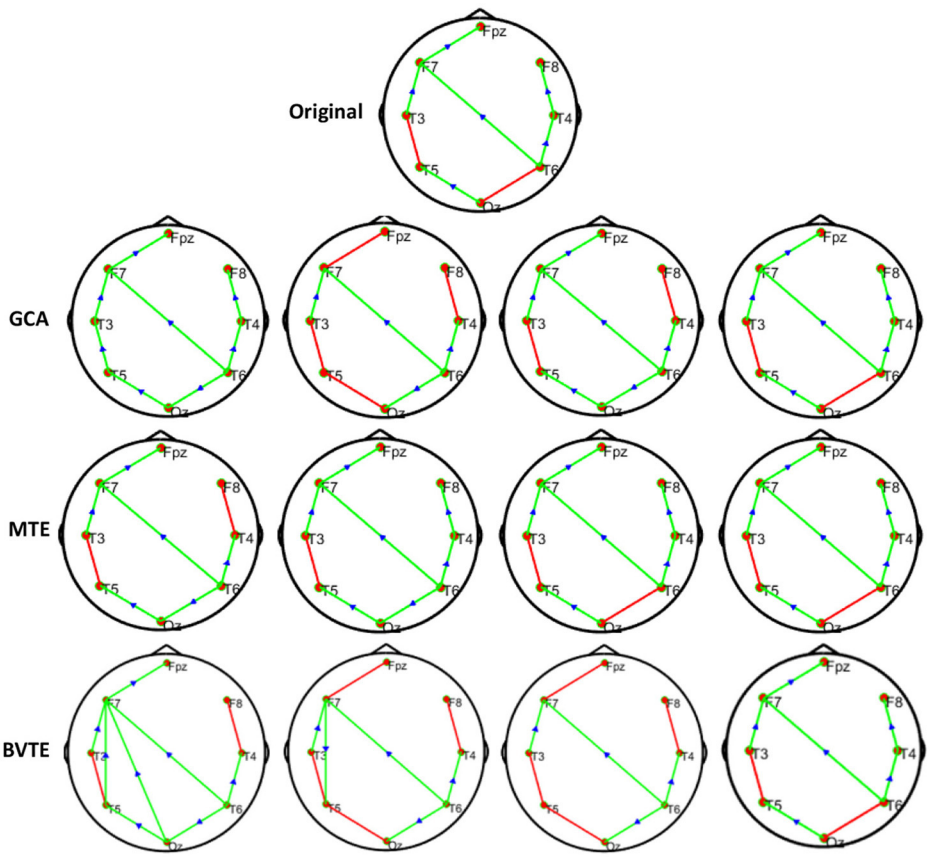


**FIGURE 2 |** Original or predefined 8 nodes simulated network and estimated linear networks by GCA, MTE, and BVTE with $Y = A \times B$.

$$TE_{X\to Y} = \sum_{y_n+1,x_n,y_n} p(y_n+1,x_n,y_n)\log_2 \left(\frac{p(y_n+1,x_n,y_n)P(y_n)}{p(x_n,y_n)p(y_n+1,y_n)}\right) \quad (6)$$

Where $x_n$, and $y_n$, are the stochastic variables obtained by sampling the processes at the present time $n$ (Gilmour et al., 2012; Wollstadt et al., 2014; Shao et al., 2015).

TE estimator can detect both linear and non-linear causality. However, because of the bivariate nature of TE, its outcome may infer spurious or redundant causality and may also miss synergistic interactions between multiple relevant sources and the target (Wollstadt et al., 2019). Hence, we need to have a tool or method that can accommodate these challenges. MTE has proven to be a better option to measure both the linear and inherent non-linear brain signals and their causal relationships effectively. Importantly, the MTE is an extension of the TE, which is a direct measure of information transfer between a source and a target process in a dynamic or composite system. Unlike TE, however, MTE does not give spurious, redundant information and also may not miss synergistic interactions (Montalto et al., 2014; James et al., 2016; Wollstadt et al., 2019).

Let at a given instance the dynamic system be composed of a source system $X$, a destination system $Y$ and remaining systems $Z = \left\{Z^k\right\}_{k=1,\ldots M-2}$. Here, we are interested in evaluating the information flow from a source system $X$ to a destination system $Y$. Then, MTE models the information flow from the source system to the destination system in the presence of the remaining systems, as shown in Equation (7).

$$TE_{X\to Y|Z} = \sum p(y_{1:n},x_{1:n-1},z_{1:n-1}) \log \frac{p(y_n|x_{1:n-1},y_{1:n-1},z_{1:n-1})}{p(y_n|y_{1:n-1},z_{1:n-1})} \quad (7)$$

Where $x$, $y$, and $z$ are the state visited by the systems $X$, $Y$, and $Z$ over time. Let $x_n$, $y_n$, and $z_n$ be the stochastic variables obtained by sampling the processes at the present time $n$. Furthermore, we denote $x_{tn}$ as the vector variable describing all the states visited by $X$ from time $t$ up to $n$ (assuming $n$ as the present time and setting the origin of time at $t = 1$, $x_{1:n-1}$ represents the whole past history of the process $x$).

In our case, the dynamic system is composed of the brain regions, Frontal (F), Parietal (P), Temporal (T), and Occipital (O) lobes. In other words, the source system $X$ and the destination system $Y$ are the brain regions involved in a given information flow, e.g., it could be F and P or T and O. The information flow between any two brain regions is also affected by the states of remaining brain regions, which are not part of the information flow (Wang et al., 2011; Adhikari and Agrawal, 2013; Anil et al., 2015). Hence, MTE is a good estimator to measure the linear and non-linear directed information flow in the brain.

For an illustration, let's demonstrate MTE brain network algorithm analysis as shown in **Figure 1**. Here the nodes or channels represent (stochastic) processes and the arrows represent causal connections or interactions between processes. It has target of interest and relevant sources.

Thus if $Y$ is the current target of interest, then nodes highlighted in red represent the set of relevant sources $Z = \{X_1, X_3, X_4\}$, i.e., the sources that contribute to the target's current value $Y_n$. In order to estimate the MTE into the target $Y$, it requires inferring the set Z containing the relevant sources (or parents) of $Y$. Once $Z$ is inferred, we compute the MTE from a single process into the target as a conditional transfer entropy, which accounts for the potential effects of the remaining relevant sources. Formally, the MTE from a single source (e.g., $X_3$) into $Y$ is defined as the TE from $X_3$ to $Y$, conditioned on $Z$ and excluding $X_3$: $\mathrm{TE}(X_3 \to Y|Z\backslash X_3)$ as shown in **Figure 1** (Srivastava, 2002; Flecker et al., 2011; Wollstadt et al., 2019).

# VALIDATION ANALYSIS

## Simulation Study

### Simulated Network

We generated and simulated a random time series with 7 and 8 nodes/process and 500 observations (**Figures 2**, **5**). A network structure with unidirectional and bidirectional couplings and nodes with input and output degrees or domain were considered. Two network structures were simulated, i.e., linear and non-linear. Out of the linear equation, we modeled the non-linear networks by adding five different types of non-linear functions to the linear equation (Khadem and Hossein-Zadeh, 2014; Dong et al., 2015; Li et al., 2017). When estimating the MTE and the BVTE, we used the toolbox IDTx (Wollstadt et al., 2019) and GCCA-toolbox for GCA, to estimate the parameters of the MVAR models and the Akaike Information Criterion (AIC) for model order selection (Sohrabpour et al., 2016). We applied the conventional multivariate Granger Analysis for our computation and analysis for GCA. The performance of the GCA, BVTE, and MTE are statistically tested under multiple strategies including the effective connectivity, edges recovered, sensitivity, and specificity on the 8 nodes time series.

To see which method performs better by suppressing the turbulent noise condition, we added Gaussian noise (Ozaki, 2012) with a varying SNR in a range of −10, −5, 5, and 10 dB to the generated time series. With different realizations of the

**TABLE 1 |** Causal interactions parameters and explanation.

| Parameter | Description of parameter |
|---|---|
| TN | TN denotes the number of direct interactions that were not available and were truly marked as non-existent. |
| TP | TP describes causal interactions that were available and truly labeled as existent. |
| FN | FN denotes the number of causal interactions that were incorrectly marked as not existing. |
| FP | FP denotes the number of directed interactions that were incorrectly marked as existing or indicates the number of pairs that were identified to have false causal relationships. |

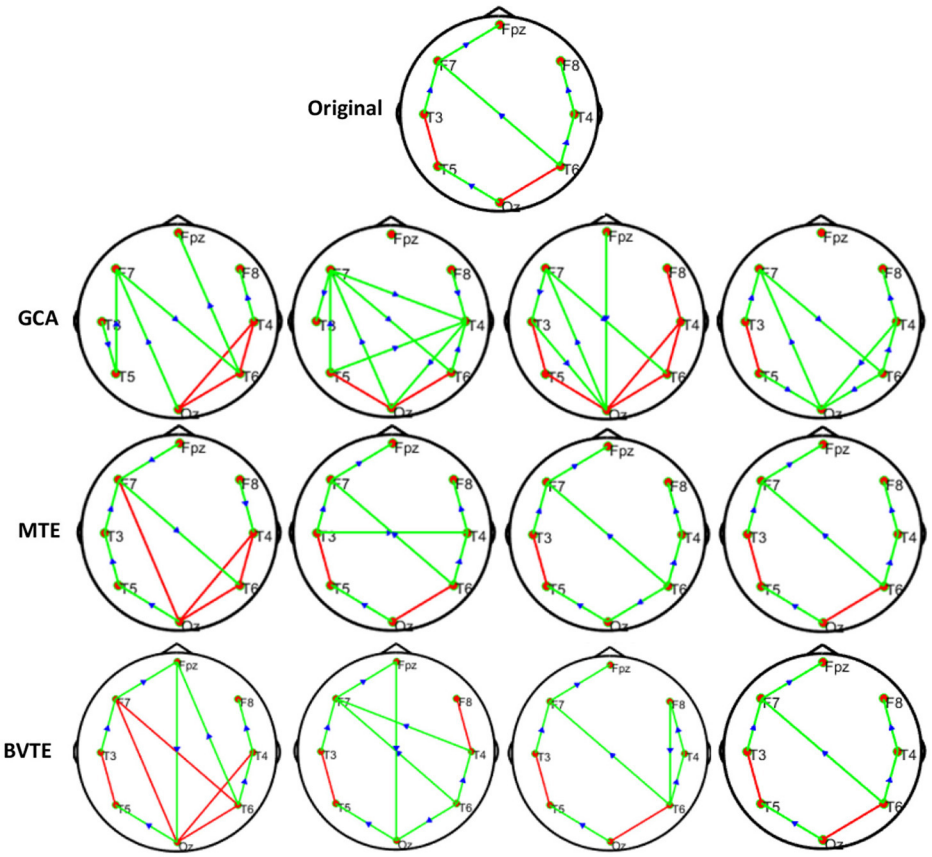


**FIGURE 3 |** Original or predefined 8 nodes simulated network and estimated non-linear networks by GCA, MTE, and BVTE with ($r = f(x)$, $r = \frac{(2.40\times 9x)}{1+\exp(-4x)}$).

driving noises, each of the network simulations was repeated 200 times for each linear and non-linear equations.

To know the percentage of available causal connections that are correctly detected as existent and the percentage of unavailable causal connections that are really detected as non-existent, the sensitivity and specificity analysis were calculated, respectively. Confusion matrix function is used for the sensitivity and specificity calculations. It is made up of a target matrix and the actual matrix. The confusion matrix compares the relationship between the target matrix and the actual matrix by comparing the rows of the target matrix with that of the actual matrix and returns four parameters (**Table 1**) including True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP).

$$Sensitivity\,(\%) = 100 \times TP/(TP + FN) \quad (8)$$

$$Specificity\,(\%) = 100 \times TN/(TN + FP) \quad (9)$$

The adjacency matrix linkage bias and network patterns are estimated using the GCA, BVTE, and MTE under various SNR conditions. Based on the simulated networks, we also compute the edges recovered and the adjacency matrix linkage bias. Adjacency matrix linkage bias can be defined as follows:

$$\Delta Y = \frac{\| Y_c - Y_b \|}{\| Y_c \|} \quad (10)$$

where $Y_c$ is the adjacency matrix linkage estimated without any added noise effect, and $Y_b$ is the corresponding parameter subjected to noise condition.

We also evaluated the strength of the networks produced by GCA, BVTE, and MTE by considering the total number of edges in the network. The 8 nodes network comprises 56 causal linkages, those edges with directed causal consistent with the originally defined edges are described as correct linkages.

## Simulation Performance

As displayed in **Figures 2**, **5**, under the linear condition, under most cases, the GCA, MTE, and BVTE could correctly estimate the network structures (**Figures 2**, **5**), respectively just the same with the original or predefined ones. Unfortunately, under the various non-linear conditions of varied SNRs, the GCA failed to capture the predefined network structure (**Figures 3**, **4**, **6**, **7**). In contrast, the MTE outperformed the GCA and BTE. **Figures 3**, **4**, **6**, **7** depict two of the non-linear simulation conditions($r = f(x)$,$r = \frac{(2.40\times 9x)}{1+\exp(-4x)}$), $r = S(x)$,$r = \frac{1}{(1+\exp(-x))}$, estimated by GCA, MTE and BVTE, respectively. These figures are similar

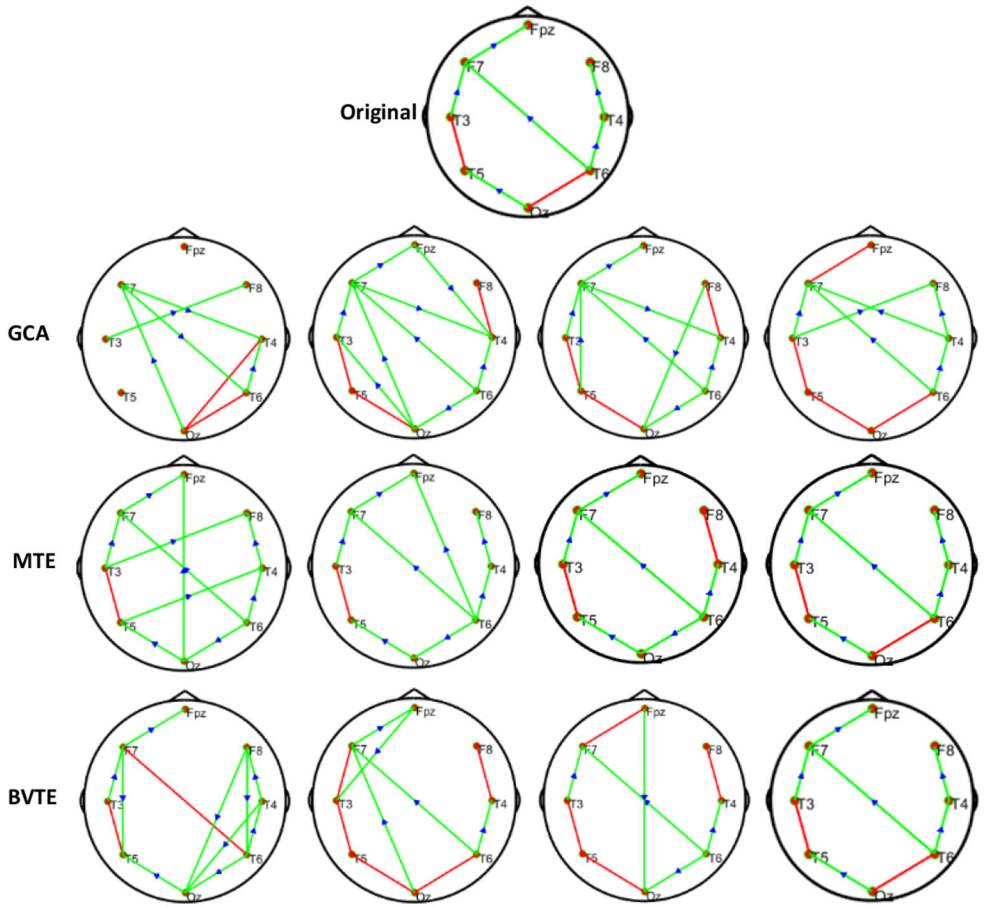


**FIGURE 4 |** Original or predefined 8 nodes simulated network and estimated non-linear networks by GCA, MTE, and BVTE with $r = S(x)$, $r = \frac{1}{(1+\exp(-x))}$.

to the other three non-linear simulation conditions. All the simulated figures have the similar structure, which includes original, GCA, MTE, and BVTE results. Besides, the results from left to right are under the SNR of −10, −5, 5, and 10 dB, by row, respectively. In each figure, the green arrows show unidirectional causal interactions and the red lines depict bidirectional connections.

To further demonstrate the advantages of MTE on the network edges recovery over GCA and BVTE, we added few more networks to the already demonstrated figures in **Figures 2**–**4** by simulating additional 7 nodes with networks of structures different from that in **Figures 2**–**4**. This is shown in **Figures 5**–**7**. It could be noticed from the figures again that MTE was able to recover the network edges better than GCA and BVTE both in linear and non-linear states.

Thereafter, **Tables 2**, **3** quantitatively display the performances of the average results from 200 runs with parameters of adjacency matrix linkage bias, edges recovered, sensitivity, and specificity under varied SNRs on the 8 nodes simulation. The values highlighted depict the estimator or method which had the least adjacency matrix linkage bias, the highest consistent linkage edges or recovery edges, and also the highest sensitivity and specificity. Out of the six simulations, the MTE outperformed the GCA and BVTE in both linear and non-linear conditions, which is validated by the independent paired *t*-test with a significance level of 0.05.

## Real P300 EEG

### Participants

This experiment included 48 right-handed (self-reported) participants, which consisted of 23 SCZ patients (10 females, age 28.87 ± 7.68) and 25 HCs (11 females, age 29.44 ± 5.75). All participants had the normal or corrected-to-normal vision. None of them had used any medication, and there had been no personal or family history of psychiatric or neurological disease. The Ethics Committee of Peking University Sixth Hospital approved this study. Before experiments, all participants gave the written informed consent with their names signed on it.

### Experimental Protocol

Before the commencement of the experiment, all participants were instructed to be seated comfortably, stay relaxed and were also asked to control their eye blinks and body movements in the experiments. A square with a thin cross in the center and a circle with a thin cross in the center were defined as the

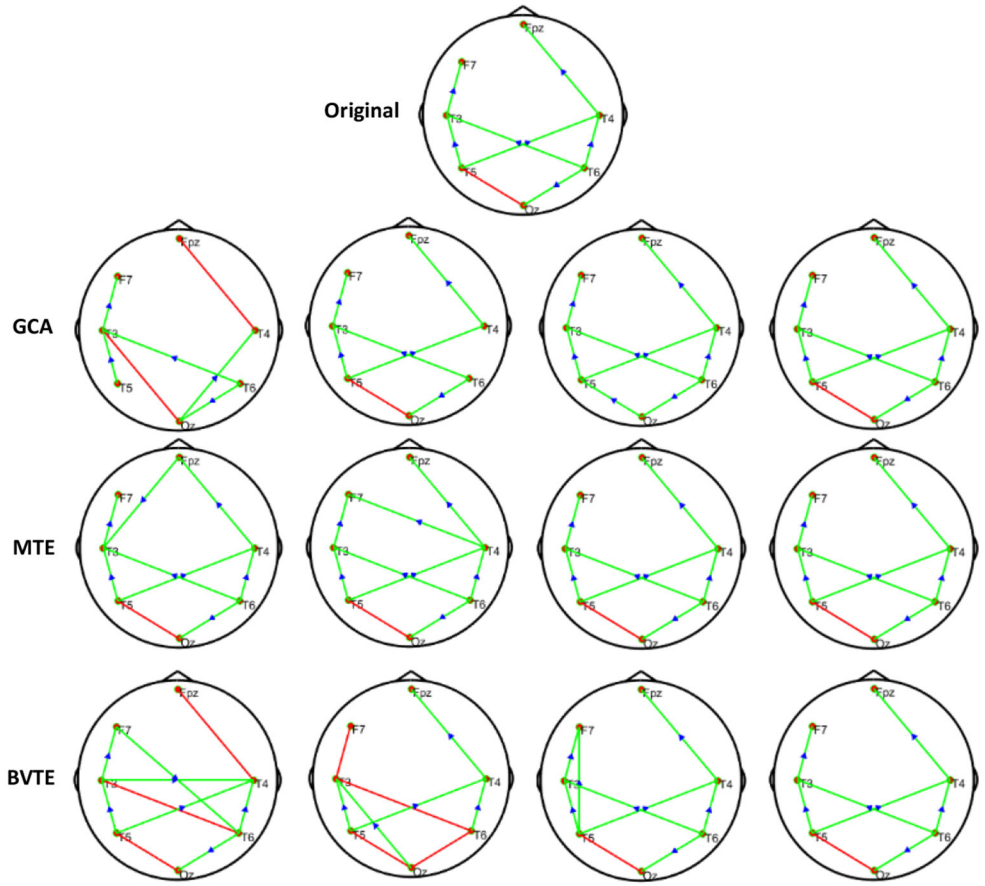


**FIGURE 5 |** Original or predefined 7 nodes simulated network and estimated linear networks by GCA, MTE, and BVTE with $Y = A \times B$.

standard and target stimulus, respectively. We included a 5-min, eye-closed resting-state session and four runs of P300 tasks during the experiments. In each P300 run, a total of 100 stimuli, 80 standards, and 20 targets, were randomly presented on the computer screen. **Figure 8** depicts the timeline of a given P300 trial. In detail, a bold-cross cue was first presented and lasted 750 ms to warn participants to focus their attention and to inform them that a standard (or target) stimulus would appear very soon. Either a standard or target stimulus then appeared on the screen for 150 ms. Participants were asked to press the "1" key on a standard keyboard when they noticed a target stimulus appeared at the same time. A 1,000-ms break was given after and the next trial began.

## EEG Recording

We recorded the EEG datasets with the Symtop amplifier (Symtop Instrument, Beijing, China) and a 16-channel Ag/AgCl (i.e., Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, and T6) electrode cap (BrainMaster, Inc., Shenzhen, China). We positioned all the electrodes used in accordance with the 10–20 international electrode placement system and digitized with a sampling rate of 1,000 Hz and online bandpass filtered at 0.05–100 Hz. Electrode AFz was used as the reference and was grounded during online recording. The total impedance during the whole task of all electrodes was kept below 5 $K\Omega$, during the recording.

## Effective Network

Since, we aimed to investigate the brain network deterioration of the SCZ in the oddball task, in this study, only the EEG datasets of the four runs of P300 tasks were included in the following analyses. To construct an effective network, we used multiple standard procedures to preprocess the task datasets. The multiple procedures comprise [0.5 Hz, 30 Hz] offline bandpass filtering, 1-s length data segment (ranging from 200 ms before and 800 ms after targets onset [−200 ms, 800 ms]), [−200 ms, 0 ms] baseline correction, artifact-trial removal using a threshold of ±100 μv, and Reference Electrode Standardization Technique (REST). Thereafter, based on the EEG time series we generated, the GCA, MTE, and BVTE were used to construct the corresponding weighted effective network for the HC and SCZ.

The effective network is a square asymmetric adjacency matrix where the number of rows and columns is equal to the number of electrodes. The GCA, MTE, and BVTE are then applied to estimate the adjacency matrix per task trial per subject. Thereafter, the final weighted rest (also task), a 16 × 16 adjacency

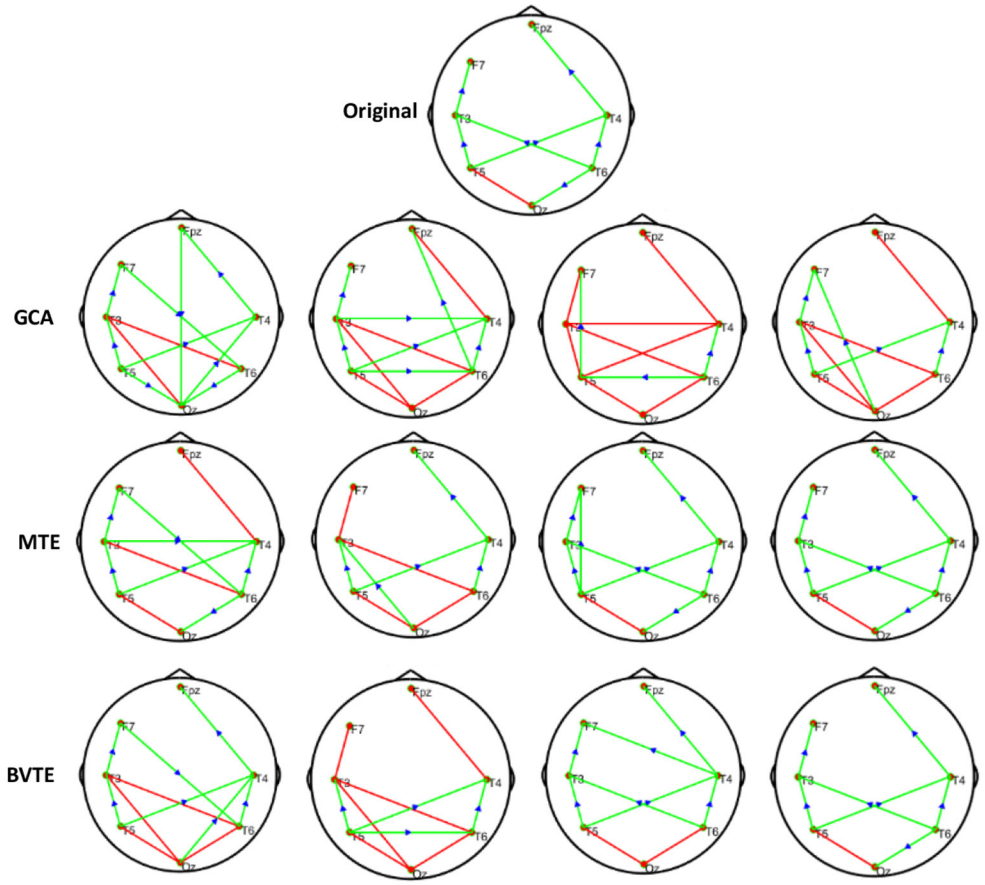


**FIGURE 6 |** Original or predefined 7 nodes simulated network and estimated non-linear networks by GCA, MTE, and BVTE with ($r = f(x)$, $r = \frac{(2.40 \times 9x)}{1+\exp(-4x)}$).

matrix, directed brain network for each subject was acquired by averaging matrices across all artifact-free segments (also task trials), and eventually, we conducted independent *t*-test to unearth the potential difference ($p < 0.05$) in the brain networks of HC and SCZ for both methods.

## Topological Differences in HC and SCZ

**Figure 9**, visually demonstrates differential network topology between HC and SCZ ($P < 0.05$, FDR corrected) estimated by the methods-GCA and MTE. As displayed in **Figure 9**, the GCA (**Figures 9A,B**) and MTE (**Figures 9C,D**) showed much denser connectivity for the HC, compared to that of the SCZ, which extended on the frontal and parietal lobes. In specific, the corresponding stronger and denser causal connectivity can be found to flow from prefrontal/frontal to parietal lobes. In addition, compared to the GCA, the MTE gives more causal linkages, shows the dense edges in the frontal lobe.

## Statistical Comparison for the Topographical Difference Between HCs and SCZ Patients

We conducted further analysis on **Figure 9** to prove our method MTE over the GCA using out degree in **Figure 10**. The node out-degree can be defined as the number of edges pointing out or going out of the node. The number of edges connecting the node with any or all other nodes is termed Node degree. If the nodes are more connected, it means they have greater degree and vice versa (Fornito et al., 2016). The degree of a node could be in-degree or out-degree. For example[1] in directed network, if we have an edge with a path from node i to node j, then Node i's out-degree is $\sum_{jgij}$ .

This has important influence on the brain network. This information flow can influence the properties of dynamical systems that evolve on the brain network, such as the synchronization of networked oscillators. Moreover, different nodes play or serve distinct topological roles in the brain network, with highly connected nodes exerting a particularly important influence over network function (Fornito et al., 2016). Thus, in our study after the construction of the differential network topology, we based our analysis on the information flow out of the node to further explain **Figure 9**.

After the out degree analysis, our proposed method-MTE still proved to be better than the conventional method GCA. In **Figure 10**, our method proved better because it could help locate

[1]Network/Graph Theory Graph-based representations Protein-Protein Interaction.

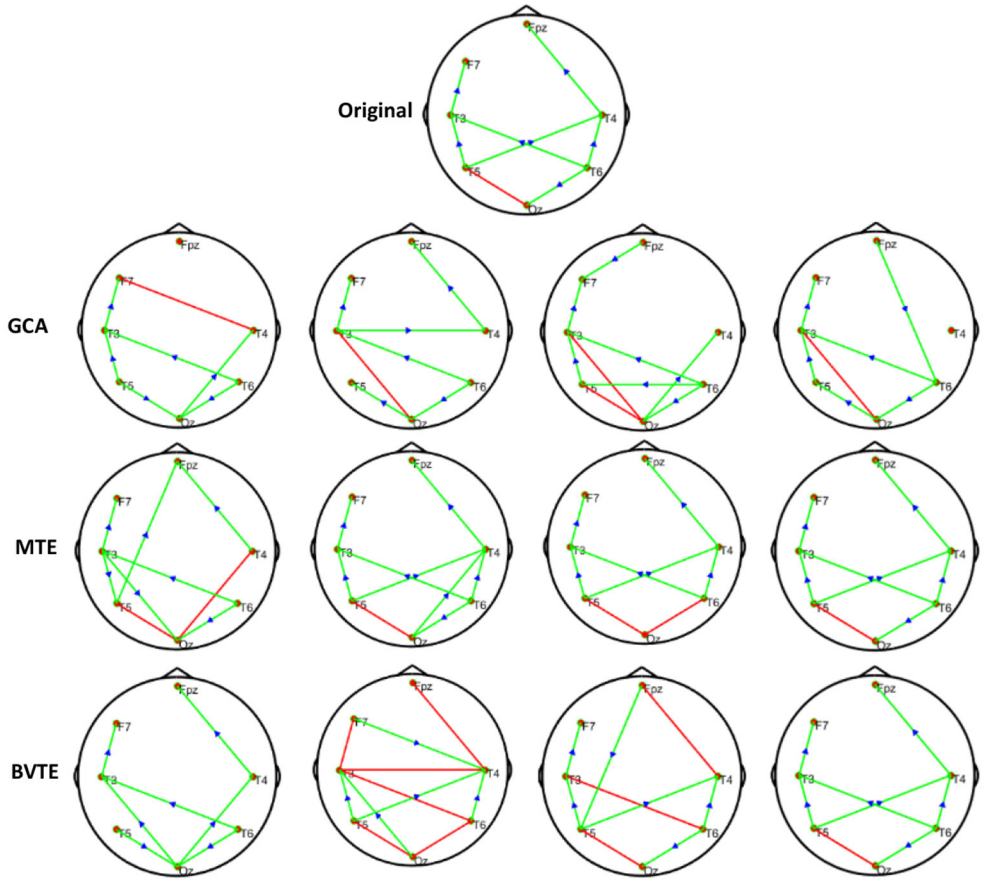


**FIGURE 7 |** Original or predefined 7 nodes simulated network and estimated non-linear networks by GCA, MTE, and BVTE with $r = S(x)$, $r = \frac{1}{(1+\exp(-x))}$.

the network channels well which the GCA method couldn't. There are significant differences between the HCs and SCZ for all the methods. However, MTE showed more outgoing degrees compared to GCA. The out degree for MTE could help locate the brain regions or channels better than the GCA and with this we could see the nodes which are highly connected and those with less or no connections. In **Figure 10**, MTE has more variation of information between all the channels compared to GCA. The colors correspond to a variation of information between the regions or channels (Van Den Heuvel and Fornito, 2014; Yang et al., 2017). GCA has the following results for its out degrees for HCs and SCZ patients, respectively:

Channels (Fp1 of HC and Fp1 of SCZ, Fp2 and Fp2, F3 and F3, F3 and F3, F4 and F4), have no difference in their channels. Meanwhile, the channels (F7 and F7, F8 and F8, C4 and C4, T3 and T3, T4 and T4, T5 and T5, T6 and T6, P3 and P3, P4 and P4, O1 and O1, O2 and O2) had a difference between them. The highest out degree for HCs is 2 for the channels- C2, C4, P3, and O1. SCZ patients had 1 as the highest out degree.

For MTE, only the channels (F8 and F8, C4 and C4) had no difference in between them. The channels (Fp1 and Fp1, Fp2 and Fp2, F7 and F7, F3 and F3, F4 and F4, T3 and T3, T4 and T4, C3 and C3, P3 and P3, P4 and P4, T5 and T5, T6 and T6, O1 and O1, O2 and O2) had a difference between their channels. In all, channels F3, P3 T3 and T5 had the highest out degrees for HCs whiles channel T4 also had the highest out degree for SCZ patients (Rubinov and Bullmore, 2013; van Straaten and Stam, 2013). The analysis above clearly show that our method MTE still had the best performance in the out degree condition. It had more information flow from out of the nodes and also more channel influence than the GCA method.

## DISCUSSION

Non-linearity characterizes our daily activities. Biological systems, such as EEG, is linear and inherently non-linear. Although linear methods are important and have obtained satisfying findings in EEG analysis, they compromise the underlying non-linearity characteristics or non-linear causal dynamics. The applications of non-linear methods in EEG analysis will, therefore, pave a way for logical steps that can be used to enhance the characterization of these signals. The GCA has the problem of model dependency, statistical and conceptual problems, and it ignores the system dynamics (Stokes et al., 2018). BVTE analysis also lead to spurious and redundant

**TABLE 2 |** A consistent number of edges recovered by GCA, BVTE, and MTE methods.

| Causal relationship function and description | Linear/Non-linear | Gaussian noise SNR (dB) | GCA | | BVTE | | MTE | |
|---|---|---|---|---|---|---|---|---|
| | | | Bias | Edges recovered | BVTE bias | Edges recovered | Bias | Edges recovered |
| $Y = A \times B$ | Linear | −10 | 0.98 ± 0.09 | 48.01 ± 1.08 | 0.99 ± 0.08 | 47.86 ± 1.09 | 0.61 ± 0.13 | 53.37 ± 0.79 |
| | | −5 | 0.95 ± 0.07 | 48.81 ± 1.04 | 0.97 ± 0.05 | 48.74 ± 1.06 | 0.59 ± 0.12 | 53.39 ± 0.78 |
| | | 5 | 0.75 ± 0.05 | 51.22 ± 1.01 | 0.76 ± 0.04 | 50.65 ± 1.03 | 0.48 ± 0.11 | 53.58 ± 0.67 |
| | | 10 | 0.62 ± 0.02 | 53.36 ± 0.78 | 0.65 ± 0.02 | 52.78 ± 0.80 | 0.36 ± 0.09 | 54.12 ± 0.41 |
| $r = C(x)$<br>$r = \cos(x) + \sin(x)$ | Non-linear | −10 | 0.99 ± 0.08 | 38.81 ± 3.23 | 0.98 ± 0.05 | 42.95 ± 3.15 | 0.64 ± 0.08 | 44.58 ± 0.10 |
| | | −5 | 0.97 ± 0.06 | 38.91 ± 3.03 | 0.82 ± 0.03 | 45.97 ± 3.01 | 0.60 ± 0.07 | 47.99 ± 0.07 |
| | | 5 | 0.78 ± 0.04 | 39.98 ± 3.01 | 0.73 ± 0.03 | 48.99 ± 2.47 | 0.52 ± 0.04 | 50.18 ± 0.04 |
| | | 10 | 0.72 ± 0.01 | 42.28 ± 2.88 | 0.64 ± 0.02 | 50.13 ± 2.70 | 0.51 ± 0.02 | 51.30 ± 0.01 |
| $r = f(x)$<br>$r = \frac{(2.40 \times 9x)}{1+\exp(-4x)}$ | Non-linear | −10 | 0.98 ± 0.08 | 42.58 ± 2.55 | 0.78 ± 0.08 | 44.89 ± 2.43 | 0.58 ± 0.09 | 47.99 ± 0.12 |
| | | −5 | 0.95 ± 0.07 | 45.69 ± 2.48 | 0.65 ± 0.05 | 47.67 ± 2.22 | 0.55 ± 0.07 | 49.68 ± 0.10 |
| | | 5 | 0.69 ± 0.03 | 47.82 ± 2.58 | 0.58 ± 0.02 | 49.32 ± 2.14 | 0.53 ± 0.04 | 51.04 ± 0.08 |
| | | 10 | 0.63 ± 0.02 | 49.21 ± 2.45 | 0.54 ± 0.01 | 51.16 ± 1.25 | 0.52 ± 0.02 | 52.06 ± 0.05 |
| $r = \cos inusoidal(x)$<br>$r = \cos(2\pi x)$ | Non-linear | −10 | 0.99 ± 0.05 | 38.78 ± 3.33 | 0.73 ± 0.08 | 46.18 ± 3.29 | 0.67 ± 0.14 | 48.96 ± 0.25 |
| | | −5 | 0.89 ± 0.04 | 43.71 ± 2.25 | 0.67 ± 0.03 | 47.71 ± 2.55 | 0.65 ± 0.13 | 48.99 ± 0.24 |
| | | 5 | 0.71 ± 0.02 | 45.01 ± 2.20 | 0.59 ± 0.02 | 48.01 ± 2.10 | 0.62 ± 0.11 | 49.70 ± 2.03 |
| | | 10 | 0.68 ± 0.01 | 46.86 ± 2.17 | 0.56 ± 0.01 | 50.86 ± 1.13 | 0.59 ± 0.07 | 51.42 ± 0.01 |
| $r = H(x)$<br>$r = \exp(\sin(2\pi x))$ | Non-linear | −10 | 0.99 ± 0.08 | 44.79 ± 1.79 | 0.68 ± 0.06 | 46.99 ± 0.32 | 0.69 ± 0.15 | 48.97 ± 0.29 |
| | | −5 | 0.98 ± 0.03 | 44.99 ± 0.99 | 0.67 ± 0.04 | 47.87 ± 0.11 | 0.68 ± 0.14 | 48.99 ± 0.09 |
| | | 5 | 0.70 ± 0.02 | 46.28 ± 0.89 | 0.64 ± 0.02 | 48.23 ± 0.72 | 0.61 ± 0.10 | 50.02 ± 0.05 |
| | | 10 | 0.59 ± 0.01 | 46.99 ± 0.61 | 0.59 ± 0.10 | 49.89 ± 0.45 | 0.58 ± 0.07 | 51.88 ± 0.02 |
| $r = S(x)$<br>$r = \frac{1}{(1+\exp(-x))}$ | Non-linear | −10 | 0.97 ± 0.09 | 47.55 ± 1.14 | 0.63 ± 0.07 | 47.67 ± 0.38 | 0.66 ± 0.11 | 48.98 ± 0.74 |
| | | −5 | 0.95 ± 0.07 | 48.32 ± 1.12 | 0.61 ± 0.03 | 49.42 ± 0.15 | 0.60 ± 0.09 | 50.12 ± 0.68 |
| | | 5 | 0.79 ± 0.05 | 49.34 ± 1.08 | 0.58 ± 0.09 | 49.78 ± 0.60 | 0.54 ± 0.07 | 51.03 ± 0.20 |
| | | 10 | 0.56 ± 0.02 | 49.99 ± 0.06 | 0.55 ± 0.07 | 51.88 ± 0.23 | 0.52 ± 0.04 | 52.45 ± 0.18 |

interactions and may miss synergistic interactions between multiple relevant sources and the target (Wollstadt et al., 2019). In the current study, we thus proposed to apply the MTE to the task EEGs of the SCZ and HC, to investigate the mechanism explaining the cognitive deficits in the SZ, from the perspective of effective connectivity.

GCA computation or estimation encounter many problems. It can either be severely biased or have high variance and these shortcomings lead to spurious, redundant, etc. results. GCA estimation or computation alone are not interpretable without examining the component behaviors of the system model even if these estimations are done correctly and also ignoring the critical components system's dynamics. On the basis of these analysis, the idea or notion of causality quantified is not compatible with the objectives of many neuroscience research investigations and this has led to highly counterintuitive and potentially misleading results with GCA (Stokes et al., 2018). GCA in time domain cannot correctly determine how strongly one time series influences the other especially when there is directional causality between two time series. In other words a larger GCA value does not necessarily mean higher real causality, or vice versa (Hu et al., 2016). Moreover, many connectivity measures like GCA that are based on the autoregressive model do not always reflect true neuronal connectivity (Schindler, 2011). TE was also formulated for the bivariate case; that is between a single source and a single target. However, in a multivariate setting, bivariate analysis may lead to false positive or false negative results inferring spurious or redundant causality or interactions and also missing synergistic interactions between important sources and the target. Usually, these many sources together send more information into the target than what could be detected from examining source contributions individually (Tanaka et al., 2013; James et al., 2016; Wollstadt et al., 2019). These findings are confirmed by our study in **Tables 2**, **3** and **Figures 2**–**7**, **9**, **10**, especially the networks revealed by the methods on the real data.

The MTE could detect both linear and non-linear signals better than the GCA and the BVTE and is able to account for all relevant sources of a target. By predefining the simulated network structure as well as the corresponding time courses, we applied the GCA, BVTE, and MTE methods to estimate the defined flow matrix and the directed networks under the influence of Gaussian noise in order of −10, −5, 5, and 10 dB, and evaluated the performance of the GCA, BVTE, and MTE under adjacency matrix linkage bias, edges recovered, sensitivity, and specificity. **Figures 2**, **5** demonstrate that the GCA, MTE, and BVTE have the potential for effectively estimating the originally defined network patterns under the linear condition of varied SNRs, respectively. However, as displayed in **Figures 3**, **4**, **6**, **7** corresponding to two of the various non-linear conditions, the GCA was not able to recover the original defined network patterns and produced many false linkages. Even though, BVTE was able to recover the predefined network but in contrast, the MTE outperforms the

**TABLE 3 |** Sensitivity and specificity analysis by GCA, BVTE, and MTE methods.

| Causal relationship function and description | Linear/Non-linear | Gaussian noise SNR(dB) | GCA | | BVTE | | MTE | |
|---|---|---|---|---|---|---|---|---|
| | | | Sensitivity | Specificity | Sensitivity | Specificity | Sensitivity | Specificity |
| $Y = A \times B$ | Linear | −10 | 89.59 ± 9.61 | 87.33 ± 2.62 | 85.53 ± 9.74 | 86.92 ± 2.62 | 91.92 ± 7.12 | 88.57 ± 5.69 |
| | | −5 | 92.64 ± 4.56 | 89.45 ± 5.66 | 89.74 ± 7.66 | 87.67 ± 5.71 | 93.98 ± 4.34 | 90.75 ± 3.32 |
| | | 5 | 94.72 ± 7.70 | 92.57 ± 6.72 | 93.83 ± 6.50 | 91.71 ± 7.43 | 95.88 ± 4.20 | 93.79 ± 5.51 |
| | | 10 | 95.98 ± 3.56 | 94.81 ± 5.52 | 94.87 ± 4.64 | 93.63 ± 7.33 | 97.99 ± 2.33 | 96.49 ± 3.61 |
| $r = C(x)$<br>$r = \cos(x) + \sin(x)$ | Non-linear | −10 | 68.34 ± 14.59 | 74.58 ± 5.34 | 85.34 ± 4.31 | 84.87 ± 5.56 | 88.54 ± 1.14 | 88.21 ± 2.42 |
| | | −5 | 75.52 ± 7.31 | 76.88 ± 6.43 | 87.43 ± 6.23 | 88.12 ± 4.40 | 91.49 ± 1.07 | 91.98 ± 2.32 |
| | | 5 | 83.74 ± 5.17 | 84.89 ± 3.26 | 90.61 ± 7.30 | 91.77 ± 4.17 | 93.91 ± 1.50 | 94.67 ± 3.12 |
| | | 10 | 91.07 ± 2.48 | 92.33 ± 1.16 | 93.11 ± 3.50 | 94.04 ± 1.82 | 96.01 ± 0.15 | 96.99 ± 1.04 |
| $r = f(x)$<br>$r = \frac{(2.40 \times 9x)}{1+\exp(-4x)}$ | Non-linear | −10 | 48.84 ± 2.41 | 72.68 ± 5.72 | 76.94 ± 3.54 | 82.87 ± 4.78 | 79.51 ± 3.86 | 85.96 ± 3.68 |
| | | −5 | 51.92 ± 1.39 | 78.96 ± 3.47 | 83.96 ± 2.87 | 87.31 ± 4.60 | 86.78 ± 2.91 | 89.10 ± 5.71 |
| | | 5 | 69.98 ± 2.89 | 87.88 ± 4.87 | 89.93 ± 1.78 | 87.72 ± 5.13 | 93.78 ± 0.98 | 91.09 ± 3.77 |
| | | 10 | 74.69 ± 2.76 | 91.21 ± 1.87 | 93.91 ± 2.19 | 90.88 ± 1.40 | 95.04 ± 1.58 | 93.16 ± 0.83 |
| $r = \cos inusoidal(x)$<br>$r = \cos(2\pi x)$ | Non-linear | −10 | 40.22 ± 21.25 | 82.40 ± 4.48 | 47.69 ± 2.71 | 83.54 ± 3.63 | 52.83 ± 1.28 | 87.67 ± 3.56 |
| | | −5 | 48.78 ± 2.96 | 89.31 ± 5.69 | 64.35 ± 1.50 | 90.09 ± 0.14 | 66.14 ± 0.89 | 92.18 ± 1.16 |
| | | 5 | 67.09 ± 4.58 | 93.17 ± 0.97 | 85.95 ± 5.66 | 91.08 ± 0.30 | 88.29 ± 4.09 | 93.99 ± 2.21 |
| | | 10 | 74.42 ± 2.84 | 94.98 ± 1.78 | 90.87 ± 1.11 | 92.20 ± 0.61 | 93.14 ± 0.89 | 94.58 ± 5.10 |
| $r = H(x)$<br>$r = \exp(\sin(2\pi x))$ | Non-linear | −10 | 49.71 ± 17.20 | 69.36 ± 4.66 | 69.88 ± 2.54 | 86.12 ± 0.41 | 72.26 ± 1.42 | 87.85 ± 1.28 |
| | | −5 | 57.12 ± 6.53 | 71.06 ± 1.77 | 84.63 ± 2.20 | 86.92 ± 0.13 | 86.07 ± 1.50 | 88.96 ± 0.88 |
| | | 5 | 68.09 ± 3.36 | 74.91 ± 0.82 | 86.12 ± 4.14 | 88.98 ± 0.94 | 89.81 ± 0.23 | 91.74 ± 1.65 |
| | | 10 | 77.82 ± 5.63 | 79.99 ± 0.74 | 91.90 ± 1.72 | 93.87 ± 2.19 | 94.51 ± 1.11 | 95.83 ± 1.32 |
| $r = S(x)$<br>$r = \frac{1}{(1+\exp(-x))}$ | Non-linear | −10 | 34.85 ± 7.86 | 66.87 ± 2.85 | 70.91 ± 3.13 | 83.73 ± 1.77 | 73.18 ± 2.63 | 86.98 ± 2.84 |
| | | −5 | 42.74 ± 5.83 | 71.93 ± 4.59 | 76.89 ± 1.25 | 86.42 ± 3.87 | 79.99 ± 0.82 | 88.79 ± 4.63 |
| | | 5 | 53.42 ± 6.73 | 76.89 ± 2.96 | 85.33 ± 0.84 | 87.75 ± 4.44 | 87.78 ± 1.14 | 90.03 ± 3.50 |
| | | 10 | 74.38 ± 7.42 | 79.61 ± 1.13 | 88.15 ± 2.37 | 89.56 ± 1.41 | 91.27 ± 1.08 | 92.11 ± 0.24 |

GCA and the BVTE under same conditions (**Figures 3**, **4**, **6**, **7**). The MTE is able to suppress the turbulent noise contaminated and efficiently estimated most of the original or predefined network linkages, which is unlike the GCA affected by the noise and thus performed badly. Specifically, the strength of edges recovered and the reduction of edges strength with bias errors, sensitivity, and specificity are shown in **Tables 2**, **3** which reveals clearly how these three methods are influenced by noise in linear and non-linear conditions. With consistency, MTE always held a good performance in all the functional indexes with less or lowest bias errors to GCA and BVTE in a mean of 200 runs. That is, in the linear and five non-linear simulations under all the SNR conditions, the MTE could recover highest linkages closed to the predefined network structure, compared to the GCA and BVTE, as well as the highest sensitivity and specificity. As illustrated previously, the MTE is capable of overcoming spurious or redundant interactions and is also able to reveal synergistic interactions between multiple relevant sources that the GCA and BVTE lack. The topological differences between the three methods indeed show clearly that the MTE method could estimate the networks better than the GCA and BVTE both in the simulation and the real task EEG computation.

A research by Bassett and Bullmore (2009) reported that the causal interactions between the components of the prefrontal-limbic system determines the global trajectories of the individual's brain activation, with the strengths and modulations of these causal interactions being potentially key components determining or underlying the differences between HC individuals and those with SCZ. Research also has it that SCZ patients have significant reduction in strength of functional connectivity and increased diversity of functional linkages. Meanwhile topologically, functional brain network has a reduction on clustering and small-worldness, probability of high-degree hubs, but increased robustness in the SCZ group. The medial parietal, premotor and cingulate, and right orbitofrontal cortical nodes of functional networks in SCZ also locally saw a reduction in degree and clustering (Lynall et al., 2010). A research conducted in Jalili and Knyazeva (2011) and Ray et al. (2017) indicated that many higher deficits in cognition in SCZ may be as a result of dysfunction of cognitive control deficits in SCZ. In a comparative analysis between SCZ and HCs, SCZ individuals demonstrated a reduced activation in the dorsolateral prefrontal cortex (DLPFC), ventrolateral prefrontal cortex (VLPFC), dorsal anterior cingulate cortex (ACC), pre-SMA, ventral premotor cortex, posterior areas in the temporal and parietal cortex, and sub-cortical areas. Further meta-analysis also revealed disrupted and decreased resting-state functional connectivity (rsFC) within the self-referential network and default mode network which play roles in the malfunction of information processing in SCZ, while the core network might act as a dysfunctional hub of regulation (Li S. et al., 2019). These meta-analysis results are consistent with our present studies in **Figures 9**, **10**.

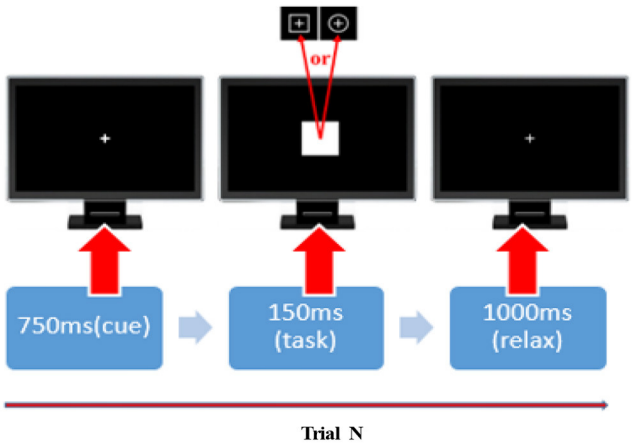


**FIGURE 8 |** The timeline of a given P300 trial. In each P300 trail, a 750-ms cue, 150-ms stimulus, and 1,000-ms break were added. The squares and circles with a thin cross in the center represent the standard and target stimuli in that order.

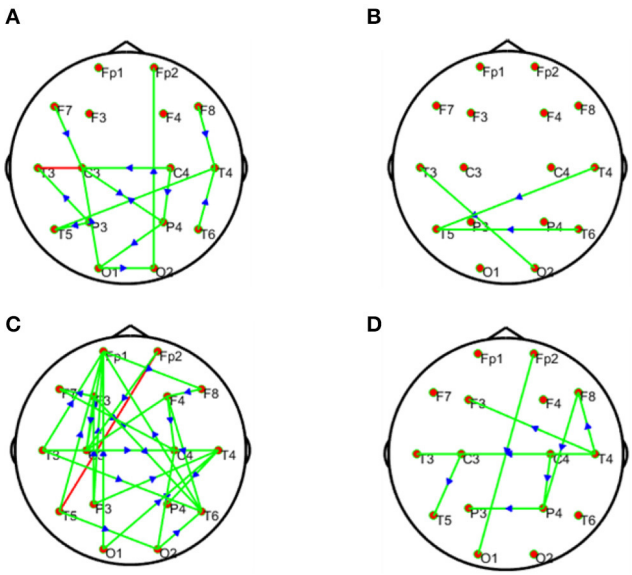


**FIGURE 9 |** Statistical analysis for the differential network topology between the HC and SCZ estimated by the GCA **(A,B)** and MTE **(C,D)**. The first column depicts that the connectivity of HC is stronger than that of SCZ, whereas the second column depicts that the connectivity of SCZ is lesser or weaker than that of HC. In each subfigure, the red and green lines depict bidirectional and unidirectional connectivity, respectively.

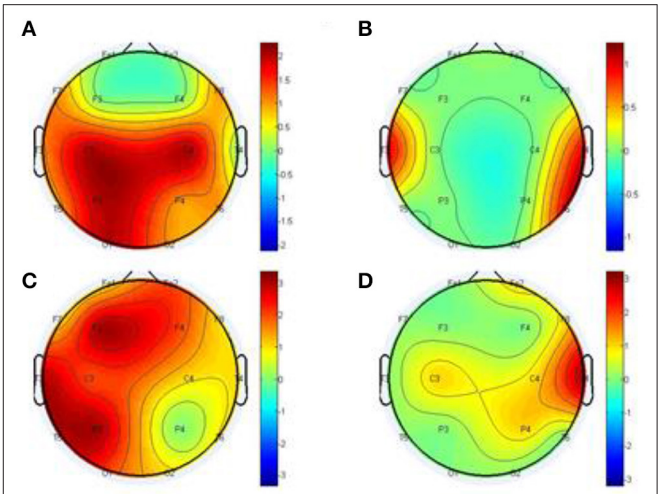


**FIGURE 10 |** Statistical comparison for the topographical difference between HCs and SCZ using out degree, estimated by the GCA **(A,B)** and MTE **(C,D)**. The first column depicts the connectivity of HC is stronger than that of SCZ, whereas the second column depicts the connectivity of SCZ is weaker than that of HC.

Based on our analysis and other findings, SCZ patients most often find it difficult to retain their attention during tasks unlike the HC. Usually the altered brain regions affect the information processing in the SCZ and these disruptions give rise to P300 malfunctions, which eventually disturbs the brain at rest in terms of abnormalities (Li F. et al., 2019). As a result of the malfunctioning of neurotransmitters, the ability of the SCZ patients to perceive reality is dumped (Karlsgodt et al., 2010; Alonso-Solís et al., 2015). In fact, people living with psychiatry or mental problems have severe brain network deterioration (Fogelson et al., 2014). The disruption of large-scale brain regions can largely account for the dysfunction of brain function in people living with the SCZ, and this disruption of the interregional connection may give rise to failure of the functional integration in the SCZ, thus paving a way for proper explanation of the abnormal behavior and cognitive impairment in patients with the SCZ (McKiernan et al., 2014; Zhang et al., 2019). Our findings in **Figures 9**, **10** indeed show the differential network topology and its comparison which show clearly the complete disruption of the multiple brain regions of the SCZ in relation to the HC agreeing with these studies. In specific, the HC showed the denser connectivity compared to that of the SCZ and these connections are extended on the frontal and parietal lobes. In essence, an alteration in causal connectivity between parts of the prefrontal cortex and the limbic system is found in Menon (2011), Qiu et al. (2014). The prefrontal cortex, the basal ganglia, and limbic system, etc. are interconnected and hence an attack of infection on one region will eventually affect the others. These above considerations drive us to conclude that the directed causal connectivity from prefrontal/frontal to parietal lobes is deteriorated, which then leads to the deficits in the P300, e.g., decreased P300 amplitudes.

Specifically, **Figures 2**–**7**, **9**, **10** again show clearly that the MTE method could estimate the networks better than the GCA and BVTE not only in the simulation (**Figures 2**–**7**, **Tables 2**, **3**), but also in the real EEG application with GCA in **Figures 9**, **10**. It holds its superiority over the GCA and BVTE in simulation and with GCA in real EEG analyses by giving a more satisfying performance. Our study and other studies (Gourévitch et al., 2006; Liu and Aviyente, 2012) have found that the GCA is not

robust enough in detecting non-linear linkages but it seems to be effective in detecting linear linkages. Also though BVTE could detect the non-linear causality better than GCA, in contrast, the MTE can address this problem. The MTE is able to handle spurious or redundant interactions and also unearth synergistic interactions between multiple relevant sources (Stokes et al., 2018; Wollstadt et al., 2019). Thus, when exploring the brain network deterioration in the SCZ patients, the MTE indeed outperforms the GCA and BVTE and seems to be a good choice.

## CONCLUSION

In summary, we testified to the fact that non-linear dynamics can give clearer information for better understanding of the causal dynamic issues surrounding EEG signals when it comes to its inherent non-linearity. Compared to the GCA and BVTE, the MTE was remarkably helpful in marking the causality either in a linear or non-linear system, which uncovered the brain dysfunction in effective connectivity for the SCZ that is deteriorated at the frontal and parietal lobes.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Peking University Sixth Hospital. The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

PX, JW, and WD conceived of and designed the experiments. JW performed the experiments. DH, CL, and YL analyzed the dataset. DH, FL, and PX wrote the manuscript. CL, WA, JB, and DY provided some useful suggestions in manuscript writing.

## FUNDING

This work was supported by the National Key Research and Development Plan of China (#2017YFB1002501), the National Natural Science Foundation of China (#61522105, #61603344, #81401484, and #81330032, #61701089), the Open Foundation of Henan Key Laboratory of Brain Science and Brain-Computer Interface Technology (No. HNBBL17001), and the Longshan academic talent research supporting program of SWUST (#17LZX692).

## REFERENCES

Adhikari, R., and Agrawal, R. K. (2013). *An Introductory Study on Time Series Modeling and Forecasting*. Riga: LAP LAMBERT Academic Publishing.

Alonso-Solís, A., Vives-Gilabert, Y., Grasa, E., Portella, M. J., Rabella, M., Sauras, R. B., et al. (2015). Resting-state functional connectivity alterations in the default network of schizophrenia patients with persistent auditory verbal hallucinations. *Schizophr. Res.* 161, 261–268. doi: 10.1016/j.schres.2014.10.047

Alvarado-González, M., Garduño, E., Bribiesca, E., Yáñez-Suárez, O., and Medina-Bañuelos, V. (2016). P300 Detection Based on EEG Shape Features. *Comput. Math. Methods Med.* 2016, 33–42. doi: 10.1155/2016/2029791

Anil, K. S., Barrett, A. B., and Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *J. Neurosci.* 35, 3293–3297. doi: 10.1523/JNEUROSCI.4399-14.2015

Bassett, D. S., and Bullmore, E. T. (2009). Human brain networks in health and disease. *Curr. Opin. Neurol.* 22, 340–347. doi: 10.1097/WCO.0b013e32832d93dd

Bonmati, E. (2018). Novel brain complexity measures based on information theory. *Entropy* 20:491. doi: 10.3390/e20070491

Bose, E., Hravnak, M., and Sereika, M. S. (2017). Vector autoregressive (VAR) models and granger causality in time series analysis in nursing research: dynamic changes among vital signs prior to cardiorespiratory instability events as an example. *Nurs. Res.* 66, 12–19. doi: 10.1097/NNR.0000000000000193

Dejman, A., Khadem, A., and Khorrami, A. (2017). "Exploring the disorders of brain effective connectivity network in ASD: a case study using EEG, transfer entropy, and graph theory," in *2017 25th Iranian Conference on Electrical Engineering (ICEE)* (Tehran), 8–13. doi: 10.1109/IranianCEE.2017.7985309

Domínguez-Iturza, N., Lo, A. C., Shah, D., Armendáriz, M., Vannelli, A., Mercaldo, V., et al. (2018). The autism and schizophrenia-associated protein CYFIP1 regulates bilateral brain connectivity. *bioRxiv* 477174. doi: 10.1101/477174

Dong, L., Zhang, Y., Zhang, R., Zhang, X., Gong, D., Valdes-Sosa, P. A., et al. (2015). Characterizing nonlinear relationships in functional imaging data using eigenspace maximal information canonical correlation analysis (emi CCA). *Neuroimage* 109, 388–401. doi: 10.1016/j.neuroimage.2015.01.006

Ehrlich, S., Geisler, D., Yendiki, A., Panneck, P., Roessner, V., Calhoun, V. D., et al. (2014). Associations of white matter integrity and cortical thickness in patients with schizophrenia and healthy controls. *Schizophr. Bull.* 40, 665–674. doi: 10.1093/schbul/sbt056

Flecker, B., Alford, W., Beggs, J. M., Williams, P. L., and Beer, R. D. (2011). Partial information decomposition as a spatiotemporal filter. *Chaos* 21, 1–11. doi: 10.1063/1.3638449

Fogelson, N., Litvak, V., Peled, A., Fernandez-del-Olmo, M., and Friston, K. (2014). The functional anatomy of schizophrenia: a dynamic causal modeling study of predictive coding. *Schizophr. Res.* 158, 204–212. doi: 10.1016/j.schres.2014.06.011

Fornito, A., Zalesky, A., and Bullmore, E. (2016). *Fundamentals of Brain Network Analysis. 1st Edn*. London, UK: Academic Press. p. 137–161.

Gaspar, P. A., Ruiz, S., Zamorano, F., Altayó, M., Pérez, C., Bosman, C. A., et al. (2011). P300 amplitude is insensitive to working memory load in schizophrenia. *BMC Psychiatry* 11:29. doi: 10.1186/1471-244X-11-29

Gilmour, T. P., Lagoa, C., Jenkins, W. K., Rao, A. N., Berk, M. A., and Venkiteswaran, K. (2012). "Transfer entropy between cortical and basal ganglia electrophysiology," in *2012 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (New York, NY).

Gourévitch, B., Bouquin-Jeannès, R. L., and Faucon, G. (2006). Linear and nonlinear causality between signals: methods, examples and neurophysiological applications. *Biol. Cybern.* 95, 349–369. doi: 10.1007/s00422-006-0098-0

Henderson, M., Harvey, S. B., Overland, S., Mykletun, A., and Hotopf,. M. (2011). Work and common psychiatric disorders. *J. R. Soc. Med.* 105, 198–207. doi: 10.1258/jrsm.2011.100231

Hristopulos, D. T., Babul, A., Babul, S., and Brucar, L., R. (2019). Resting-state directed brain connectivity patterns in adolescents from source-reconstructed EEG signals based on information flow rate. *bioRxiv* 608299. doi: 10.1101/608299

Hu, S., Cao, Y., Zhang, J., and Kong, W. (2016). Shortcomings/limitations of blockwise granger causality and advances of blockwise new causality. *IEEE Trans. Neural Netw. Learn. Syst.* 27, 2588–2601. doi: 10.1109/TNNLS.2015.2497681

Jalili, M., and Knyazeva, M. G. (2011). EEG-based functional networks in schizophrenia. *Comput. Biol. Med.* 41, 1178–1186. doi: 10.1016/j.compbiomed.2011.05.004

James, R. G., Barnett, N., and Crutchfield, J., P. (2016). Information flows? A critique of transfer entropies. *Phys. Rev. Lett.* 116:238701. doi: 10.1103/PhysRevLett.116.238701

Karlsgodt, K. H., Sun, D., and Cannon, T. D. (2010). Structural and functional brain abnormalities in schizophrenia. *Curr. Dir. Psychol. Sci.* 19, 226–231. doi: 10.1177/0963721410377601

Khadem, A., and Hossein-Zadeh, G. A. (2014). Estimation of direct nonlinear effective connectivity using information theory and multilayer perceptron. *J. Neurosci. Methods* 229, 53–67. doi: 10.1016/j.jneumeth.2014.04.008

Krusienski, D. J., Sellers, E. W., Cabestaing, F., Bayoudh, S., McFarland, D. J., Vaughan, T. M., et al. (2006). A comparison of classification techniques for the P300 speller. *J. Neural Eng.* 3, 299–305. doi: 10.1088/1741-2560/3/4/007

Lainscsek, C., Hernandez, M. E., Weyhenmeyer, J., Sejnowski, T. J., and Poizner, H. (2013). Non-linear dynamical analysis of EEG time series distinguishes patients with Parkinson's disease from healthy individuals. *Front. Neurol.* 4:200. doi: 10.3389/fneur.2013.00200

Lee, Y. J., Zhu, Y. S., Xu, Y. H., Shen, M. F., Zhang, H. X., and Thakor, N. V. (2001). Detection of non-linearity in the EEG of schizophrenic patients. *Clin. Neurophysiol.* 112, 1288–1294. doi: 10.1016/S1388-2457(01)00544-2

Li, F., Wang, J., Jiang, Y., Si, Y., Peng, W., Song, L., et al. (2018). Top-down disconnectivity in schizophrenia during P300 tasks. *Front. Comput. Neurosci.* 12:33. doi: 10.3389/fncom.2018.00033

Li, F., Wang, J., Liao, Y., Yi, C., Jiang, Y., Si, Y., et al. (2019). Differentiation of schizophrenia by combining the spatial EEG brain network patterns of rest and task P300. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 594–602. doi: 10.1109/TNSRE.2019.2900725

Li, P., Huang, X., Li, F., Wang, X., Zhou, W., Liu, H., et al. (2017). Robust Granger analysis in Lp norm space for directed EEG network analysis. *IEEE Trans. Neural Syst. Rehabil. Eng.* 25, 1959–1969. doi: 10.1109/TNSRE.2017.2711264

Li, S., Hu, N., Zhang, W., Tao, B., Dai, J., Gong, Y., et al. (2019). Dysconnectivity of multiple brain networks in schizophrenia : a meta- analysis of resting-state functional connectivity. *Front. Psychiatry* 10:482. doi: 10.3389/fpsyt.2019.00482

Liu, J., Li, M., Pan, Y., Lan, W., Zheng, R., Wu, F.-X., et al. (2017). Complex brain network analysis and its applications to brain disorders: a survey. *Complexity* 2017, 1–27. doi: 10.1155/2017/3014163

Liu, Y., and Aviyente, S. (2012). Quantification of effective connectivity in the brain using a measure of directed information. *Comput. Math. Methods Med.* 2012:635103. doi: 10.1155/2012/635103

Lynall, M. E., Bassett, D. S., Kerwin, R., McKenna, P. J., Kitzbichler, M., Muller, U., et al. (2010). Functional connectivity and brain networks in schizophrenia. *J. Neurosci.* 30, 9477–9487. doi: 10.1523/JNEUROSCI.0333-10.2010

Madulara, M. D., Francisco, P. A. B., Nawang, S., Arogancia, D. C., Cellucci, C. J., Rapp, P., et al. (2012). Eeg transfer entropy tracks changes in information transfer on the onset of vision. *Int. J. Mod. Phys. Conf. Ser.* 17, 9–18. doi: 10.1142/S201019451200788X

Mastrovito, D., Hanson, C., and Hanson, S. J. (2018). Differences in atypical resting-state effective connectivity distinguish autism from schizophrenia. *NeuroImage Clin.* 18, 367–376. doi: 10.1016/j.nicl.2018.01.014

McKiernan, K., Pearlson, G. D., Garrity, A. G., Calhoun, V. D., Lloyd, D., Kiehl, K., et al. (2014). Aberrant 'Default Mode' functional connectivity in schizophrenia. *Am. Psychiatry J.* 164, 450–457. doi: 10.1176/ajp.2007.164.3.450

Mehta, K., and Kliewer, J. (2016). "Directed information measures for assessing perceived audio quality using EEG," in *2015 49th Asilomar Conference on Signals, Systems and Computers* (Pacific Grove, CA), 123–127.

Mehta, K., and Kliewer, J. (2018). Directional and causal information flow in EEG for assessing perceived audio quality. *IEEE Trans. Mol. Biol. Multi-Scale Commun.* 1–16. *arXiv: 1802.06327*.

Menon, V. (2011). Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn. Sci.* 15, 483–506. doi: 10.1016/j.tics.2011.08.003

Montalto, A., Faes, L., and Marinazzo, D. (2014). MuTE : a MATLAB toolbox to compare established and novel estimators of the multivariate transfer entropy. *PLoS ONE* 9:e109462. doi: 10.1371/journal.pone.0109462

Novelli, L., Wollstadt, P., Mediano, P., Wibral, M., and Lizier, J. T. (2019). Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. *Netw. Neurosci.* 3, 827–847. doi: 10.1162/netn_a_000

Ozaki, T. (2012). *Time Series Modeling of Neuroscience Data (Chapman & Hall/CRC Interdisciplinary Statistics), 1st Edn*. CRC Press. p. 286–305.

Patel, K. R., Cherian, J., and Gohil, K. (2014). Schizophrenia: overview and treatment options. *P T.* 39, 638–645.

Pereda, E., Quiroga, R. Q., and Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals. *Prog. Neurobiol.* 77, 1–37. doi: 10.1016/j.pneurobio.2005.10.003

Pérez-Vidal, A. F., Garcia-Beltran, C. D., Martínez-Sibaja, A., and Posada-Gómez, R. (2018). Use of the stockwell transform in the detection of P300 evoked potentials with low-cost brain sensors. *Sensors* 18, 1–14. doi: 10.3390/s18051483

Qiu, Y. Q., Tang, Y. X., Chan, R. C., Sun, X. Y., and He, J. (2014). P300 aberration in first-episode schizophrenia patients: a meta-analysis. *PLoS ONE* 9:e97794. doi: 10.1371/journal.pone.0097794

Ray, K. L., Lesh, T. A., Howell, A. M., Salo, T. P., Ragland, J. D., MacDonald, A. W., et al. (2017). Functional network changes and cognitive control in schizophrenia. *NeuroImage Clin.* 15, 161–170. doi: 10.1016/j.nicl.2017.05.001

Rubinov, M., and Bullmore, E. (2013). Schizophrenia and abnormal brain network hubs. *Dial. Clin. Neurosci.* 15, 339–349.

Sabesan, S., Tsakalis, K., Spanias, A., and Iasemidis, L. (2010). "A robust estimation of information flow in coupled nonlinear systems," in *Springer Optimization and Its Applications*, Vol. 38 (New York, NY: Springer International Publishing), 271–283.

Schindler, K. (2011). Equivalence of granger causality and transfer entropy: a generalization. *Appl. Math. Sci.* 5, 3637–3648.

Schreiber, T. (2006). Measuring information transfer – aim: improve on standard use of mutual information. *Phys. Rev. Lett.* 85, 461–464. doi: 10.1103/PhysRevLett.85.461

Selskii, A. O., Hramov, A. E., Pisarchik, A. N., Moskalenko, O. I., and Zhuravlev, M., O. (2017). The nonlinear association analysis of the EEG brain data in the process of bistable image perception. *PHYSCON 2017* (Florence), 17–19.

Shao, S., Guo, C., Luk, W., and Weston, S. (2015). "Accelerating transfer entropy computation," in *2014 International Conference on Field-Programmable Technology* (Shanghai), 60–67. doi: 10.1109/FPT.2014.7082754

Shovon, M. H. I., Nandagopal, N., Vijayalakshmi, R., Du, J. T., and Cocks, B. (2017). Directed connectivity analysis of functional brain networks during cognitive activity using transfer entropy. *Neural Process. Lett.* 45, 807–824. doi: 10.1007/s11063-016-9506-1

Sohrabpour, A., Ye, S., Worrell, G. A., Zhang, W., and He, B. (2016). Noninvasive electromagnetic source imaging and granger causality analysis: an electrophysiological connectome (eConnectome) approach. *IEEE Trans. Biomed. Eng.* 63, 2474–2487. doi: 10.1109/TBME.2016.2616474

Somani, S., and Shukla, J. (2012). The P300 wave of event-related-potential. *Res. Rev. J. Med. Heal. Sci.* 3, 33–42.

Sørensen, B. E., and Causality, G. (2005). *1 1.1 Granger Causality*, 1–4.

Srivastava, M. S. (2002). *Methods of Multivariate Statistics, 1st Edn*. New York, NY: Wiley-Interscience. p. 79–95.

Stokes, P. A., and Purdon, P., L. (2018). A study of problems encountered in Granger causality analysis from a neuroscience perspective. *Proc. Natl. Acad. Sci. U.S.A.* 115:E6964. doi: 10.1073/pnas.1809324115

Tanaka, K., Mizuno, Y., Tanaka, T., and Kitajo, K. (2013). "Detection of phase synchronization in EEG with Bivariate Empirical Mode Decomposition," in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (Osaka), 3–7.

Ure, J. A., Corral, R., and Wainwright, E. (2018). Schizophrenia and brain networks. *Neuro. Neurosurg.* 1, 2–10. doi: 10.15761/NNS.1000102

Van Den Heuvel, M. P., and Fornito, A. (2014). Brain networks in schizophrenia. *Neuropsychol. Rev.* 24, 32–48. doi: 10.1007/s11065-014-9248-7

van Straaten, E. C., and Stam, C. J. (2013). Structure out of chaos : functional brain network analysis with EEG, MEG, and functional MRI. *Eur. Neuropsychopharmacol.* 23, 7–18. doi: 10.1016/j.euroneuro.2012.10.010

Venkatesh, P., and Grover, P. (2016). "Is the direction of greater Granger causal influence the same as the direction of information flow?,"in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing*

*(Allerton)* (Monticello, IL), 672–679. doi: 10.1109/ALLERTON.2015.7447069

Wang, C., Yu, H., Grout, W. R., Ma, K., and Chen, J. H. (2011). "Analyzing information transfer in time-varying multivariate data," in *2011 IEEE Pacific Visualization Symposium* (Hong Kong), 99–106.

Wollstadt, P., Martínez-Zarzuela, M., Vicente, R., Díaz-Pernas, F. J., and Wibral, M. (2014). Efficient transfer entropy analysis of non-stationary neural time series. *PLoS ONE* 9:e102833. doi: 10.1371/journal.pone.0102833

Wollstadt, P., Patricia, Lizier, Joseph, Vicente, Raul, et al. (2019). IDTxl: The Information Dynamics Toolkit xl : a Python package for the efficient analysis of multivariate information dynamics in networks. *J. Open Source Softw.* 4:108. doi: 10.21105/joss.01081

Yang, J., Hu, C., Guo, N., Dutta, J., Vaina, L. M., Johnson, K. A., et al. (2017). Partial volume correction for PET quantification and its impact on brain network in Alzheimer's disease. *Sci. Rep.* 7:13035. doi: 10.1038/s41598-017-13339-7

Zhang, X., Wang, L., Ding, Y., Huang, L., and Cheng, X. (2019). Brain network analysis of schizophrenia based on the functional connectivity. *Chinese J. Electron.* 28, 535–541. doi: 10.1049/cje.2019.03.017

Zhang, Z., Liao, W., Zuo, X. N., Wang, Z., Yuan, C., Jiao, Q., et al. (2011). Resting-state brain organization revealed by functional covariance networks. *PLoS ONE* 6:e28817. doi: 10.1371/journal.pone.0028817

Zhao, Y., Billings, S. A., Wei, H., He, F., and Sarrigiannis, P. G. (2013). A new NARX-based Granger linear and nonlinear casual influence detection method with applications to EEG data. *J. Neurosci. Methods* 212, 79–86. doi: 10.1016/j.jneumeth.2012.09.019

Zhu, Q., Huang, J., and Xu, X. (2018). Non-negative discriminative brain functional connectivity for identifying schizophrenia on resting-state fMRI. *Biomed. Eng.* 17, 1–15. doi: 10.1186/s12938-018-0464-x

**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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 15 January 2020
doi: 10.3389/fncom.2019.00093

# The Self-Face Paradigm Improves the Performance of the P300-Speller System

*Zhaohua Lu, Qi Li*, Ning Gao and Jingjing Yang*

*School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China*

**Objective:** Previous studies have shown that the performance of the famous face P300-speller was better than that of the classical row/column flashing P300-speller. Furthermore, in some studies, the brain was more active when responding to one's own face than to a famous face, and a self-face stimulus elicited larger amplitude event-related potentials (ERPs) than did a famous face. Thus, we aimed to study the role of the self-face paradigm on further improving the performance of the P300-speller system with the famous face P300-speller paradigm as the control paradigm.

**Methods:** We designed two facial P300-speller paradigms based on the self-face and a famous face (Ming Yao, a sports star; the famous face spelling paradigm) with a neutral expression.

**Results:** ERP amplitudes were significantly greater in the self-face than in the famous face spelling paradigm at the parietal area from 340 to 480 ms (P300), from 480 to 600 ms (P600f), and at the fronto-central area from 700 to 800 ms. Offline and online classification results showed that the self-face spelling paradigm accuracies were significantly higher than those of the famous face spelling paradigm at superposing first two times ($P < 0.05$). Similar results were found for information transfer rates ($P < 0.05$).

**Conclusions:** The self-face spelling paradigm significantly improved the performance of the P300-speller system. This has significant practical applications for brain-computer interfaces (BCIs) and could avoid infringement issues caused by using images of other people's faces.

**Keywords: brain-computer interface (BCI), event-related potential, famous face, P300-speller, self-face**

OPEN ACCESS

***Edited by:***
*Pei-Ji Liang,*
*Shanghai Jiao Tong University, China*

***Reviewed by:***
*Erwei Yin,*
*China Astronaut Research and Training Center, China*
*Radwa Khalil,*
*Jacobs University Bremen, Germany*

****Correspondence:***
*Qi Li*
*liqi@cust.edu.cn*

***Received:*** *14 May 2019*
***Accepted:*** *20 December 2019*
***Published:*** *15 January 2020*

***Citation:***
*Lu Z, Li Q, Gao N and Yang J (2020) The Self-Face Paradigm Improves the Performance of the P300-Speller System. Front. Comput. Neurosci. 13:93. doi: 10.3389/fncom.2019.00093*

## INTRODUCTION

A brain-computer interface (BCI) is a communication technology based on brain activity. BCIs allow severely disabled patients, especially patients with amyotrophic lateral sclerosis, to send messages or control external devices without physical actions (Thompson et al., 2013; Rosenfeld and Wong, 2017; Lazarou et al., 2018). BCIs can also help restore function in patients with severe motor disabilities, including patients with spinal cord injury, stroke, neuromuscular disorder, and limb amputation (Takeuchi et al., 2015; Carelli et al., 2017; Wang et al., 2019). In recent years, some studies have used BCIs for enhancing clinical communication assessments in patients with disorders of consciousness (Wang et al., 2017; Jeunet et al., 2018). BCIs are commonly based on electroencephalogram (EEG) that is recorded non-invasively via electrodes placed on the surface of the head (Waldert, 2016).

The P300 event-related potential (ERP) induced by an oddball paradigm is commonly used in non-invasive BCI systems (Bernat et al., 2001). Farwell and Donchin (1988) first applied the P300 potential to a BCI system; they achieved a character-spelling system based on the P300, which was called the P300-speller system. The users attend to a cell of the matrix (that is, a target character) and count the number of times it is intensified. In this system, the probability of the intensified row/column containing the target character is 1/6 (a matrix of 6 rows and 6 columns), which is an oddball event, which therefore would induce P300 potentials; the system can then output a character by analyzing the P300 potentials. However, the system was not satisfactory due to its low speed and variable accuracy.

A number of studies have attempted to design different paradigms to improve the performance of the P300-speller system (Allison and Pineda, 2003, 2006; Sellers et al., 2006; Salvaris and Sepulveda, 2009; Li et al., 2019). Kaufmann et al. (2011) introduced the famous face paradigm into the P300-speller system and found that its performance was markedly superior to that of the conventional P300-speller system, because the face stimulus also induced other ERPs (e.g., the N170) in addition to an increased P300 amplitude, which enhanced the waveform difference between the target and non-target characters. Subsequently, Jin et al. (2012) compared the performance of P300-speller system between the stimulus types involving a famous face, character flashing, and character movement, and the results showed that the system performed significantly better under the famous face condition than under the other two conditions. Recently, Speier et al. (2017) compared the stimulus types in an online classification of the P300-speller, and the results showed that famous faces stimuli yielded superior results than that with both standard and character inversion stimuli. Some researchers have attempted to optimize the face paradigm to improve the performance of the P300-speller system. For example, Jin et al. (2014b) designed a new stimulus presentation based on facial expression changes, to reduce adjacent interference annoyance and fatigue. Li et al. (2015) combined chromatic properties and the famous face spelling paradigm, which improved the performance of the P300-speller system.

Studies on human face recognition have shown that the brain has specialized cognitive processing for one's own face as compared with other faces. When participants searched for their own face vs. another face, they consistently processed their own face faster than other faces (Tong and Nakayama, 1999). Prior fMRI studies have shown that neural activity was enhanced over the frontal central area for self-face recognition as compared to other face recognition (Kircher et al., 2001). Some ERP studies on human face recognition have shown that the self-face induced greater ERP amplitudes than did other faces. The P300 is more sensitive to the self-face than to other faces (Ninomiya et al., 1998). For example, several studies have found that one's own face elicits a larger P300 amplitude than does a famous face (Caharel et al., 2005; Sui et al., 2006; Miyakoshi et al., 2008; Keyes et al., 2010; Tacikowski et al., 2011). The N170 is face-specific component that reflects facial perception (Bentin and Deouell, 2000; Schweinberger et al., 2002; Herzmann et al., 2004; Carbon et al., 2005). In Caharel et al.'s (2005) study on face processing, the self-face induced a larger N170 amplitude than did famous and unknown faces, distinguishing the self-face from famous and unknown faces. Other studies have also found that the self-face induced a larger N170 amplitude than did other faces (Miyakoshi et al., 2008; Keyes et al., 2010).

Thus, existing studies of face recognition have suggested that the brain is more active in response to the self-face than to a famous face. In the present study, we designed a new spelling paradigm based on self-face stimuli, in which we replaced the famous face with the self-face, to investigate whether the use of the self-face could improve the performance of the P300-speller system. The control paradigm was that of the famous face spelling paradigm. We analyzed the ERP waveforms induced in the self-face and famous face spelling paradigms and compared the classification accuracies between the two spelling paradigms.

# MATERIALS AND METHODS

## Subjects

A total of 20 subjects participated in the study; of these, one group ($n = 10$, three men, aged 20–28 years, mean 24.4 years) participated in the offline experiment, and the other group ($n = 10$, six men, aged 22–29 years, mean 25.6 years) participated in the online experiment. The subjects did not have any known neurological disorders and had a normal or corrected-to-normal vision. This study was carried out in accordance with the recommendations of the ethics committee of Changchun University of Science and Technology, which approved the protocol. All subjects gave written informed consent in accordance with the Declaration of Helsinki. All subjects were native Chinese speakers but were familiar with the Western characters used in the display.

## Spelling Paradigms

We designed two P300-speller paradigms based on the conventional P300-speller paradigm. For each paradigm, 36 characters were presented in a 6 × 6 matrix subtended at a 13.4° × 19.4° (24 × 1.5 cm) visual angle on a 19-in screen with a refresh rate of 60 Hz (**Figure 1**). In the first paradigm, the rows or columns of the characters were covered with pictures of the subject's self-face while they were intensified (self-face spelling paradigm, as shown in **Figure 1**; the subject has provided permission to publish his facial photograph in **Figure 1**). In the control spelling paradigm (the famous face spelling paradigm), the characters were covered with the famous face, and the paradigm's setup was the same as that of the self-face spelling paradigm.

We chose a picture of Ming Yao, a sports star, as the famous face. The subjects' self-face was photographed with a digital camera for the self-face paradigm. All facial images were frontal and showed a neutral expression. These photographs were processed to remove the background and everything below the neck in Adobe Photoshop (Adobe Systems, Inc. San Jose, CA, USA).

In our study, the characters were intensified according to the rows and columns of a virtual matrix (**Figure 1**, right). In the

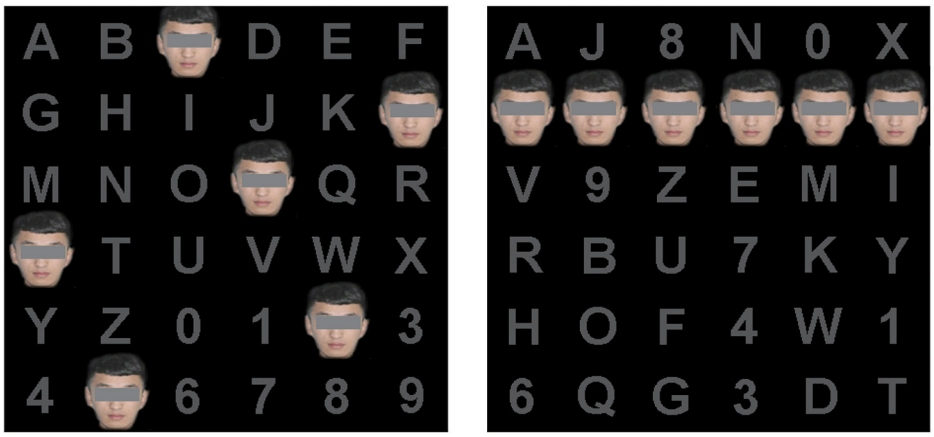


**FIGURE 1 |** The spelling paradigm. The left figure is the actual spelling matrix, and the right figure is a virtual matrix of the spelling paradigm. The figure shows the self-face paradigm in which the facial photograph is that of a subject.

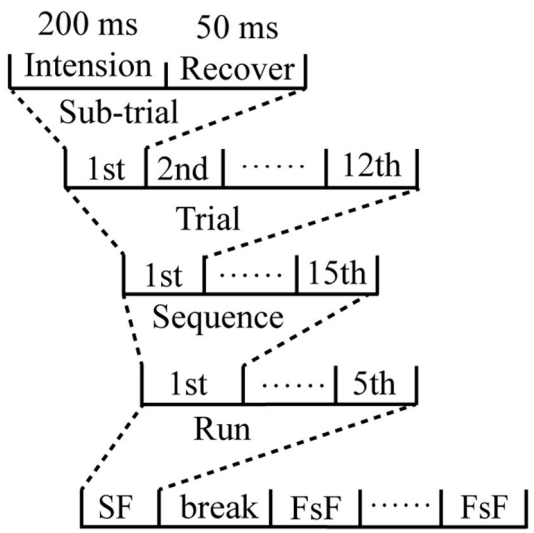


**FIGURE 2 |** Diagrammatic representation of the time-course of the experiment.

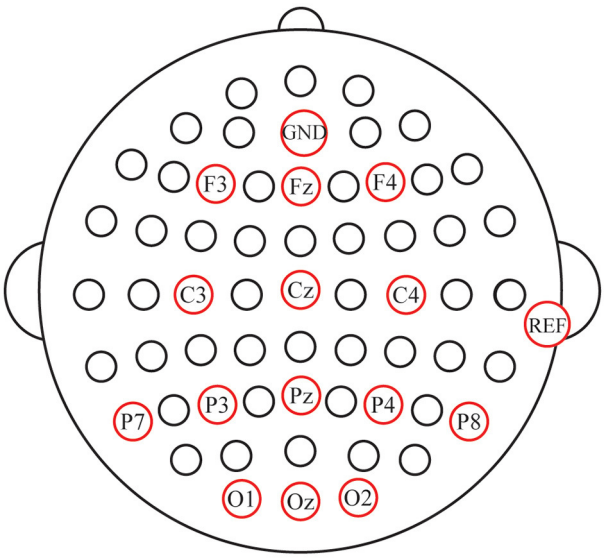


**FIGURE 3 |** Configuration of electrode positions.

virtual matrix, the characters were randomly rearranged into a new matrix in which the characters of the same row or the same column in the traditional matrix were positioned as far away as possible. Therefore, the rows or columns were six random characters in the actual matrix (**Figure 1**, left), which mitigates the problem of adjacency flashing (Townsend et al., 2010). The rows and columns of the virtual matrix flashed consecutively in a pseudo-random order. The stimulus onset asynchrony of each paradigm was set to 250 ms, in which each character was covered with a picture of a face for 200 ms and then reverted to a gray character for 50 ms.

## Procedure

Each subject sat in a comfortable chair, ∼70 cm from the front of the computer monitor, in a shielded room. During data acquisition, subjects were asked to relax and avoid unnecessary movement. The subjects' task was to focus on the target character and silently count the number of times the target characters were covered with faces during stimulus presentation.

In the offline experiment, one flash of a row or column was referred to as a sub-trial. The flash of a row or column that included the target character was defined as a target sub-trial, and the flash of a row or column without the target character was defined as a non-target sub-trial. Six rows and six columns flashed once (12 flashes) as a trial, and the trial was repeated 15 times as a sequence. Thus, each sequence consisted of 180 flashes of rows or columns to output a target character. During the experiment, each spelling paradigm was conducted four times, and each time, a five-character word was spelled out, which was

considered a run (**Figure 2**). The runs of the two paradigms were counted alternately to control for potential habituation effects. Participants were allowed to take a 5-min break between runs.

In the online experiment, each subject completed training and testing phases for the famous face and self-face spelling paradigms. In the training phase, there were four runs, and each run contained 20 sequences (whereby one character was revealed per sequence); that is, there were 20 characters in a run and a total of 80 characters in the training phase for each spelling paradigm, which were used to obtain the classifier. The test phase output a total of 30 characters by the trained classifier. In addition, trials were only repeated twice in each sequence for both the training and testing phases.

## Data Acquisition

EEG signals were recorded with a NeuroScan amplifier (SynAmps 2, NeuroScan Inc., and Abbotsford, Australia). All signals were digitized at a rate of 250 Hz, and band-pass filtered between 0.1 and 100 Hz. Fourteen-channel (Fz, F3, F4, C3, Cz, C4, P7, P8, P3, P4, Pz, O1, Oz, and O2, **Figure 3**) EEG data were recorded with the AFz as the ground and the right mastoid as the reference electrode position. Horizontal eye movements were measured by deriving the electrooculogram (EOG) from a pair of horizontal EOG (HEOG) electrodes placed at the outer canthi of both the left and right eyes. Vertical eye movements and eye blinks were detected by deriving an EOG signal from a pair of vertical EOG (VEOG) electrodes placed ∼1 cm above and below the subject's left eye. The impedance was maintained below 5 KΩ.

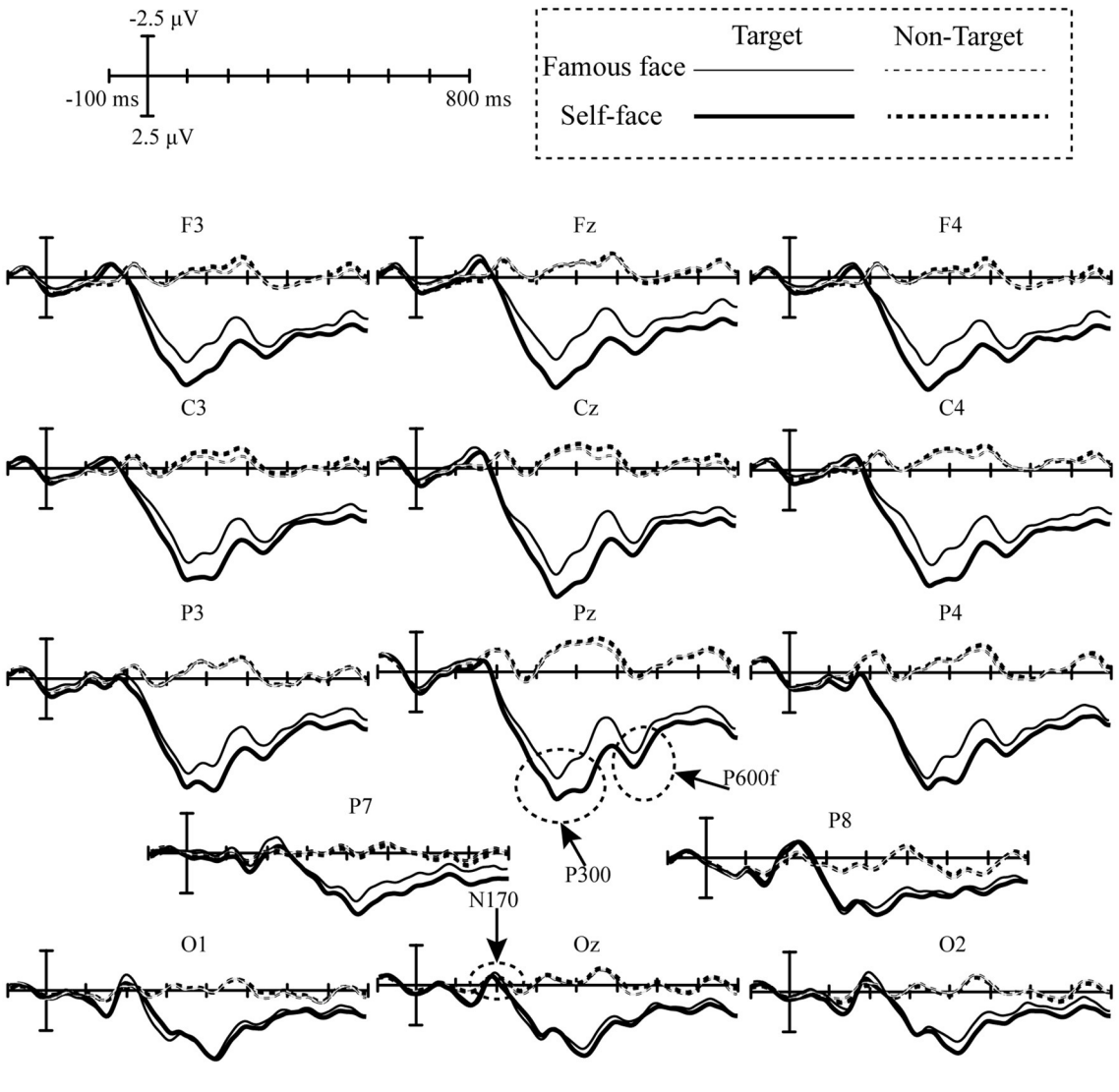


**FIGURE 4 |** Superimposed grand-averaged event-related potentials elicited by the target and non-target stimuli over 14 electrodes in the self-face and famous face spelling paradigms.

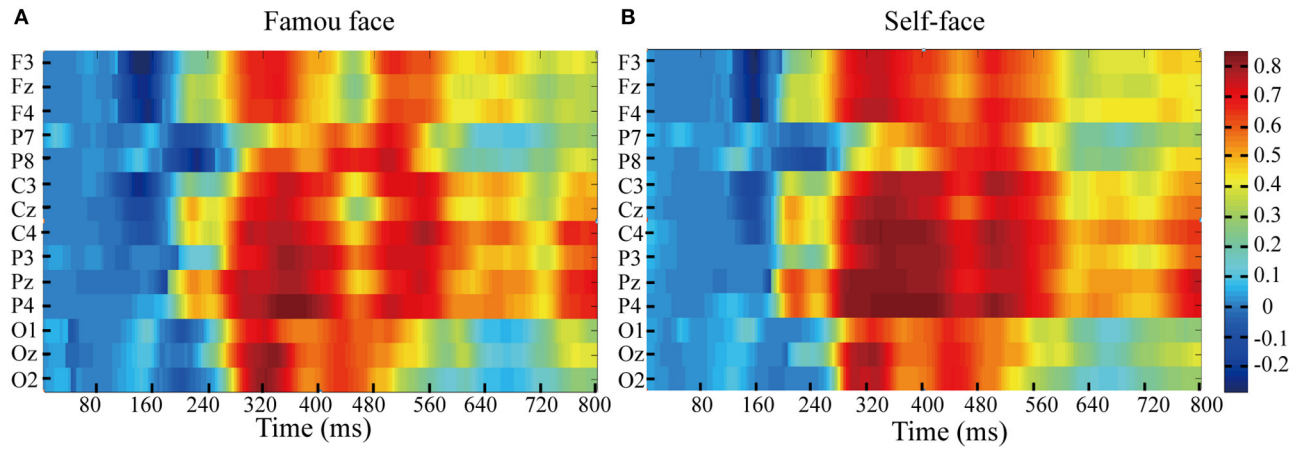


**FIGURE 5 |** $R^2$ values of ERPs in response to the target and non-target stimuli between 0 and 800 ms from EEG data of all subjects in the famous face and self-face spelling paradigms. **(A)** $R^2$ values of ERPs for the famous face spelling paradigm. **(B)** $R^2$ values of ERPs for the self-face spelling paradigm.

## Feature Extraction Procedure

For offline data, the classification performance of the speller depends not only on the amplitude of ERPs elicited by the target stimulus but also on the difference in ERP amplitudes elicited by the target and non-target stimuli. Thus, the analysis of $R^2$ values can provide the mathematic foundation for selecting channels and the features of each channel. The r-squared is calculated by formula (1)

$$r^2 = \left( \frac{\sqrt{N_1 N_2}}{N_1 + N_2} \times \frac{mean(x_1) - mean(x_2)}{std(x_1 \bigcup x_2)} \right)^2 \quad (1)$$

where $N_1$ and $N_2$ represent the sample size of the target and non-target stimuli, respectively; $x_1$and $x_2$ are features vector of the target and non-target stimuli, respectively.

According to the results of the r-squared values, ERP data of different time windows were down-sampled from 250 to 62.5 Hz by selecting every four samples, and the feature vector was $N_p \times N_c$, where $N_p$ represents the sample points within the selected time window, and $N_c$ represents the number of channels. For online data, the EEG data were first filtered between 0.1 and 30 Hz using a third-order Butterworth bandpass filter, then down-sampled from 250 to 50 Hz. We extracted the EEG data from 200 to 800 ms after stimuli onset as the vector feature.

## Classification Scheme

Bayesian linear discriminant analysis (BLDA) was used to classify the EEG data in the experiment. BLDA is an extension of Fisher's linear discriminant analysis that avoids over-fitting. The details of the algorithm have been described elsewhere (Hoffmann et al., 2008; Jin et al., 2014a). We used 4-fold cross-validation to calculate the individual accuracy in the offline experiment.

## Information Transfer Rate

Information transfer rate (ITR) is generally used to evaluate the communication performance of a BCI system and is a standard measure that accounts for accuracy, the number of possible selections, and the time required to make each selection (Thompson et al., 2013). The ITR (bits $min^{-1}$) can be calculated as follows:

$$ITR = \frac{60(P \log_2(P) + (1-P) \log_2 \frac{1-P}{N-1} + \log_2 N)}{T} \quad (2)$$

where $P$ denotes the probability of recognizing a character, $T$ is the time taken to recognize a character, and $N$ is the number of classes ($N = 36$).

## Data Analysis

A one-way repeated measure ANOVA with the within-subjects two factors of spelling paradigm (self-face and famous face spelling paradigms) and electrodes (electrodes were based on the waveform of ERPs elicited by target stimuli) was used to compare the difference in ERP amplitudes between self-face and famous face spelling paradigms acquired by subtracting the waveforms elicited by non-target stimuli from that by target stimuli. The comparison of classification accuracy and ITR in offline and online experiments was conducted by a paired $T$-test. The statistical analyses were conducted using the SPSS version 19.0 software package (SPSS Inc., Chicago, IL, USA).

# RESULTS

## ERP Results

**Figure 4** displays the superimposed grand-averaged waveform elicited by target and non-target stimuli in the self-face and famous face spelling paradigms. A clear negative peak was observed at O1, Oz, and Oz between 150 and 200 ms, which is indicative of the N170 potential. In addition, we observed a clear positive peak at all electrodes between 200 and 500 ms, which is indicative of the P300 potential, and the other positive peak was observed between 500 and 600 ms, at F3, Fz, F4, C3, Cz, C4, P3, Pz, and P4, which is similar to the P600f potential.

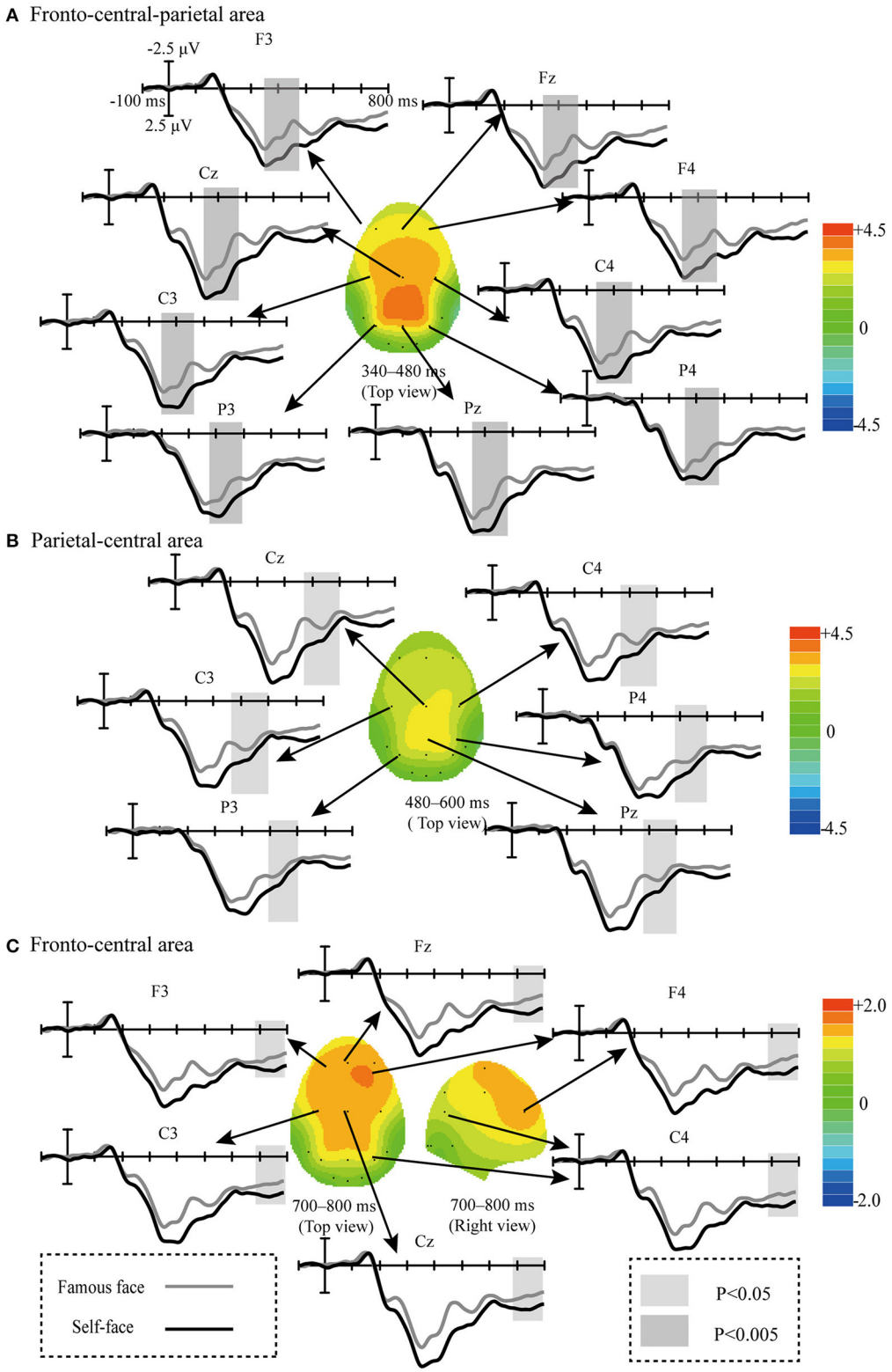


**FIGURE 6 |** Comparison of waveforms ($ERP_{Target} - ERP_{Non-target}$) elicited by the target and non-target stimuli in the self-face and famous face spelling paradigms, and scalp topographies from difference waveforms. Difference waveforms were calculated by subtracting the ERPs of the famous face spelling paradigm from those of the self-face spelling paradigm. **(A)** The fronto-central-parietal area at 340–480 ms. **(B)** The parietal-central area at 480–600 ms. **(C)** The fronto-central area at 700–800 ms.

Feature differences in the ERPs elicited by target and non-target stimuli in the famous face and self-face spelling paradigms were indicated by the r-squared values (**Figure 5**). As seen in **Figure 5**, we observed that the feature differences in the ERPs elicited by target and no-target stimuli were mainly between 200 and 800 ms at all electrodes for both the famous face and self-face spelling paradigms. To represent the positive and negative deflections of ERP amplitude and to allow for richer visual information, we set the $R^2$ value corresponding to the negative ERP amplitude value as a negative value.

**Figure 6** displays the scalp topographic regions that corresponded to significant differences between the waveforms elicited in the self-face and famous face spelling paradigms. Significant differences were observed in three regions corresponding to three time periods after stimulus presentation, as follows: the fronto-central-parietal area from 340 to 480 ms [$F_{(1,9)} = 14.54$, $P < 0.005$; **Figure 6A**]; the parietal-central area from 480 to 600 ms [$F_{(1,9)} = 8.018$, $P < 0.05$; **Figure 6B**]; and the fronto-central area from 700 to 800 ms [$F_{(1,9)} = 6.023$, $P < 0.05$; **Figure 6C**].

## Classification Results

Based on the results of the r-squared values, we compared the classification accuracies based on two feature vectors, as follows: the feature vector A was 25 × 12 (time window of 200–700 ms, channels F3, Fz, F4, C3, Cz, C4, P3, Pz, P4, O1, Oz, and O4); the feature vector B was 45 × 14 (time window of 0–800 ms, 14 channels). The results of classification accuracies based on feature A and feature B are shown in **Figure 7**, which shows the average accuracies across all subjects at each sequence in famous face and self-face spelling paradigms. There was no significant difference in accuracy between feature A and feature B in the two spelling paradigms.

Previous work has shown that the frequency band for the P300 is mainly between 1 and 10 Hz (Basar-Eroglu et al., 1992) and different band passes have been used to filter EEG data to acquire better classification accuracy, such as 1–4, 1–12, and 1–30 Hz (Jin et al., 2017). In this study, we compared the classification accuracies at the first three superpositions (superposition times represent the number of trials, that is, the repeating times of 6 rows/columns flashing) between 1–4, 1–12, and 1–30 Hz for the famous face and self-face spelling paradigms (**Figure 8**). We found that the average accuracy at 1–12 Hz was larger than that at 1–4 Hz, and the average accuracy at 1–30 Hz was larger than that at 1–4/1–12 Hz for the first three superpositions in the two spelling paradigms except for the accuracies between 1–12 and 1–30 Hz at two superpositions in the famous face spelling paradigm. The paired $t$-test results revealed a significant difference for classification accuracy between 1–4 and 1–12 Hz/1–30 Hz in the famous face and self-face spelling paradigms.

**Figure 9** shows the individual and average offline accuracies in the two face spelling paradigms based on the feature vector B and a 1–30 Hz frequency band filter. The accuracies increased with the increase in the number of superpositions in both paradigms; the average spelling accuracy of the self-face spelling paradigm was greater than that in the famous face spelling paradigm at 1–15 superpositions. The average number of superpositions when the accuracies reached 100% for all subjects was 2 in the self-face spelling paradigm; thus, we conducted a $t$-test on the accuracies only for the first two superpositions between the self-face and famous face paradigms. We found significant differences between the self-face and famous face spelling paradigms at both one superposition ($t = -2.331$, $P < 0.05$; **Figure 10A**) and two superpositions ($t = -2.25$, $P < 0.05$; **Figure 10B**).

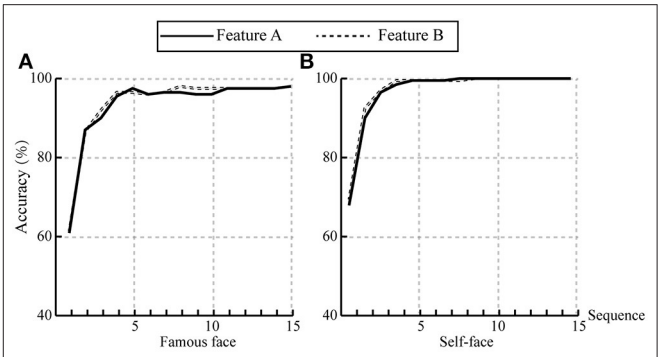


**FIGURE 7 |** The comparison of classification accuracies based on feature vector A and feature vector B. **(A)** The average accuracies across all subjects in the famous face spelling paradigm. **(B)** The average accuracies across all subjects in the self-face spelling paradigm.

**Table 1** shows the ITRs for each subject and the averages in the self-face and famous face spelling paradigms. The best ITR result, 31.4 bits $min^{-1}$ at one superposition, was found with the self-face spelling paradigm. The average ITR was greater at two superpositions than at one superposition. The paired $t$-tests showed that the ITR was significantly greater in the self-face paradigm than in the famous face paradigm at one superposition ($t = -2.414$, $P = 0.039 < 0.05$) and two superpositions ($t = -2.345$, $P = 0.044 < 0.05$).

The online accuracies and ITRs of each subject for the famous face and self-face spelling paradigms are shown in **Table 2**. We found that the average accuracy and ITR in the self-face spelling paradigm were higher than those in the famous face spelling paradigm. Paired $t$-tests showed that there were significant differences in the accuracy and ITR between the two spelling paradigms (accuracy: $t = -2.643$, $P < 0.05$; ITR: $t = -3.140$, $P < 0.05$).

# DISCUSSION

In the present study, we proposed a new P300-speller using self-face stimulus and assessed the grand-average ERP waveforms elicited by target stimuli in the new and control spelling paradigms, analyzed the different ERP waveforms and the scalp topographies corresponding to significantly different waveforms elicited by the target minus non-target stimuli, and compared the classification accuracy and ITR of offline and online experiments between the self-face and famous face spelling paradigms.

## ERPs

Previous work has found that the performance of the P300-speller system could be improved by enhancing the

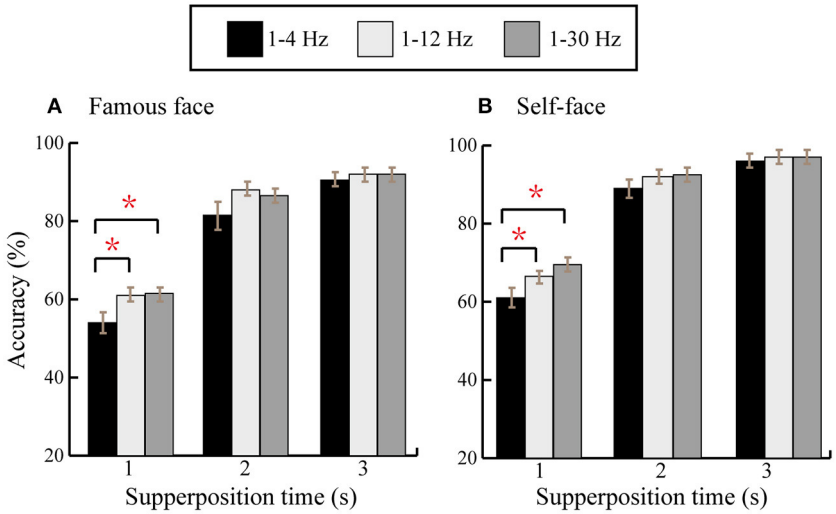


**FIGURE 8 |** Average offline classification accuracies across all subjects at the first three superpositions for 1–4, 1–12, and 1–30 Hz. **(A)** The comparison of accuracies between three frequency band filters in the famous face spelling paradigm. **(B)** The comparison of accuracies between three frequency band filters in the self-face spelling paradigm. *A significant difference in accuracy between two frequency bands.

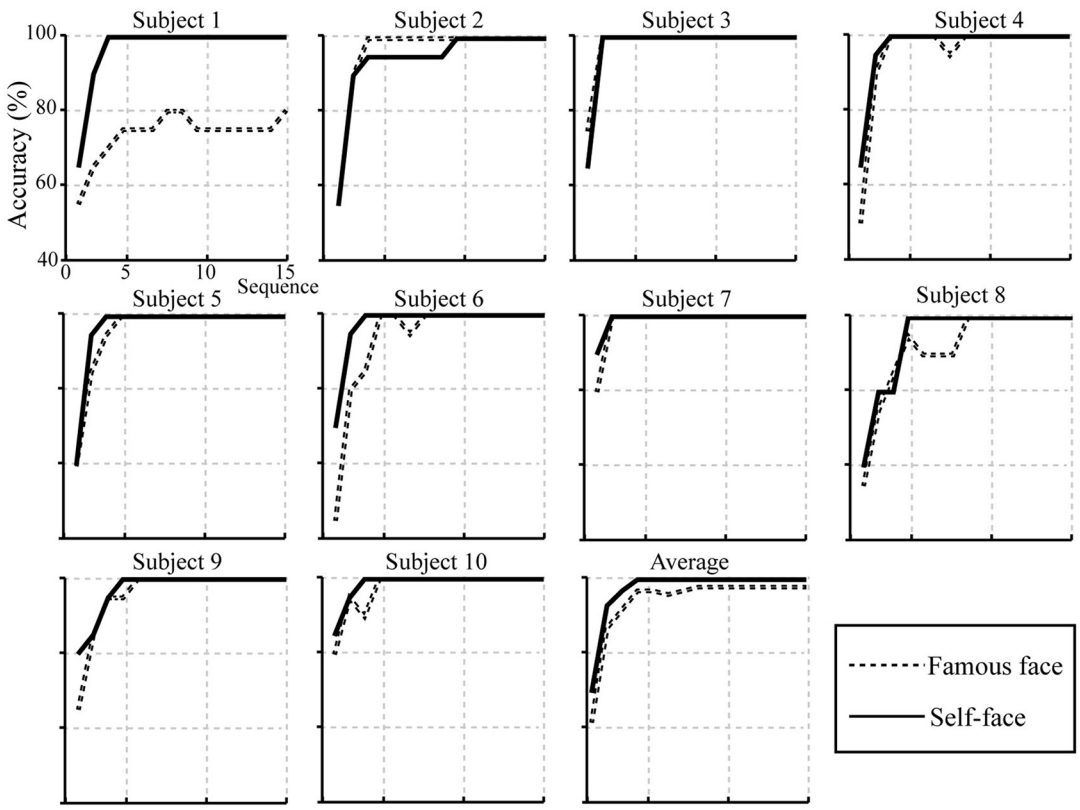


**FIGURE 9 |** Individual and average accuracies of the self-face and famous face spelling paradigms for 10 subjects.

difference between target trials and non-target trials (Jin et al., 2012). Therefore, we compared the waveforms ($ERP_{Target}$ – $ERP_{Non-target}$) elicited during the two face paradigms and found a significant difference between the two. The first significantly different waveform was from 340 to 480 ms over the fronto-central-parietal area (**Figure 6**), i.e., the P300. The P300 is not only associated with attention and cognitive processing (Polich, 2007) but also reflects the involvement of higher-order cognitive

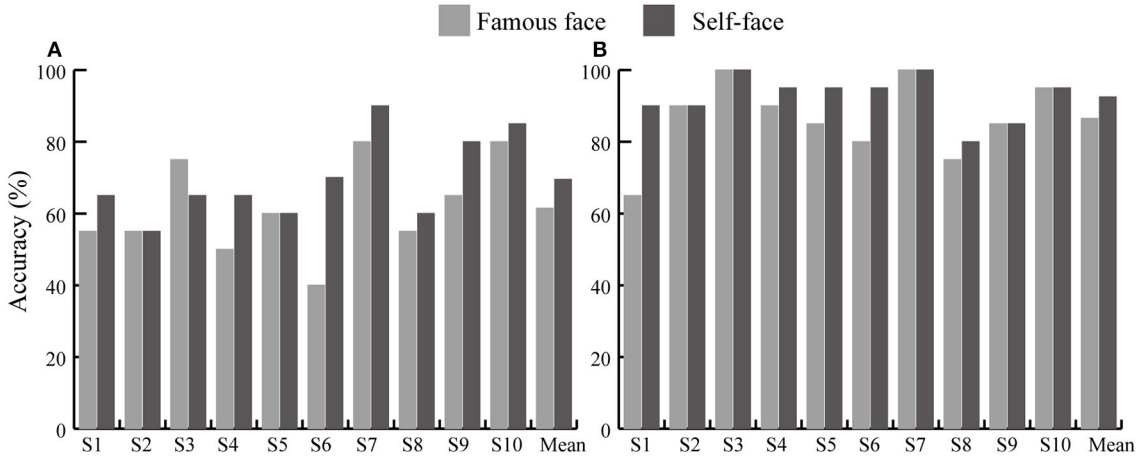


**FIGURE 10 |** Accuracies of each subject and mean accuracy of 10 subjects at one superposition and two superpositions for the famous face and self-face spelling paradigms. **(A)** Accuracies at one superposition. **(B)** Accuracies at two superpositions.

functions, including self-relevance (for one's own face, e.g., Ninomiya et al., 1998; Tanaka et al., 2006). Ninomiya et al. (1998) found that the P300 amplitude in response to one's own face was significantly larger than that in response to other stimuli. The authors, therefore, suggested that enhancement of the P300 in response to one's own face is not only due to an orienting response to a physically deviant stimulus but also due to the additional effect of relevance to the subject. Thus, the P300 can serve as an index of self-relevance, whereby higher self-relevance corresponds to a larger P300 amplitude (Kok, 2001). In Miyakoshi et al.'s study, the P300 amplitude elicited by the self-face stimulus was greater than that elicited by a famous face, and the P300 could distinguish the self-face from a famous face, and the authors, therefore, suggested that the P300 amplitude was sensitive to self-relevance (Miyakoshi et al., 2008). Therefore, the larger amplitude P300 in the self-face spelling paradigm than in the famous face spelling paradigm may be due to the higher self-relevance of the self-face than of the famous face for subjects.

The second significant difference in positive waveform was observed from 480 to 600 ms at the parietal-central area (**Figure 6**); this was similar to the P600f, which is related to processes involved in the recollection of faces (Eimer, 2000; Curran and Hancock, 2007). Some studies have suggested that perception of an individual's face may induce spontaneous activation of the characteristic and information associated with the individual (Bargh et al., 1996; Todorov and Uleman, 2002). The ERPs between 500 and 700 ms with a larger amplitude in response to a familiar face as compared to an unfamiliar face may indicate that the perception of the familiar face automatically generated more of one's personal traits or other episodic information than the perception of an unfamiliar face (Sui et al., 2006). Curran and Hancock (2007) also reported that a familiar face elicited a larger positive waveform between 500 and 700 ms (P600f) than did a stranger's face. Thus, we speculate that the larger P600f amplitude observed in the self-face spelling paradigm than in the famous face spelling paradigm indicates that the self-face induced more recollection, including characteristic or episodic information about the self than did the famous face.

The third significant difference in positive waveforms was from 700 to 800 ms at the fronto-central area (**Figure 6**). In ERP studies of face recognition, attending to the self-face induced a larger amplitude waveform between 600 and 800 ms at the prefronto-central area than did attending to a familiar face; it was speculated that this component was affected by the allocation of attentional resources in face recognition (Sui et al., 2006). Miyakoshi et al. (2008) found that the self-face was more likely to attract the attention of participants than a familiar face. In our study, the increased amplitude between 700 and 800 ms for the self-face than for the famous face paradigm may indicate that subjects paid more attention to their own faces.

In addition, our results showed that there was no significant difference in the N170 amplitude between the two spelling paradigms. This may be due to differences in experimental design (Keyes et al., 2010; Alonso-Prieto et al., 2015). Alonso-Prieto et al. (2015) reported that the sensitivity of the N170 to faces with different levels of familiarity is affected by the experimental settings, such as faces with different facial angles or faces with emotional information. For example, there was a difference in the N170 between a famous face and the self-face in studies of the influence of facial angle (Miyakoshi et al., 2008) and of emotional expression (Caharel et al., 2005), while Tacikowski et al. (2011) found no difference in the N170 amplitude between the self-face and a famous face when using frontal and neutral face images. In our study, the famous face and self-face comprised frontal and neutral images; thus, our results are consistent with those of Tacikowski et al. In addition, the type of familiarity of the face has also been found to affect the sensitivity of the N170 (Alonso-Prieto et al., 2015). For example, Sui et al. (2006) found that the N170 did not differ between self-faces and familiar faces (classmates), while Keyes et al. (2010) showed an increased N170 amplitude to the self-face relative to familiar faces (good friends).

**TABLE 1 |** The information transfer rate of each subject for the famous face and self-face spelling paradigms at one and two superpositions.

| | One superposition | | Two superpositions | |
|---|---|---|---|---|
| **Subject** | **Famous face** | **Self-face** | **Famous face** | **Self-face** |
| Subject 1 | 14.0 | 18.3 | 13.3 | 22.8 |
| Subject 2 | 14.0 | 14.0 | 22.8 | 22.8 |
| Subject 3 | 23.1 | 18.3 | 27.5 | 27.5 |
| Subject 4 | 12.0 | 18.3 | 22.8 | 25.2 |
| Subject 5 | 16.1 | 16.1 | 20.7 | 25.2 |
| Subject 6 | 8.4 | 20.6 | 18.7 | 25.2 |
| Subject 7 | 25.7 | 31.4 | 27.5 | 27.5 |
| Subject 8 | 14.0 | 16.1 | 16.8 | 18.7 |
| Subject 9 | 18.3 | 25.7 | 20.7 | 20.7 |
| Subject 10 | 25.7 | 28.4 | 25.2 | 25.2 |
| Avg. ± SD | 17.1 ± 5.9 | 20.7 ± 5.8 | 21.6 ± 4.6 | 24.1 ± 2.8 |
| *p*-value | $t = -2.414$; $p = 0.039$ | | $t = -2.345$; $p = 0.044$ | |

*The unit of information transfer rate is bit/min.*

**TABLE 2 |** The online accuracies and ITRs for all subjects in the famous face and self-face spelling paradigms.

| | Accuracies (%) | | ITRs (bit/min) | |
|---|---|---|---|---|
| **Subject** | **Famous face** | **Self-face** | **Famous face** | **Self-face** |
| Subject 1 | 96.7 | 100.0 | 31.9 | 33.6 |
| Subject 2 | 80.0 | 93.3 | 22.8 | 29.8 |
| Subject 3 | 60.0 | 66.7 | 14.3 | 17.0 |
| Subject 4 | 70.0 | 76.7 | 18.3 | 21.3 |
| Subject 5 | 73.3 | 83.3 | 19.8 | 24.4 |
| Subject 6 | 80.0 | 73.3 | 22.8 | 19.8 |
| Subject 7 | 86.7 | 93.3 | 26.1 | 29.8 |
| Subject 8 | 90.0 | 96.7 | 27.9 | 31.9 |
| Subject 9 | 83.0 | 90.0 | 24.3 | 27.9 |
| Subject 10 | 80.0 | 80.0 | 22.8 | 22.8 |
| Avg. ± SD | 80.0 ± 10.4 | 85.3 ± 11.0 | 23.1 ± 5.0 | 25.8 ± 5.4 |
| *p*-value | $t = -2.643$, $P < 0.05$ | | $t = -3.140$, $P < 0.05$ | |

In the present study, the reason we found no difference in the N170 between the two paradigms may be that the difference in familiarity level between the famous face (Ming Yao) and the self-face may not have been enough to induce a statistically significant difference in N170 amplitude.

## Classification Accuracies and ITR

Offline classification results showed that the average accuracies of the self-face spelling paradigm were higher than those of the famous face spelling paradigm at all numbers of superpositions (**Figure 9**). A significant difference was found between the self-face and famous face spelling paradigm at one superposition ($P < 0.05$; **Figure 10A**) and at two superpositions ($P < 0.05$; **Figure 10B**). The offline accuracies demonstrated that use of the self-face improved the performance of the facial spelling paradigm because the self-face stimulus induced larger ERP components than did the famous face. In addition, the improvement and stability of spelling accuracy required stimuli to be repeated several times because of the low signal-to-noise ratios; however, increasing the number of repetitions may reduce the spelling speed. Thus, the ITR depended on both classification accuracy and speed character output, which is an important statistical metric for the performance of the P300-speller system. Our results indicated that the ITR of the self-face spelling paradigm was significantly greater than that of the famous face spelling paradigm at the first two superpositions ($P < 0.05$). The best result, 31.4 bits $min^{-1}$ for subject 7, was obtained with the self-face spelling paradigm, in which subject 7 achieved 90% accuracy with one superposition only. Yet, the average ITR at two superpositions was larger than that at one superposition, and the standard deviation at one superposition was greater than that at two superpositions in both spelling paradigms (**Table 1**). This indicated that the spelling stability and performance is better at two superpositions. Therefore, in the online experiment, we set the trial to repeat only twice (that is, two superposition for 6 rows/columns) to acquire the accuracies and ITRs of character spelling. The online results showed that accuracy and ITR of the self-face spelling paradigm were significantly larger than those of the famous face spelling paradigm (**Table 2**). In summary, the proposed self-face spelling paradigm significantly improved the performance of the P300-speller system.

In addition, we compared the offline classification accuracies based on different feature vectors and frequency band passes. For feature vector A (25 × 12) and feature vector B (45 × 14), there was no significant difference at all superposition times, which indicates that the feature vector from amplitude difference between target and non-target stimuli can acquire classification results that are comparable to the feature vector in the 0–800 ms time window and at all channels (**Figure 7**). The classification results based on three frequency band passes showed that the best classification result was at 1–30 Hz at first three superpositions in both spelling paradigms (**Figure 8**), which indicated that a filter of 1–30 Hz could be a good choice for the classification accuracy of the P300-speller system.

## Future Work

The analysis of ERPs, classification accuracies, and ITRs between the two spelling paradigms showed that the self-face stimulus elicited significantly increased ERP amplitudes compared to the famous face stimulus and improved the spelling accuracy and ITR of the P300-speller system. Moreover, the use of self-face also avoided the copyright issues caused by using a famous face. Thus, the proposed self-face paradigm promotes practical applications of BCIs system. Some recent studies have shown that the brain responded more positively to a happy face and which could elicit increased ERP amplitudes, compared to a neutral face stimulus (Denefrio et al., 2017; Lu et al., 2019). In future work, we intend to use the subject's own happy face to investigate whether the self-face with happy emotion can further improve the performance and practicability of the P300-speller system.

## CONCLUSION

This study investigated whether the use of the self-face could improve the performance of the P300-speller system as compared to the use of a famous face. We found a significant improvement in classification accuracy and ITR for the self-face spelling paradigm at the first two superpositions, as compared to the famous face spelling paradigm, which may have a significant impact on increasing the speed and accuracy of spelling. Moreover, this has significance in practical BCI applications because the use of a famous face may involve copyright infringement problems.

## DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/supplementary material.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics committee of Changchun University of Science and Technology. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

## AUTHOR CONTRIBUTIONS

QL and ZL designed the experiment, wrote the manuscript, and revised the manuscript. ZL and NG implemented the experiment and accomplished the data processing. ZL and JY analyzed the experimental results and revised the manuscript. All authors read and approved the final manuscript.

## FUNDING

This work was financially supported by the National Natural Science Foundation of China (grant numbers 61806025 and 61773076), Jilin Scientific and Technological Development Program (grant numbers 20190302072GX and 20180519012JH), and Scientific Research Project of Jilin Provincial Department of Education during the 13th Five-Year Plan Period (grant number JJKH20190597KJ).

## ACKNOWLEDGMENTS

The authors further wish to thank all individuals who participated in our study.

## REFERENCES

Allison, B. Z., and Pineda, J. A. (2003). ERPs evoked by different matrix sizes: implications for a brain computer interface (BCI) system. *IEEE Trans. Neural Syst. Rehabil. Eng.* 11, 110–113. doi: 10.1109/TNSRE.2003.814448

Allison, B. Z., and Pineda, J. A. (2006). Effects of SOA and flash pattern manipulations on ERPs, performance, and preference: implications for a BCI system. *Int. J. Psychophysiol.* 59, 127–140. doi: 10.1016/j.ijpsycho.2005.02.007

Alonso-Prieto, E., Pancaroglu, R., Dalrymple, K. A., Handy, T., Barton, J. J., and Oruc, I. (2015). Temporal dynamics of the face familiarity effect: bootstrap analysis of single-subject event-related potential data. *Cogn. Neuropsychol.* 32, 266–282. doi: 10.1080/02643294.2015.1053852

Bargh, J. A., Chen, M., and Burrows, L. (1996). Automaticity of social behavior: direct effects of trait construct and stereotype-activation on action. *J. Pers. Soc. Psychol.* 71, 230–244. doi: 10.1037/0022-3514.71.2.230

Basar-Eroglu, C., Basar, E., Demiralp, T., and Schurmann, M. (1992). P300-response: possible psychophysiological correlates in delta and theta frequency channels. A review. *Int. J. Psychophysiol.* 13, 161–179. doi: 10.1016/0167-8760(92)90055-G

Bentin, S., and Deouell, L. Y. (2000). Structural encoding and identification in face processing: erp evidence for separate mechanisms. *Cogn. Neuropsychol.* 17, 35–55. doi: 10.1080/026432900380472

Bernat, E., Shevrin, H., and Snodgrass, M. (2001). Subliminal visual oddball stimuli evoke a P300 component. *Clin. Neurophysiol.* 112, 159–171. doi: 10.1016/S1388-2457(00)00445-4

Caharel, S., Courtay, N., Bernard, C., Lalonde, R., and Rebai, M. (2005). Familiarity and emotional expression influence an early stage of face processing: an electrophysiological study. *Brain Cogn.* 59, 96–100. doi: 10.1016/j.bandc.2005.05.005

Carbon, C. C., Schweinberger, S. R., Kaufmann, J. M., and Leder, H. (2005). The Thatcher illusion seen by the brain: an event-related brain potentials study. *Brain Res. Cogn. Brain Res.* 24, 544–555. doi: 10.1016/j.cogbrainres.2005.03.008

Carelli, L., Solca, F., Faini, A., Meriggi, P., Sangalli, D., Cipresso, P., et al. (2017). Brain-computer interface for clinical purposes: cognitive assessment and rehabilitation. *Biomed. Res. Int.* 2017:1695290. doi: 10.1155/2017/1695290

Curran, T., and Hancock, J. (2007). The FN400 indexes familiarity-based recognition of faces. *Neuroimage* 36, 464–471. doi: 10.1016/j.neuroimage.2006.12.016

Denefrio, S., Simmons, A., Jha, A., and Dennis-Tiwary, T. A. (2017). Emotional cue validity effects: The role of neurocognitive responses to emotion. *PLoS ONE* 12:e0179714. doi: 10.1371/journal.pone.0179714

Eimer, M. (2000). Event-related brain potentials distinguish processing stages involved in face perception and recognition. *Clin. Neurophysiol.* 111, 694–705. doi: 10.1016/S1388-2457(99)00285-0

Farwell, L. A., and Donchin, E. (1988). Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalogr. Clin. Neurophysiol.* 70, 510–523. doi: 10.1016/0013-4694(88)90149-6

Herzmann, G., Schweinberger, S. R., Sommer, W., and Jentzsch, I. (2004). What's special about personally familiar faces? A multimodal approach. *Psychophysiology* 41, 688–701. doi: 10.1111/j.1469-8986.2004.00196.x

Hoffmann, U., Vesin, J. M., Ebrahimi, T., and Diserens, K. (2008). An efficient P300-based brain-computer interface for disabled subjects. *J. Neurosci. Methods* 167, 115–125. doi: 10.1016/j.jneumeth.2007.03.005

Jeunet, C., Lotte, F., Batail, J. M., Philip, P., and Franchi, J. A. M. (2018). Using recent BCI literature to deepen our understanding of clinical neurofeedback: a short review. *Neuroscience* 378, 225–233. doi: 10.1016/j.neuroscience.2018.03.013

Jin, J., Allison, B. Z., Kaufmann, T., Kubler, A., Zhang, Y., Wang, X., et al. (2012). The changing face of P300 BCIs: a comparison of stimulus changes in a P300 BCI involving faces, emotion, and movement. *PLoS ONE* 7:e49688. doi: 10.1371/journal.pone.0049688

Jin, J., Allison, B. Z., Zhang, Y., Wang, X., and Cichocki, A. (2014a). An ERP-based BCI using an oddball paradigm with different faces and reduced errors in critical functions. *Int. J. Neural Syst.* 24:1450027. doi: 10.1142/S0129065714500270

Jin, J., Daly, I., Zhang, Y., Wang, X., and Cichocki, A. (2014b). An optimized ERP brain-computer interface based on facial expression changes. *J. Neural. Eng.* 11:036004. doi: 10.1088/1741-2560/11/3/036004

Jin, J., Zhang, H., Daly, I., Wang, X., and Cichocki, A. (2017). An improved P300 pattern in BCI to catch user's attention. *J. Neural Eng.* 14:036001. doi: 10.1088/1741-2552/aa6213

Kaufmann, T., Schulz, S. M., Grunzinger, C., and Kubler, A. (2011). Flashing characters with famous faces improves ERP-based brain-computer interface performance. *J. Neural Eng.* 8:056016. doi: 10.1088/1741-2560/8/5/056016

Keyes, H., Brady, N., Reilly, R. B., and Foxe, J. J. (2010). My face or yours? Event-related potential correlates of self-face processing. *Brain Cogn.* 72, 244–254. doi: 10.1016/j.bandc.2009.09.006

Kircher, T. T., Senior, C., Phillips, M. L., Rabe-Hesketh, S., Benson, P. J., Bullmore, E. T., et al. (2001). Recognizing one's own face. *Cognition* 78, B1–B15. doi: 10.1016/S0010-0277(00)00104-9

Kok, A. (2001). On the utility of P3 amplitude as a measure of processing capacity. *Psychophysiology* 38, 557–577. doi: 10.1017/S0048577201990559

Lazarou, I., Nikolopoulos, S., Petrantonakis, P. C., Kompatsiaris, I., and Tsolaki, M. (2018). EEG-based brain-computer interfaces for communication and rehabilitation of people with motor impairment: a novel approach of the 21 (st) century. *Front. Hum. Neurosci.* 12:14. doi: 10.3389/fnhum.2018.00014

Li, Q., Liu, S., Li, J., and Bai, O. (2015). Use of a green familiar faces paradigm improves P300-speller brain-computer interface performance. *PLoS ONE* 10:e0130325. doi: 10.1371/journal.pone.0130325

Li, Q., Lu, Z., Gao, N., and Yang, J. (2019). Optimizing the performance of the visual P300-speller through active mental tasks based on color distinction and modulation of task difficulty. *Front. Hum. Neurosci.* 13:130. doi: 10.3389/fnhum.2019.00130

Lu, Z., Li, Q., Gao, N., Yang, J., and Bai, O. (2019). Happy emotion cognition of bimodal audiovisual stimuli optimizes the performance of the P300 speller. *Brain Behav.* 9:e01479. doi: 10.1002/brb3.1479

Miyakoshi, M., Kanayama, N., Nomura, M., Iidaka, T., and Ohira, H. (2008). ERP study of viewpoint-independence in familiar-face recognition. *Int. J. Psychophysiol.* 69, 119–126. doi: 10.1016/j.ijpsycho.2008.03.009

Ninomiya, H., Onitsuka, T., Chen, C. H., Sato, E., and Tashiro, N. (1998). P300 in response to the subject's own face. *Psychiatry Clin. Neurosci.* 52, 519–522. doi: 10.1046/j.1440-1819.1998.00445.x

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clin. Neurophysiol.* 118, 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Rosenfeld, J. V., and Wong, Y. T. (2017). Neurobionics and the brain-computer interface: current applications and future horizons. *Med. J. Aust.* 206, 363–368. doi: 10.5694/mja16.01011

Salvaris, M., and Sepulveda, F. (2009). Visual modifications on the P300 speller BCI paradigm. *J. Neural Eng.* 6:046011. doi: 10.1088/1741-2560/6/4/046011

Schweinberger, S. R., Pickering, E. C., Jentzsch, I., Burton, A. M., and Kaufmann, J. M. (2002). Event-related brain potential evidence for a response of inferior temporal cortex to familiar face repetitions. *Brain Res. Cogn. Brain Res.* 14, 398–409. doi: 10.1016/S0926-6410(02)00142-8

Sellers, E. W., Krusienski, D. J., McFarland, D. J., Vaughan, T. M., and Wolpaw, J. R. (2006). A P300 event-related potential brain-computer interface (BCI): the effects of matrix size and inter stimulus interval on performance. *Biol. Psychol.* 73, 242–252. doi: 10.1016/j.biopsycho.2006.04.007

Speier, W., Deshpande, A., Cui, L., Chandravadia, N., Roberts, D., and Pouratian, N. (2017). A comparison of stimulus types in online classification of the P300 speller using language models. *PLoS ONE* 12:e0175382. doi: 10.1371/journal.pone.0175382

Sui, J., Zhu, Y., and Han, S. (2006). Self-face recognition in attended and unattended conditions: an event-related brain potential study. *Neuroreport* 17, 423–427. doi: 10.1097/01.wnr.0000203357.65190.61

Tacikowski, P., Jednorog, K., Marchewka, A., and Nowicka, A. (2011). How multiple repetitions influence the processing of self-, famous and unknown names and faces: An ERP study. *Int. J. Psychophysiol.* 79, 219–230. doi: 10.1016/j.ijpsycho.2010.10.010

Takeuchi, N., Mori, T., Nishijima, K., Kondo, T., and Izumi, S. (2015). Inhibitory transcranial direct current stimulation enhances weak beta event-related synchronization after foot motor imagery in patients with lower limb amputation. *J. Clin. Neurophysiol.* 32, 44–50. doi: 10.1097/WNP.0000000000000123

Tanaka, J. W., Curran, T., Porterfield, A. L., and Collins, D. (2006). Activation of preexisting and acquired face representations: the N250 event-related potential as an index of face familiarity. *J. Cogn. Neurosci.* 18, 1488–1497. doi: 10.1162/jocn.2006.18.9.1488

Thompson, D. E., Blain-Moraes, S., and Huggins, J. E. (2013). Performance assessment in brain-computer interface-based augmentative and alternative communication. *Biomed. Eng. Online* 12:43. doi: 10.1186/1475-925X-12-43

Todorov, A., and Uleman, J. S. (2002). Spontaneous trait inferences are bound to actors' faces: evidence from a false recognition paradigm. *J. Pers. Soc. Psychol.* 83, 1051–1065. doi: 10.1037/0022-3514.83.5.1051

Tong, F., and Nakayama, K. (1999). Robust representations for faces: evidence from visual search. *J. Exp. Psychol. Hum. Percept. Perform.* 25, 1016–1035. doi: 10.1037/0096-1523.25.4.1016

Townsend, G., LaPallo, B. K., Boulay, C. B., Krusienski, D. J., Frye, G. E., Hauser, C. K., et al. (2010). A novel P300-based brain-computer interface stimulus presentation paradigm: moving beyond rows and columns. *Clin. Neurophysiol.* 121, 1109–1120. doi: 10.1016/j.clinph.2010.01.030

Waldert, S. (2016). Invasive vs. non-invasive neuronal signals for brain-machine interfaces: will one prevail? *Front. Neurosci.* 10:295. doi: 10.3389/fnins.2016.00295

Wang, F., He, Y., Qu, J., Cao, Y., Liu, Y., Li, F., et al. (2019). A brain-computer interface based on three-dimensional stereo stimuli for assisting clinical object recognition assessment in patients with disorders of consciousness. *IEEE Trans. Neural. Syst. Rehabil. Eng.* 27, 507–513. doi: 10.1109/TNSRE.2019.2896092

Wang, F., He, Y., Qu, J., Xie, Q. Y., Lin, Q., Ni, X. X., et al. (2017). Enhancing clinical communication assessments using an audiovisual BCI for patients with disorders of consciousness. *J. Neural Eng.* 14:046024. doi: 10.1088/1741-2552/aa6c31

**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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 22 January 2020
doi: 10.3389/fncom.2019.00095

# Spectro-Temporal Processing in a Two-Stream Computational Model of Auditory Cortex

*Isma Zulfiqar[1]*, Michelle Moerel[1,2,3] and Elia Formisano[1,2,3]*

*[1] Maastricht Centre for Systems Biology, Maastricht University, Maastricht, Netherlands, [2] Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, Netherlands, [3] Maastricht Brain Imaging Center, Maastricht, Netherlands*

Neural processing of sounds in the dorsal and ventral streams of the (human) auditory cortex is optimized for analyzing fine-grained temporal and spectral information, respectively. Here we use a Wilson and Cowan firing-rate modeling framework to simulate spectro-temporal processing of sounds in these auditory streams and to investigate the link between neural population activity and behavioral results of psychoacoustic experiments. The proposed model consisted of two *core* (A1 and R, representing primary areas) and two *belt* (*Slow* and *Fast,* representing rostral and caudal processing respectively) areas, differing in terms of their spectral and temporal response properties. First, we simulated the responses to amplitude modulated (AM) noise and tones. In agreement with electrophysiological results, we observed an area-dependent transition from a temporal (synchronization) to a rate code when moving from low to high modulation rates. Simulated neural responses in a task of amplitude modulation detection suggested that thresholds derived from population responses in *core* areas closely resembled those of psychoacoustic experiments in human listeners. For tones, simulated modulation threshold functions were found to be dependent on the carrier frequency. Second, we simulated the responses to complex tones with missing fundamental stimuli and found that synchronization of responses in the *Fast* area accurately encoded pitch, with the strength of synchronization depending on number and order of harmonic components. Finally, using speech stimuli, we showed that the spectral and temporal structure of the speech was reflected in parallel by the modeled areas. The analyses highlighted that the *Slow* stream coded with high spectral precision the aspects of the speech signal characterized by slow temporal changes (e.g., prosody), while the *Fast* stream encoded primarily the faster changes (e.g., phonemes, consonants, temporal pitch). Interestingly, the pitch of a speaker was encoded both spatially (i.e., tonotopically) in *Slow* area and temporally in *Fast* area. Overall, performed simulations showed that the model is valuable for generating hypotheses on how the different cortical areas/streams may contribute toward behaviorally relevant aspects of auditory processing. The model can be used in combination with physiological models of neurovascular coupling to generate predictions for human functional MRI experiments.

**Keywords: auditory cortex, sound processing, dynamic neuronal modeling, temporal coding, rate coding**

OPEN ACCESS

***Edited by:***
*Maurizio Mattia,*
*Istituto Superiore di Sanità (ISS), Italy*

***Reviewed by:***
*Emili Balaguer-Ballester,*
*Bournemouth University,*
*United Kingdom*
*Alejandro Tabas,*
*Max Planck Institute for Human Cognitive and Brain Sciences,*
*Germany*

****Correspondence:***
*Isma Zulfiqar*
*isma.zulfiqar@maastrichtuniversity.nl*

***Received:*** *16 September 2019*
***Accepted:*** *23 December 2019*
***Published:*** *22 January 2020*

***Citation:***
*Zulfiqar I, Moerel M and Formisano E (2020) Spectro-Temporal Processing in a Two-Stream Computational Model of Auditory Cortex. Front. Comput. Neurosci. 13:95. doi: 10.3389/fncom.2019.00095*

## INTRODUCTION

The processing of sounds in primate auditory cortex (AC) is organized in two anatomically distinct streams: a *ventral* stream originating in areas located rostrally to the primary auditory core and projecting to the ventral regions of the frontal cortex, and a *dorsal* stream originating in areas located caudally to the primary core and projecting to dorsal frontal regions. Processing in these separate streams is hypothesized to underlie auditory cognition and has been linked respectively to specialized mechanisms of sound analysis for deriving semantic information ("what" processing) or processing sound location and sound movement ("where" processing) (Kaas et al., 1999; Romanski et al., 1999; Belin and Zatorre, 2000; Kaas and Hackett, 2000; Rauschecker and Tian, 2000; Tian et al., 2001; Arnott et al., 2004). Interestingly, the basic response properties (e.g., frequency tuning, latencies, temporal locking to the stimulus) of neurons in areas of dorsal and ventral auditory streams show marked differences (Rauschecker et al., 1996; Bendor and Wang, 2008; Oshurkova et al., 2008; Nourski et al., 2013, 2014), and differences have been reported even for neurons from areas within the same (dorsal) stream (Kuśmierek and Rauschecker, 2014). A consistent observation is that neurons in the rostral field, in comparison to primary and surrounding auditory areas, exhibit longer response latencies and narrower frequency tuning (Recanzone et al., 2000; Tian et al., 2001; Bendor and Wang, 2008; Camalier et al., 2012), whereas neurons in the caudal fields respond with shorter latencies, comparable to or even shorter than those in A1, and have broader frequency tuning (Recanzone et al., 2000; Kuśmierek and Rauschecker, 2014). How this organization of neuronal properties within AC contributes to the processing of spectro-temporally complex sounds remains unclear and poses an interesting question for computational endeavors (Jasmin et al., 2019).

Recent results of neuroimaging studies in humans have put forward the hypothesis that fine-grained spectral properties of sounds are analyzed optimally in ventral auditory regions, whereas fine-grained temporal properties are analyzed optimally in dorsal regions (Schönwiesner and Zatorre, 2009; Santoro et al., 2014). It is, however, unlikely that the neural processing of spectral and temporal properties of sounds is carried out through completely independent mechanisms. Several psychophysical phenomena such as pitch perception based on temporal cues (Houtsma and Smurzynski, 1990; Bendor et al., 2012) or the frequency dependence of amplitude modulation (AM) detection thresholds (Sek and Moore, 1995; Kohlrausch et al., 2000) suggest an interdependence between neural processing mechanisms for spectral and temporal properties.

Therefore, in this study, we aim to introduce a simple, stimulus-driven computational framework for modeling the spectral and temporal processing of sounds in AC and examine the role of the different processing streams. We use the firing rate model of Wilson and Cowan (Wilson Cowan Cortical Model, WCCM; Wilson and Cowan, 1972, 1973; Cowan et al., 2016) which simulates complex cortical computations through the modeling of dynamic interactions between excitatory and inhibitory neuronal populations. Over the years, WCCM has been successfully implemented for simulating neuronal computations in the visual cortex (Ermentrout and Cowan, 1979; Wilson and Kim, 1994; Wilson, 1997). More recently, WCCM has been applied to the AC as well to describe the propagation of activity in the interconnected network of cortical columns and to generate predictions about the role of spontaneous activity in the primary AC (Loebel et al., 2007), and the role of homeostatic plasticity in generating traveling waves of activity in the AC (Chrostowski et al., 2011). Furthermore, WCCM has been proposed for modeling stimulus-specific adaptation in the AC (May et al., 2015; Yarden and Nelken, 2017) and to generate experimentally verifiable predictions on pitch processing (Tabas et al., 2019), etc. While WCCMs are less detailed than models of interconnected neurons, they may provide a right level of abstraction to investigate functionally relevant neural computations, probe their link with psychophysical observations, and generate predictions that are testable using invasive electrocorticography (ECoG) as well as non-invasive electro- and magneto-encephalography (EEG, MEG) and functional MRI (fMRI) in humans.

Here, we used the WCCM to simulate the dynamic cortical responses (population firing rates) in the AC to both synthetic and natural (speech) sounds. After filtering from the periphery, the proposed model processes the spatiotemporally structured (i.e., tonotopic) input in two primary auditory *core* areas. The output of the core areas is then fed forward to two secondary auditory *belt* areas, which differ in terms of their processing of spectral and temporal information and thereby represent the dorsal and ventral auditory processing streams. In a number of simulations, we used this model to examine the coding of amplitude modulated (AM) broadband noise and tones using metrics derived from the electrophysiology (firing rate and temporal synchronization with the stimulus). We also simulated three psychoacoustic experiments to study the role of the multiple information streams that may underlie behavioral AM detection thresholds observed for noise (Bacon and Viemeister, 1985) and tones (Kohlrausch et al., 2000), as well as pitch perception with missing fundamental stimuli (Houtsma and Smurzynski, 1990). Lastly, we investigated the processing of speech stimuli in the model in order to generate predictions on how this cortical spectro-temporal specialization (represented by the four areas) may encode the hierarchical structure of speech.

## MATERIALS AND METHODS

### Model Design and Architecture

**Figure 1A** provides an anatomical schematic of the modeled cortical areas with approximate locations shown on the left supratemporal plane. **Figure 1B** illustrates the overall architecture of the model, consisting of a *peripheral* processing stage and a *cortical* processing stage. The *peripheral* processing stage simulates the peripheral auditory processing in two steps. First, the tonotopic response of the cochlea is estimated using a set of band-pass filters (Gammatone filterbank, $N = 100$) (Patterson, 1986; Patterson et al., 1992). The gains of the filters represent the transfer function of the outer and middle

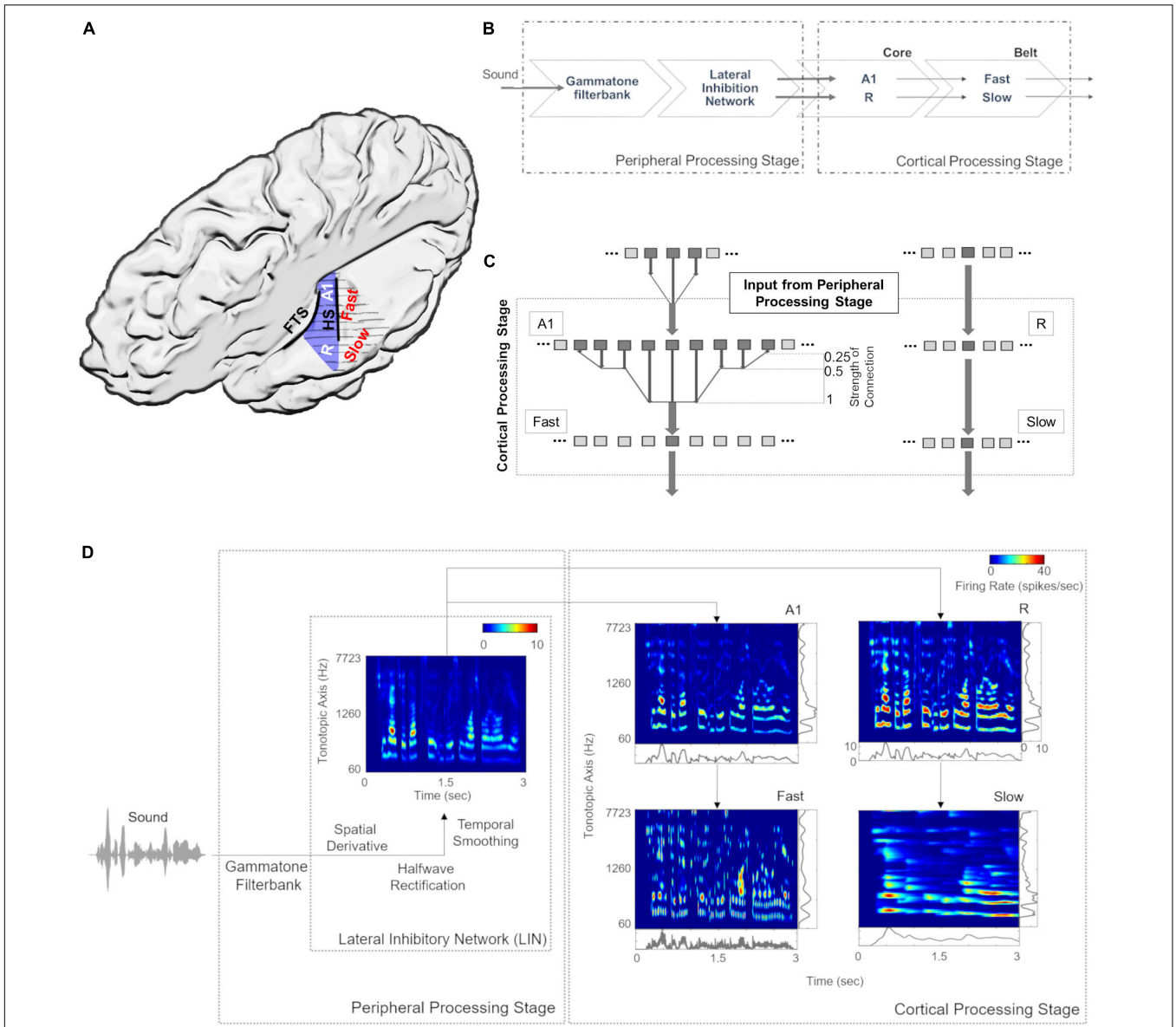


**FIGURE 1 |** Model design and architecture. **(A)** Anatomical schematic of the modeled areas shown on top view of the left supratemporal plane (with the parietal cortex removed). Heschl's sulcus (HS) and first transverse sulcus (FTS) are marked to provide anatomical references while Heschl's Gyrus is highlighted in blue. **(B)** The sound waveform is filtered with a Gammatone filterbank and passed through a Lateral Inhibitory Network (LIN) in the peripheral processing stage, which serves as input to the cortical stage. The neural responses of the simulated core areas (A1, R) are fed forward as input to two simulated belt areas (*Slow* and *Fast*), which differ from each other in their spectral and temporal properties. **(C)** Connections between model stages are shown. The output of Lateral Inhibitory Network (LIN) projects to excitatory units of A1 and R, which in turn project to excitatory units of *Fast* and *Slow*, respectively. While the convergence through A1 to the *Fast* area is high (i.e., many excitatory units of A1 provide input to a single unit of the *Fast* area), convergence through R to the *Slow* area is low (i.e., the units in areas R and *Slow* receive input from only one unit). **(D)** Model output for a sample speech sound is shown at different stages of processing as a spectrogram. The panels at right and bottom of the output of cortical processing stage show mean firing rates across time and tonotopic axis respectively.

ear (4th order Gammatone filterbank implementation by Ma et al., 2007). Following the results from psychoacoustics, the center frequencies of the filters are equally spaced on an $ERB_N$ number scale and their bandwidth increases with center frequency, so as to have a constant auditory filter bandwidth (Glasberg and Moore, 1990). Thus, bandwidth of the 100 rectangular filters is set as 1 ERB (Equivalent Rectangular Bandwidth, based on psychoacoustic measures; for a review of critical bandwidth as a function of frequency, see Moore, 2003). The filter frequencies are centered from 50 to 8000 Hz, equally spaced with a distance of 0.3 Cams (on the $ERB_N$ number scale, $ERB_N$ is the ERB of the auditory filters estimated for young people with normal hearing; Glasberg and Moore, 1990).

Second, the basilar response of the Gammatone filterbank is spectrally sharpened using a Lateral Inhibitory Network (LIN) implemented in three steps by taking a spatial (tonotopic)

derivative, half-wave rectification and temporal integration (Chi et al., 2005). The output of extreme filters (i.e., first and last filter) is removed to avoid any boundary effects of filtering, thus reducing the output of the *peripheral* processing stage to 98 units (60–7723 Hz).

For the *cortical* processing stage, the filtered tonotopic cochlear input is processed in two primary auditory *core* areas (A1 and R) and then fed forward to two secondary auditory *belt* areas (*Slow* and *Fast*; **Figure 1**). These four areas approximate the known architecture of human (Galaburda and Sanides, 1980; Rivier and Clarke, 1997; Wallace et al., 2002) and non-human primates (Hackett et al., 1998; Kaas and Hackett, 2000; Read et al., 2002) AC. Simulated areas primarily differ in their temporal and spectral (spatial) response properties. Specifically, neuronal units in the *Fast* area (approximating caudomedial-caudolateral areas) are characterized by fast temporal dynamics and coarse spectral tuning, whereas units in the *Slow* area (approximating middle lateral-anterolateral areas) are characterized by slow temporal dynamics and fine spectral tuning. It is important to note that these units represent an abstraction at the level of neural population behavior and are not always indicative of single-neuron properties.

In addition, we introduce an interdependence between temporal and spatial (tonotopic) processing within the two *belt* areas, as the variable that determines the temporal dynamics of the responses varies with frequency. Consequently, the units corresponding to lower frequencies in the tonotopic axis respond more slowly than those corresponding to higher frequencies (see Scott et al., 2011; Simpson et al., 2013; Heil and Irvine, 2017). Each simulated area comprises 98 units, which are modeled by excitatory and inhibitory unit pairs. Each of the excitatory core units receives tonotopic input from the corresponding frequency-matched *peripheral* stage. This input only targets the excitatory units of A1 and R. Excitatory responses of A1 and R act as tonotopic input for *Fast* and *Slow* areas, respectively (**Figure 1C**). The output (excitatory responses) at different stages of the model is shown in **Figure 1D**.

## The WCCM

Neuronal units of the cortical areas were simulated using the WCCM in MATLAB (The MathWorks, Inc.). The WCCM is a recurrent firing rate model where neural population processes are modeled by the interaction of excitatory and inhibitory responses. The model dynamics are described by Wilson (1999):

$$\tau\frac{dE_n(t)}{dt} = -E_n(t) + S_E\left(\sum_m w_{EE_{mn}}E_n(t) - \sum_m w_{IE_{mn}}I_n(t) + P_n(t)\right) \quad (1)$$

$$\tau\frac{dI_n(t)}{dt} = -I_n(t) + S_I\left(\sum_m w_{EI_{mn}}E_n(t) - \sum_m w_{II_{mn}}I_n(t)\right) \quad (2)$$

where $E_n$ and $I_n$ are the mean excitatory and inhibitory firing rates at time $t$ at tonotopic position $n$, respectively. $P_n$ is the external input to the network and $\tau$ is the time constant. The sigmoidal function $S$, which describes the neural activity (Sclar et al., 1990), is defined by the following Naka-Rushton function:

$$S(P) = \frac{MP^2}{\theta^2 + P^2} \quad (3)$$

$\theta$ is the semi-saturation constant and $M$ is the maximum spike rate for high-intensity stimulus $P$. The excitatory and inhibitory units are connected in all possible combinations (E–E, E–I, I–E, I–I). The spatial spread of synaptic connectivity between the units $m$ and $n$ is given by the decaying exponential $w_{ij}$ ($i$, $j$= E, I) function:

$$w_{ij_{mn}} = b_{ij}\,exp\left(\frac{-|m-n|}{\sigma_{ij}}\right) \quad (4)$$

In Equation (4), $b_{ij}$ is the maximum synaptic strength and $\sigma_{ij}$ is a space constant controlling the spread of activity. The equations were solved using Euler's method with a time step of 0.0625 ms.

## Parameter Selection and Optimization

Model parameters were selected and optimized based on the following procedure. First, the stability constraints of the model, as derived and implemented by Wilson (1999) were applied. Second, parameters range were chosen so that the model operates in active transient mode, which is appropriate to simulate activity in sensory areas (Wilson and Cowan, 1973). In active transient mode, recurrent excitation triggers the inhibitory response, which in turn reduces the network activity. The balance of excitation and inhibition was achieved by fixing the parameters as described in **Table 1** (for the derivation of these parameters see Wilson, 1999). As shown in previous modeling endeavors (Loebel et al., 2007; May et al., 2015), it is crucial to understand the behavior generated through the interaction of various model properties rather than the exact values of the parameters. In our case, we are interested in the interaction of spectral selectivity and temporal dynamics in neural populations constrained by known physiological response properties of the AC. Thus, while most of the parameters were fixed, further tuning was performed to find the combination of spatial spread ($\sigma$), connectivity between areas and time constant ($\tau$) such that the areas reflected the general spectral and temporal constraints, as derived from the electrophysiology literature (see following subsections).

### Spatial Resolution of the Model

Model parameters, spatial spread ($\sigma$) and connectivity between areas, were determined by matching the sharpness of the model's resulting frequency tuning curves (FTCs) with values reported in the literature. FTCs represent the best frequency of auditory cortical neurons as well as their frequency selectivity (i.e., the sharpness of frequency tuning; Schreiner et al., 2000). In primate AC, the sharpness of neuronal FTCs varies from sharp to broad. Quality factor ($Q$) has been used to express the sharpness of the FTCs $\left(Q = \frac{Best\ Frequency}{Bandwidth}\right)$. The $Q$-values for sharply and broadly tuned auditory cortical neurons have been reported to be around 12 and 3.7, respectively (Bartlett et al., 2011). Also, the core areas

**TABLE 1 |** Fixed parameters of the model.

| Parameters | Values |
|---|---|
| $M$ | 100 |
| $\theta$ inhibition | 60 |
| $\theta$ excitation | 80 |
| $b_{EE}$ | 1.5 |
| $b_{EI} = b_{IE}$ | 1.3 |
| $b_{II}$ | 1.5 |
| $\sigma_{II}$ | 10 |

*M is the maximum spike rate, θ the is semi-saturation constant. Parameters $b_{EE}$, $b_{II}$ $b_{EI}$, and $b_{IE}$, represent the maximum synaptic strength between excitatory units, between inhibitory units, from excitatory to inhibitory units, and vice versa, respectively. All the listed parameter values are same across the four simulated areas.*

**TABLE 2 |** Model parameters across the four simulated areas.

| Parameters | Values | | | |
|---|---|---|---|---|
| | **A1** | **R** | ***Slow*** | ***Fast*** |
| τ (ms) | 10 | 20 | 300–200 | 3–1 |
| $\sigma_{EE}$ | 40 | 40 | 20 | 200 |
| $\sigma_{EI} = \sigma_{IE}$ | 160 | 160 | 80 | 300 |

*For the four simulated areas, the values for varying parameters, time constant τ (reported over the tonotopic axis from low to high best frequencies of the units), spatial spread parameter σ (EE, EI/IE) are listed.*

have been described as having narrower tuning bandwidths than belt regions (Recanzone et al., 2000). In order to generate narrow FTCs of A1, R, and *Slow* areas and broad FTCs for *Fast* area, we iteratively changed spread of activity within the simulated area (final values are listed in **Table 2**). When changing the spread of activity (σ) within an area did not affect the *Q* of the area, the connectivity across the areas was manipulated. It should be noted that the projections act as a filter, which is then convolved with the spatial input per unit time. To avoid any boundary effects, symmetric kernel filters (odd number of elements) were used and the central part of the convolution was taken as a result. Final connectivity across regions (i.e., distribution of input units projecting from one area to another) is shown in **Figure 1C**.

The narrower tuning in the *Slow* area results from the smaller spread of excitation (σEE, see **Table 2**), and from the one-to-one projection from R units (**Figure 1C**). The broader tuning in the *Fast* area is simulated by a many-to-one projection from the Gammatone filterbank to a single unit of A1 (three to one) and from A1 to the *Fast* areas (nine to one). The strength of these connections is shown in **Figure 1C**. The FTCs across areas are quantified using *Q* at half-maximum bandwidth. The units tuning in the simulated A1 and R areas have mean $Q = 6.32$ (std = 1.43), units in the *Fast* area have mean $Q = 4$ (std = 0.87), while units in the *Slow* have $Q = 8.35$ (std = 2.1). In line with the experimental observations (Kuśmierek and Rauschecker, 2009), the *Q*-values increased with increasing center frequencies, while maintaining the general trend of broad tuning in *Fast* and narrow tuning in *Slow* area. **Figure 2** shows FTCs across the four simulated areas for a single unit with best frequency at 4.3 kHz.

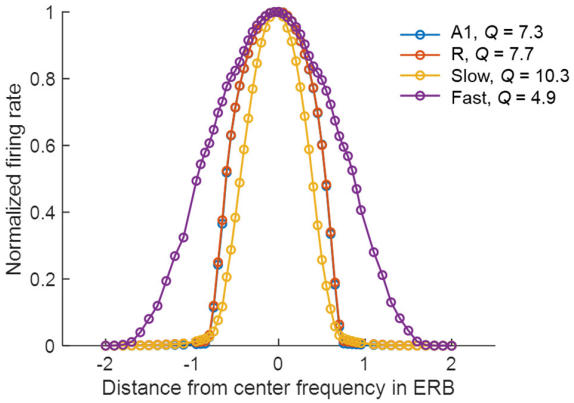


**FIGURE 2 |** Frequency tuning curves (FTCs) of the unit with best frequency at 4.3 kHz across simulated areas. Areas A1 (blue line) and R (red line) are sharply tuned, with *Q* of 7.3 and 7.7, respectively. The *Slow* area (yellow line) has the sharpest tuning curves with *Q* of 10.3, while *Fast* (purple line) has the broadest tuning with *Q* of 4.9. *Q* is measured as the ratio of the best frequency and the half-maximum bandwidth in Hz.

## Temporal Resolution of the Model

Temporal structure represents an important aspect of natural acoustic signals, conveying information about the fine structure and the envelope of the sounds (Giraud and Poeppel, 2012). In several species, a gradient of temporal responses has been observed in AC, with higher stimulus-induced phase locking (synchrony) and lower latencies in area AI compared to adjacent areas (AI vs. AII in cats: Bieser and Müller-Preuss, 1996; Eggermont, 1998; AI vs. R and RT in monkeys: Bendor and Wang, 2008). Correspondingly, model parameters determining the temporal properties of population responses in the simulated areas were adjusted to match such electrophysiological evidence. **Table 2** shows the resulting time constant τ for the simulated areas. Note that the values of parameter τ do not represent the latency of the first spike measured for single neurons but affect the response latencies and dynamics at a population level.

### *Temporal latencies*

As neurons in core area R have longer latencies than A1 (Bendor and Wang, 2008), we selected a higher value of τ for simulated R than A1. Based on the evidence of the caudomedial field showing similar latencies to A1 (Recanzone et al., 2000; Kuśmierek and Rauschecker, 2014), we adjusted τ of the *Fast* area so that the area is as fast as A1. In contrast, we set τ of the *Slow* area such that this region generates a more integrated temporal response, with the firing rate taking longer to reach the semi-saturation point. These τ values, in combination with the spatial connectivity constraints, cause the simulated belt area to display a spectro-temporal tradeoff. Additionally, in both *Slow* and *Fast* areas τ decreases linearly along the spatial axis (maximum and minimum values are reported in **Table 2**) with increasing best frequency, following electrophysiological evidence of interaction of the temporal and frequency axis where shorter latencies have been found to be correlated with high best frequencies in macaques (Scott et al., 2011).

***Temporal synchrony***

To further refine parameter τ, next we examined stimulus-driven phase locking of the simulated neural activity. Electrophysiological measurements report synchronization in the neural response to the sound carrier and envelope for a limited range of frequencies, and the upper limit of this phase locking has been found to decrease along the auditory pathway (Joris et al., 2004). At the level of cortex, while the strongest synchronization is reported for modulation rates up to 50 Hz (AM stimuli: Liang et al., 2002, Clicks: Nourski et al., 2013), weaker synchronization to even higher rates (up to 200 Hz) has been observed for a subset of units (Steinschneider et al., 1980; Bieser and Müller-Preuss, 1996; Lu et al., 2001; Nourski et al., 2013). In light of the evidence above, we adjusted τ to mimic this behavior and have strongest temporal synchronization for the low range of modulation rates (up to 50 Hz), with some residual synchronization to higher rates.

## Model Evaluation

The model performance was evaluated in three stages. First, we simulated the electrophysiological coding of AM (for both noise and tone carriers). Second, we evaluated the model's ability to predict results of human psycho-acoustical tasks, including the determination of amplitude modulation detection threshold functions, tMTFs and perception of missing fundamental. Lastly, we used speech stimuli to investigate the representation of pitch and AM features of a complex sound across the simulated areas. All artificial stimuli (AM noise, AM tones and missing fundamental complex tones) were generated using MATLAB with a sampling rate of 16 kHz and 1 s duration). Speech stimuli were taken from LDC TIMIT database (Garofolo et al., 1993). In all cases, the key readouts of the model were synchronization to stimulus features and firing rates. The pitch estimates matched against model output, where relevant, were computed using the YIN algorithm (de Cheveigné and Kawahara, 2002).

### Coding of AM Stimuli: Evidence From Electrophysiology

To evaluate the model's coding of AM, sinusoidally amplitude modulated (sAM) stimuli were used. AM sounds were defined by $(1+m\sin 2\pi gt)^{*}carrier$, where $m$ is the modulation depth, $g$ is the modulation rate and $t$ is time. The modulation rates were chosen to be 2–9 Hz (linearly spaced), and 10–1000 Hz (logarithmically spaced). Broadband noise was used as carrier to study the response of all units working together while pure tones (500 Hz–3 kHz–5 kHz) were employed to evaluate carrier-specific effects on amplitude modulation coding.

To quantify synchronization of responses to the temporal structure of AM sounds, we employed two measures from the electrophysiology literature (Eggermont, 1991; Joris et al., 2004; Bendor and Wang, 2008): vector strength $\left(VS = \frac{Strength\ of\ Fourier\ Component\ at\ the\ Modulation\ Rate}{Average\ Firing\ Rate}\right)$ (Goldberg and Brown, 1969), and rate modulation transfer function (rMTF), which is the average firing rate as a function of modulation rate. VS was computed for all modulation rates (and three harmonics), for both tone and noise carriers, across the four simulated areas. We considered a simulated area as being synchronized to a modulation rate when VS was greater than 0.1 (this is an arbitrary threshold chosen to compare phase-locking across conditions and areas).

rMTFs were calculated from the average firing rates (i.e., the Fourier component at 0 Hz) and normalized for all areas. For the computation of rMTFs, the modulation depth is fixed at 100% across all AM stimuli. For noise carriers, the computation of the VS and rMTF is based on the mean across all 98 excitatory channels. For the tone carriers, only the channel maximally tuned to the carrier frequency is considered.

### Simulating Psychoacoustical Observations

The model was tested using three paradigms approximating human psychoacoustic studies. The first two experiments simulated temporal modulation transfer functions (tMTFs: quantifying the modulation depth required to detect different modulation rates) for broadband noise (Bacon and Viemeister, 1985) and tones (Kohlrausch et al., 2000). The third experiment simulated pitch identification with missing fundamental stimuli (Houtsma and Smurzynski, 1990).

For the simulated tMTFs, AM sounds with incremental modulation depths (from 1 to 100%) were presented to the model and the oscillations in the model's output were measured. In the psychoacoustic measurements, the lowest modulation depth at which subjects can detect the modulation is considered the detection threshold. In the model, using synchronization as output measure, the lowest value of modulation depth at which the output is synchronized to the modulation rate (i.e., the strongest Fourier component was at the modulation rate) is considered as the detection threshold for that AM rate. This procedure was repeated for all the modulation rates and, for all simulated areas. For noise carriers, the mean across the excitatory units across each area is analyzed and compared to data collected by Bacon and Viemeister (1985). The model response was simulated for modulation rates at 2–9 Hz (linearly spaced), and 10–1000 Hz (logarithmically spaced).

For AM tones, the analysis of the waveform shows spectral energy at the carrier frequency and at the carrier frequency ± modulation rate. These accompanying frequency components are called "spectral sidebands" of the carrier frequency. If the modulation rate is high enough, these sidebands activate distinctively different auditory channels than the carrier frequency and can be detected audibly apart from the carrier frequency. Thus, for the tone carriers (1 and 5 kHz) the active part of the population (comprising the best frequency channel and spectral sidebands) was used to compute tMTFs based on temporal synchronization to the modulation rate (temporal code) and detection of sidebands (spatial code). As before, for the temporal code, the lowest value of modulation depth at which the output is synchronized to the modulation rate (i.e., the strongest Fourier component was at the modulation rate) is considered as the detection threshold for that AM rate. For the spatial code, the modulation depth at which the side-band amplitude (mean firing rate over time) is at least 5, 10, 15, or 20% of the peak firing rate (firing rate of the channel with CF closest to carrier frequency) are calculated. The best (lowest) value of modulation

depth is chosen from both coding mechanisms. The combination of these coding mechanisms is then compared to tMTFs (at 30 dB loudness) reported by Kohlrausch et al. (2000). The modulation rates tested were 10–1600 Hz (logarithmically spaced).

Pitch of missing fundamental complex tones has been shown to be coded by temporal and spatial codes, depending on the order of harmonics and frequency of missing fundamental (Bendor et al., 2012). Here we replicated this finding by simulating the model response to complex tones with low order (2–10) and high order harmonics (11–20) and varying missing fundamental frequency from 50 to 800 Hz. The synchronization to the missing $F_0$, measured in VS, is computed from the mean responses over time in each of the four simulated areas. Furthermore, to evaluate the role of synchronization in pitch perception, we simulated model responses to complex tones with unresolved harmonics of a missing fundamental frequency by approximating a pitch identification experiment by Houtsma and Smurzynski (1990). The missing fundamental tone complexes vary in two aspects: the number of harmonic components (2–11) and the lowest harmonic component (10 and 16) while the fundamental frequency ($F_0$) is fixed at 200 Hz. For each combination of lowest harmonic component and number of components in the harmonic complex, we computed the synchronization to the $F_0$ (in VS) and mean firing rates for all four regions.

### Model Responses to Speech

Model responses to the speech stimuli were analyzed in two stages. The speech stimuli (630 sentences, all spoken by different speakers; mean duration 3.4 s) were randomly selected from the LDC TIMIT database (Garofolo et al., 1993). To study how key temporal features of speech waveforms are represented in the modeled areas, we compared the temporal modulations in the output of all four simulated areas to the temporal modulations of the input signals. To this end, we computed the input-output magnitude spectrum coherence (*mscohere* in MATLAB with a 2048 point symmetric hamming window and overlap of 1500 samples) between the input speech signal (after LIN) and the output of all four areas. The coherence values are then scaled across the four areas using the mean spatial activity along the tonotopic axis (i.e., the mean firing rate over time for all sounds). To highlight the difference in spectrum coherence between the spectro-temporal processing streams in the model, the difference between the scaled input-output coherence is computed to compare the two *core* (R–A1) regions to each other and the two *belt* areas (*Slow*–*Fast*).

# RESULTS

## Coding of AM Stimuli

We investigated the model's AM coding using both broadband noise and tone carriers. By using broadband noise as carrier, we simulated general responses for each of the four areas, and then used pure tone carriers to study the dependence of the synchronization and rate coding on the tonotopic location (i.e., the best frequency of the units).

### Sinusoidal AM Noise

**Figure 3** shows the response of the four simulated cortical areas (A1, R, *Fast*, and *Slow*) as a function of the modulation rate of sinusoidally amplitude modulated (sAM) noise. We analyzed the mean response of all units for each area. Across regions, the response synchronization (measured as VS) decreases with increasing modulation rate (solid lines in **Figures 3A–D** for A1, R, *Fast,* and *Slow* areas respectively). The decrease in synchronization is observed to be rapid above an area-specific modulation rate (8 Hz for A1, R and *Fast* areas, 2 Hz for *Slow*). Taking the lower limit for synchronization as VS = 0.1, the highest AM rate to which the areas synchronize is 54 Hz in A1, 33 Hz in R, 4 Hz in *Slow* and 54 Hz in *Fast*. Overall, the observed responses to modulation rates show a low-pass filter profile.

Instead, the firing rate [rate Modulation Transfer Functions (rMTFs), dash-dotted lines] shows different behavior across the four areas in response to AM noise. For A1, R and *Fast* areas (**Figures 3A,C** respectively), the firing rate does not change for lower modulation rates (until 10 Hz for A1 and *Fast*, until 6 Hz for R) and then rapidly increases until a maximum limit (54 Hz for A1, R and *Fast*) and does not further change in response to higher modulation rates. In contrast, the firing rate in the *Slow* area (**Figure 3D**) shows a band-pass profile between 6 and 100 Hz, peaking at ∼20 Hz.

### Sinusoidal AM Tones

Next, we explored the frequency dependence of AM processing. As the use of broadband noise as a carrier provides no information about the temporal properties of different frequency channels along the tonotopic axis, we simulated model responses to AM pure tone carriers. **Figure 4** shows response synchronization (VS, left column) and firing rate (rMTFs, right column) across cortical areas as a function of AM rate, separately for units best responding to a low (solid lines), middle (dashed lines), and high (dash-dotted lines) frequency pure tone carriers (500, 1k and 3k Hz respectively). For each area, the responses in the model's frequency channel matching the tone carrier are shown. The synchronization shows a low-pass filter profile consistently for all three carriers. With increasing carrier frequency, the A1, R, and *Slow* areas (**Figures 4A,C,E**) are synchronized (VS cut-off at 0.1) to higher modulation rates (A1: 33 Hz for 500 Hz, 54 Hz for 1 kHz and 3 kHz, R: 26 Hz for 500 Hz, 33 Hz for 1 kHz and 3 kHz, *Slow*: 3 Hz for 500 Hz, 4 Hz for 1 and 3 kHz). This behavior is consequence of the relationship between the temporal and spatial axis (a property of the model), with temporal latencies reducing with increasing center frequencies of the units allowing phase-locking to higher modulation. The *Fast* area (**Figure 4G**) shows a similar cutoff for all carriers at 54 Hz. The rMTFs (**Figures 4B,D,F,H** for areas A1, R, *Slow,* and *Fast* respectively), however, show more complex and varied behavior for different carriers (including monotonically increasing, band-pass, and band-stop behavior). This behavior is in line with rMTFs from electrophysiological studies, where instead of singular behavior (like low-pass filter profile reported for tMTFs), rMTFs show variety of response profiles (Schreiner and Urbas, 1988; Bieser and Müller-Preuss, 1996; Liang et al., 2002; Bendor and Wang, 2008).

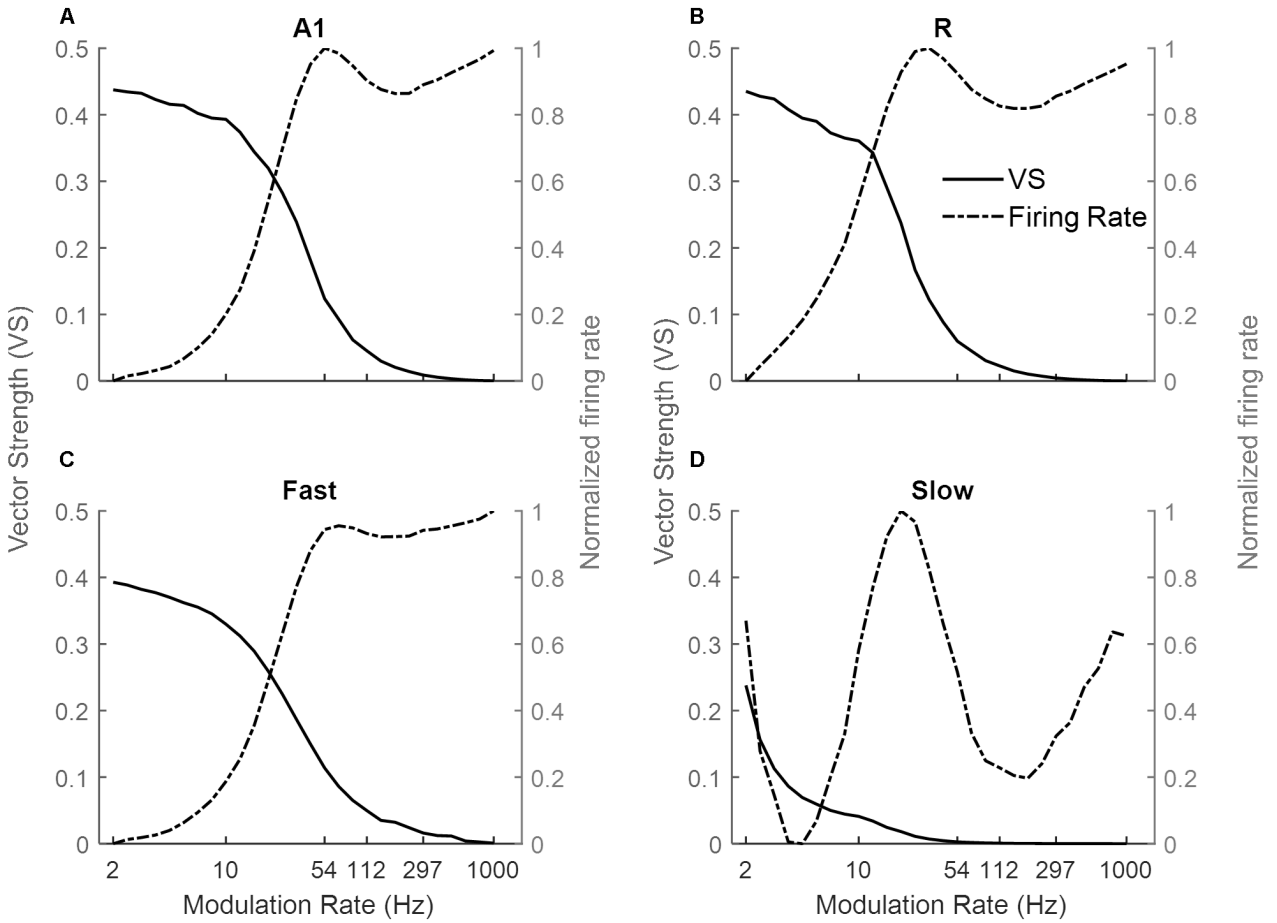


**FIGURE 3 |** Model responses to sAM noise across simulated areas. A dual coding mechanism for modulation rates, i.e., temporal (measured as Vector Strength, VS, solid lines) and rate codes (quantified as the rate Modulation Transfer Functions, rMTFs, dash-dotted lines), are shown for A1, R, *Fast*, and *Slow* areas in **(A–D)** respectively. In A1, R, and *Fast* areas, the synchronization decreases for higher modulation rates and is complimented by increasing firing rate. While very little synchronization is observed in the *Slow* area, the respective rMTF shows an interesting band-pass profile.

## Simulating Psychoacoustic Observations

Next, the model was tested using three experimental paradigms similar to those employed in human behavioral studies. The first two experiments tested the temporal modulation transfer functions (tMTFs characterizing the modulation depth required to detect different modulation rates) for broadband noise (Bacon and Viemeister, 1985) and tones (Kohlrausch et al., 2000). The third experiment examined the effects of the number of harmonics in pitch identification with missing fundamental stimuli (Houtsma and Smurzynski, 1990).

### Temporal Modulation Transfer Functions for Broadband White Noise

Similar to the behavioral task of Bacon and Viemeister (1985), we measured responses of the model to AM sounds with variable modulation depth and recorded the minimum modulation depth where the output signal was synchronized to the modulation rate (i.e., the strongest Fourier component was at the modulation rate) of the AM noise. **Figure 5** illustrates the simulation results (solid colored lines), along with human psychoacoustic data (dash-dotted black lines with circles, adapted from Bacon and Viemeister, 1985). Lower values depict higher sensitivity to the modulation rates. A1 and R show lower thresholds for slower than for faster modulation rates. In the *Fast* area, the detection profile is similar to A1 and R, but the minimum detection depth is higher than in the other areas. The broad tuning of the *Fast* area reduces the precision of the temporal structure of the input signal. Thus, the *Fast* area performs worse than the other areas across modulation rates. In the *Slow* area, modulation detection is observed to be limited to rates below 10 Hz. Thus, the *core* areas outperformed the *belt* areas in the detection of amplitude modulations. The modulation depth detection profile of the *core* areas resembles the results from human psychophysics suggesting that primary auditory cortical processing may underlie tMTFs reported in psychophysics. In comparison with synchronization, rate coding is difficult to quantify as observed before with varying response profiles for rMTFs along the frequency axis (**Figures 4F,H**). The difference between our simulations and psychophysical findings at faster rates may be explained by the fact that our simulations only considered coding through

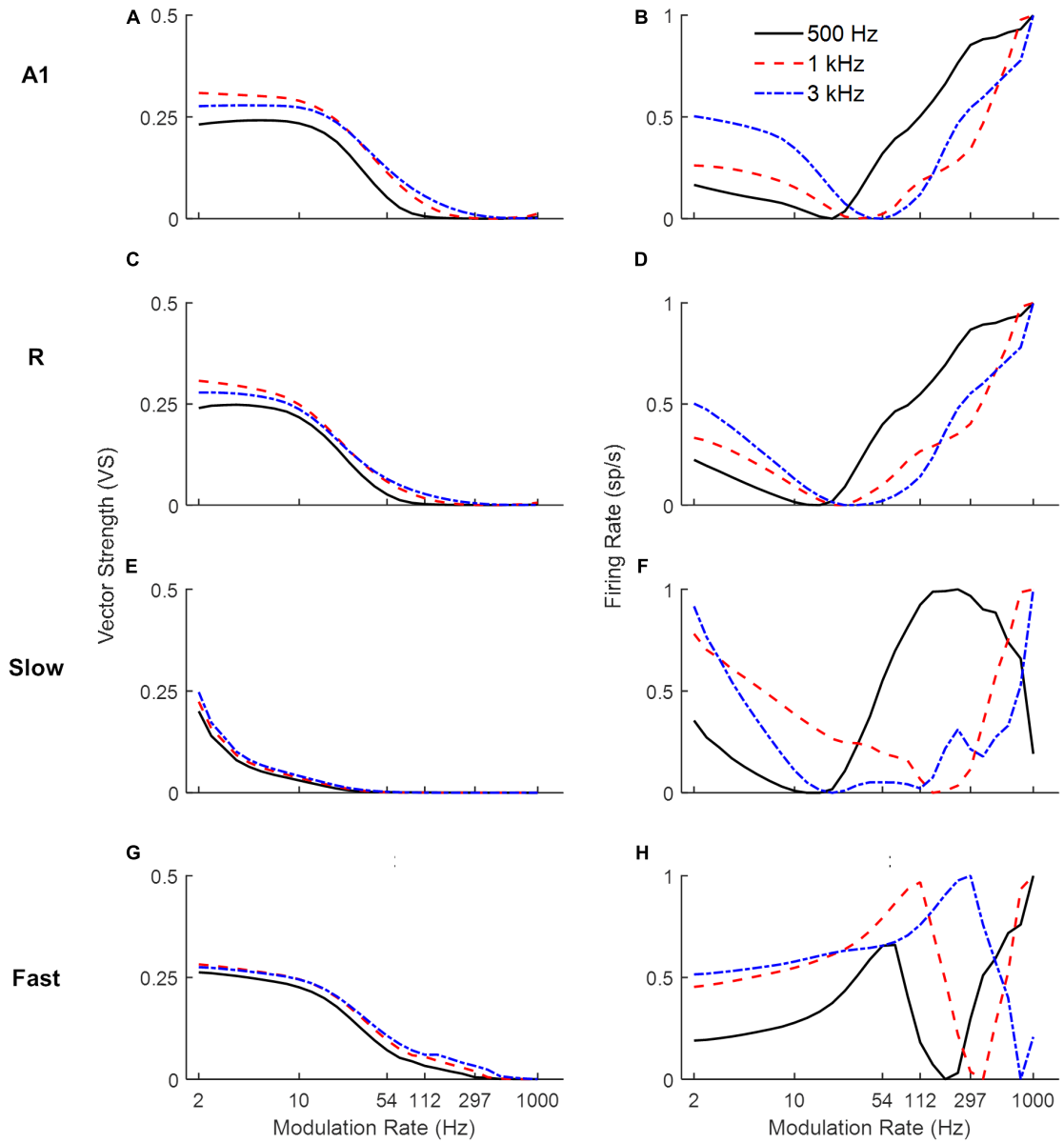


**FIGURE 4 |** Model responses to sAM tones across simulated areas. A dual coding mechanism for modulation rates, i.e., temporal (measured as Vector Strength, VS, left panels) and rate codes (quantified as the rate Modulation Transfer Functions, rMTFs, right panels), are shown for A1, R, *Fast* and *Slow* areas in respective panels (A1: **A,B**, R: **C,D**, *Slow*: **E,F**, *Fast*: **G,H**). For the three different carriers, synchronization to higher modulation rates is observed with increasing carrier frequencies across areas **(A,C,E,G)**. Rate coding, however, shows more varied profiles with different carriers **(B,D,F,H)**.

response synchronization and ignored the contribution of rate coding contributing to the detection of higher modulation rates.

## Temporal Modulation Transfer Functions of Sinusoidal Carriers

We then investigated the model's detection threshold function of sAM tones. Psychoacoustic studies have shown that human performance does not change across the lower modulation rates, becomes worse for a small range and then improves after the sidebands introduced by the modulation become detectable (Sek and Moore, 1995; Kohlrausch et al., 2000; Moore and Glasberg, 2001; Simpson et al., 2013). We obtained model responses to sAM tones as a combination of temporal and spatial codes. To characterize an area's modulation detection threshold represented by temporal code, the lowest modulation depth at which the best frequency unit or the spectral sideband synchronized to the modulation rate was chosen. Additionally, the spatial code was quantified by detection of spectral sideband. **Figure 6** shows the lowest modulation depth for which A1 (solid lines in **Figures 6A,C**) and R (solid lines **Figures 6B,D**) code modulation rates of sAM tones and the psychoacoustic data for 1 and 5 kHz sinusoidal carriers at 30 dB (dash-dotted lines with circles, Kohlrausch et al., 2000).

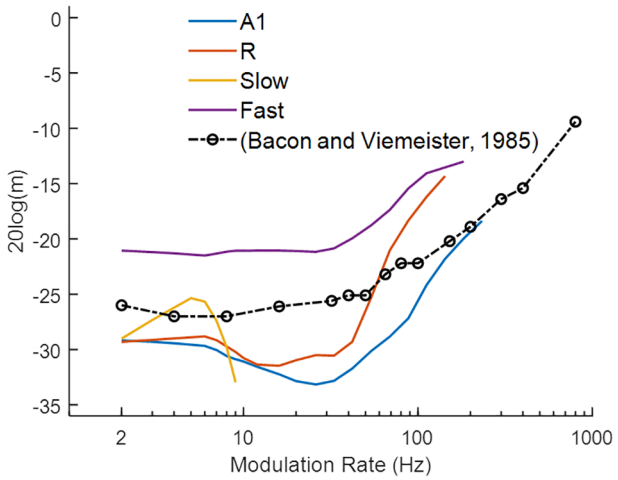


**FIGURE 5 |** Modulation detection with sAM noise. The temporal Modulation Transfer Functions (tMTFs), illustrating the minimum depth required to detect the amplitude modulation in sAM noise, are shown for the four model areas (in colored lines) and for a psychoacoustic study (black line and circles; adapted from Bacon and Viemeister, 1985). Lower values depict higher sensitivity to modulation rate. Modulation depth, m (dB) of the signal is plotted on y-axis.

The initial increase in depth values indicates the contribution of temporal coding of the modulation rates that gets worse with higher modulation rates. With increasing modulation rates, however, the spectral sidebands dissociate from the carrier channel and the contribution of spectral coding is observed. The modulation depths at which the sideband amplitude (mean firing rate over time) is detectable (multiple threshold cut-offs are shown where sideband activity is 5, 10, 15, and 20% of the firing rate of the channel with CF closest to carrier frequency) are also shown in **Figure 6**. No synchronization is observed in the *Slow* and *Fast* areas. Overall, model results show a clear frequency dependence as detection of higher rates was observed for the higher carrier (maximum for A1: 500 Hz for 1 kHz carrier, 1.2 kHz for 5 kHz carrier; R: 1.2 kHz for 1 kHz carrier, 1.6 kHz for 5 kHz carrier). The modulation detection by the model slightly worsened with increasing modulation rate but improved (lower m values) as the sidebands introduced by the modulation became detectable (after 100 Hz for the 1 kHz carrier in A1 and R, after 400 Hz for 5 kHz carrier in A1). This improvement of AM detection threshold for high AM rates is in accordance with human psychophysics, where observations show a decrease in performance with increasing modulation rates is followed by a performance increase accompanied with side-band detection (Sek and Moore, 1995; Kohlrausch et al., 2000; Moore and Glasberg, 2001; Simpson et al., 2013). Additionally, matching the model results, human psychophysics show improved performance (i.e., detection of higher rates) with increasing carrier frequencies.

## Pitch of Missing Fundamental Sounds

Missing fundamental sounds are harmonic complexes that, despite lacking energy at the fundamental frequency ($F_0$), induce the percept of a pitch corresponding to $F_0$ (Yost, 2010; Oxenham, 2012). If the harmonic components in the missing fundamental sound are resolved (i.e., each component produces a response on the basilar membrane that is distinct from that of neighboring harmonic components), the pitch information can be extracted through a spectral (spatial) mechanism, or a temporal mechanism if harmonics are unresolved, or a combination of the two (Yost, 2009). Bendor et al. (2012) have shown that low $F_0$ sounds with higher-order harmonics are primarily represented by temporal mechanisms. Thus, we tested the effect of harmonic order on the detection of missing $F_0$ through temporal synchrony across simulated areas. **Figure 7** shows synchronization (temporal code, measured as VS) to missing $F_0$ of complex tones with lower-order and higher-order harmonics in panels A and B respectively. Stronger synchronization is observed for higher-order harmonics compared to lower-order harmonics for lower missing $F_0$ complex tones in A1, R, and *Fast* areas. The effect is most pronounced in the *Fast* area. However, the synchronization drops with increasing missing $F_0$, and very little to none synchronization is observed after 400 Hz irrespective of the order of harmonics in the complex tone.

For low pitch missing fundamental sounds, psychophysics experiments employing sounds with unresolved harmonics have shown that humans are better at identifying a missing fundamental pitch when the sound consisted of lower (lowest harmonic = 10) compared to higher unresolved harmonics (lowest harmonic = 16), yet the performance reaches a plateau as more harmonics components are included for the sound consisting of lower but not higher-order harmonics (Houtsma and Smurzynski, 1990). To evaluate whether temporal mechanisms play a role in these findings we simulated a pitch identification experiment (Houtsma and Smurzynski, 1990) and explored the effects of the number of harmonic components and lowest order harmonic in the missing fundamental complex tone on the model's behavior. As already established, simulated populations could only successfully synchronize to lower missing $F_0$ (**Figure 7**), thus the task employed complex tones with low missing $F_0$ (200 Hz). **Figure 8** shows the model's synchronization (VS) to the missing $F_0$ (200 Hz and the first three harmonics) across the simulated regions (in blue lines), along with the results from the psychophysics experiment (in red lines, data adapted from Houtsma and Smurzynski, 1990).

While we did not observe any differences due to harmonic order in VS measured in A1, R, and *Slow* areas (**Figures 8A,B,D**), the *Fast* area (**Figure 8C**) showed clear dissociation in synchronization code when the lowest order harmonic changed from 10 to 16. That is, the synchronization to the missing $F_0$ in the *Fast* area was stronger when the lowest order harmonic was 10. Additionally, for both complex tones, the performance of the *Fast* area improved with an increasing number of components. The improvement in synchronization was rapid when the number of components changed from 2 to 4 for the lowest order harmonic at 10. These observations are in line with the pitch identification data shown in the red lines. Thus, neural response properties similar to those of the *Fast* area are optimized to temporally detect the $F_0$ from missing fundamental sounds, and responses in the *Fast* area follow human behavior.

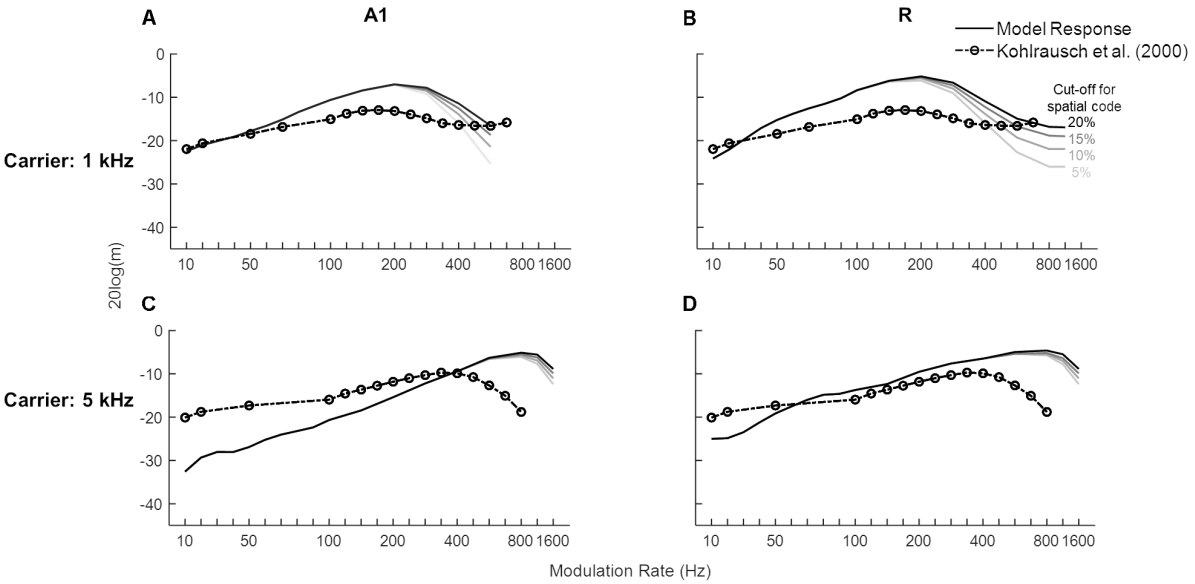


**FIGURE 6 |** Modulation detection with sAM tones. The solid lines show the temporal Modulation Transfer Functions (tMTFs), illustrating the minimum depth required to detect the amplitude modulation in sAM tones (1 kHz in top panels, 5 kHz in bottom panels), are shown for the two *core* areas (A1 in **A,C**, R in **B,D**). The model output is a combination of temporal and spatial codes for modulation detection. Variation in the spatial code is shown at four different cut-off values, represented by the solid lines in different gray-scales. Data from a psychoacoustic study are shown in dash-dotted lines with circles (adapted from Kohlrausch et al., 2000). Lower values depict higher sensitivity to modulation rate. Modulation depth, m (dB) of the signal is plotted on y-axis.

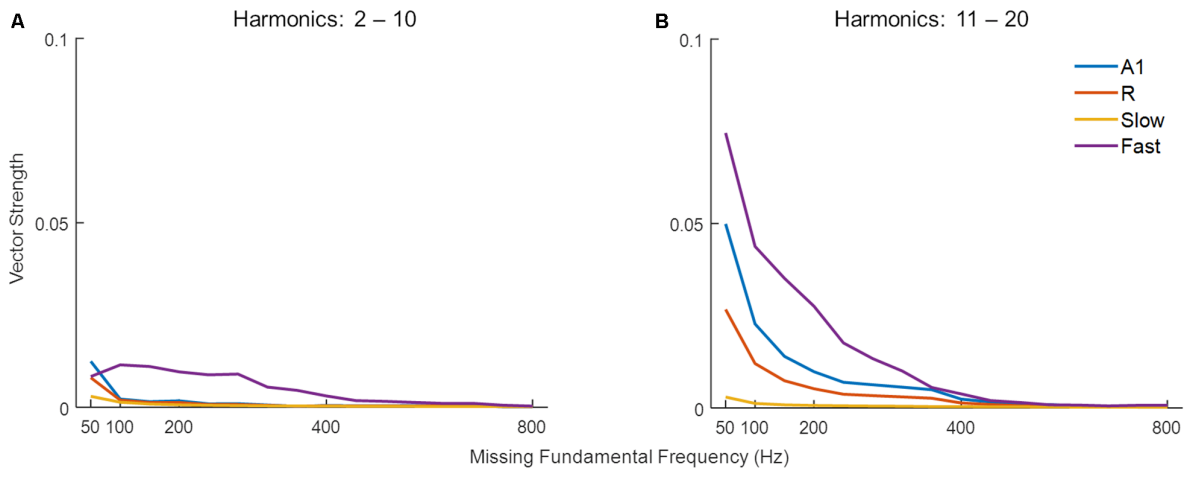


**FIGURE 7 |** Synchronization to missing fundamental frequency across harmonic order. The model performance in detecting missing fundamental of complex tones (measured as vector strength) with **(A)** low-order harmonics, and **(B)** high-order harmonics. Simulated responses in the four areas are shown in different colors.

Unlike synchronization, the simulated firing rate (**Supplementary Figure S1**) did not show a pattern that matched the behavioral data. Specifically, the simulated firing rate increased monotonically as a function of the number of components in the complex tone, irrespective of the lowest order harmonic.

## Model Responses to Speech

Speech signals encode information about intonation, syllables, and phonemes through different modulation rates. We explored the processing of speech sounds across simulated cortical areas to study the importance of simple spectro-temporal cortical properties, as reported by electrophysiology and represented by the model, in coding these temporal features of speech. To this end, we analyzed model output in response to 630 speech stimuli by computing the magnitude spectrum coherence between these sounds (the output of the LIN stage) and the simulated model responses for each of the four areas. **Figure 9** shows the normalized coherence plots (scaled by the normalized time-averaged activity). In all regions, we observed model synchronization to slow changes in the stimuli (<20 Hz).

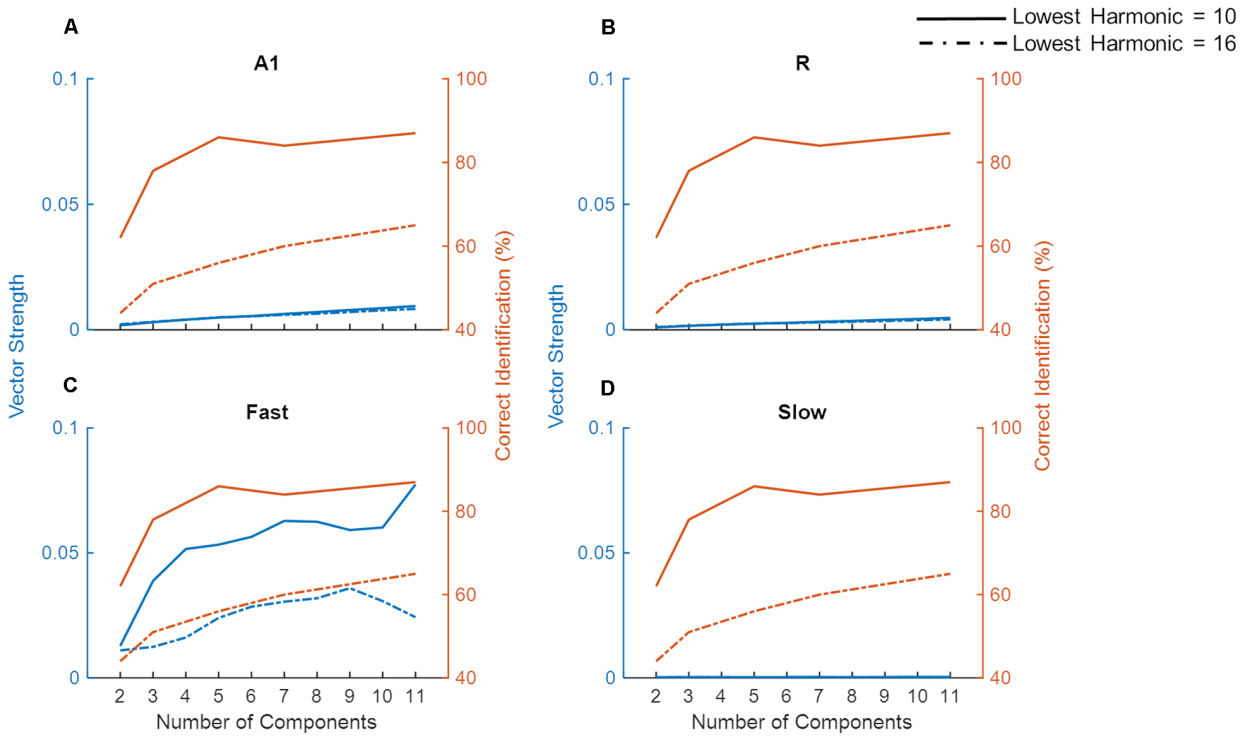


**FIGURE 8 |** Model performance on a missing fundamental task. The model performance in detecting missing fundamental of complex tones (synchronization to missing fundamental frequency at 200 Hz, measured as Vector Strength) is shown for areas A1, R, *Fast*, and *Slow* (blue lines in **A–D**, respectively). Human behavioral data on pitch identification (%) task (Houtsma and Smurzynski, 1990) is plotted in orange lines. Solid lines show complex tones with lowest harmonic at 10 while the dash-dotted lines show the lowest harmonic component at 16.

Next, in order to highlight differences in the temporal response properties between regions, we computed difference plots for the simulated core and belt areas. While we observed no differences in coding of temporal features between A1 and R, **Figure 10** shows that differences are present in the *belt* stream (comparing the coding of temporal features in the *Fast* to those in the *Slow* area). The difference between the coherence (*Slow–Fast*) across 630 stimuli (mean: -0.0332, SEM: 0.0041) was used to compute the data distribution in four percentiles (65, 75, 85, and 95%). These percentiles are shown along the color bar in **Figure 10** (with the distribution) to provide a threshold for the significance to the difference between input-output coherence of the *Slow* and *Fast* area. Shades of blue show stronger input-output coherence in the *Slow* area, while the warmer colors indicate stronger input-output coherence in the *Fast* stream. The *Slow* area represents the slower changes (4–8 Hz) in the speech envelope better than the *Fast* area. The *Fast* area, on the other hand, highlights faster changes in the temporal structure of speech in two frequency ranges (30–70 Hz, and around 100–200 Hz).

We hypothesized that the higher of these two frequency ranges (100–200 Hz) may reflect the presence of temporal pitch information in the *Fast* area. The temporal code for pitch in the simulated areas was estimated by computing short-time the Fourier Transform (window length: 300 ms, overlap: 200 ms) over length of the signal. The resulting power spectral density estimates showed temporal synchronization to the frequencies approximating the pitch in A1, R and *Fast* areas over time. For the purpose of comparison across simulated areas, the pitch estimates and contour obtained for voiced portions of the sounds (using the YIN algorithm) were correlated with the oscillatory activity of individual simulated areas for all 630 speech stimuli. Mean correlation values were A1: 0.46 (SEM: 0.02), R: 0.47 (SEM: 0.02), *Slow*: -0.14 (SEM 0.01), *Fast*: 0.59 (SEM 0.01), and showed that the *Fast* area best represented the pitch information through synchronization to instantaneous $F_0$.

**Figure 11** highlights the presence of a dual mechanism for coding pitch, as pitch information is present in both spectral (i.e., spatially, by different units) and temporal (by different oscillatory activity) model responses for a sample sound (male speaker, sentence duration 3.26 s; selected from LDC TIMIT database; Garofolo et al., 1993). In **Figure 11A**, the time-averaged response to the speech sentence across the tonotopically-organized channels in the four simulated areas is shown. In all the areas, a peak in the response profile can be observed in those frequency channels that matched the $F_0$ of the speaker (best estimate computed using YIN algorithm: 109 Hz). This spectral (i.e., spatial) representation of the speech signal's pitch is strongest in the *Slow* area and weakest in the *Fast* area. A1 and R show similar profiles with respect to each other. Contour tracking of pitch in the *Fast* area with the sample sound (correlation 0.74) is shown in **Figure 11B** (pitch contour of the speech signal

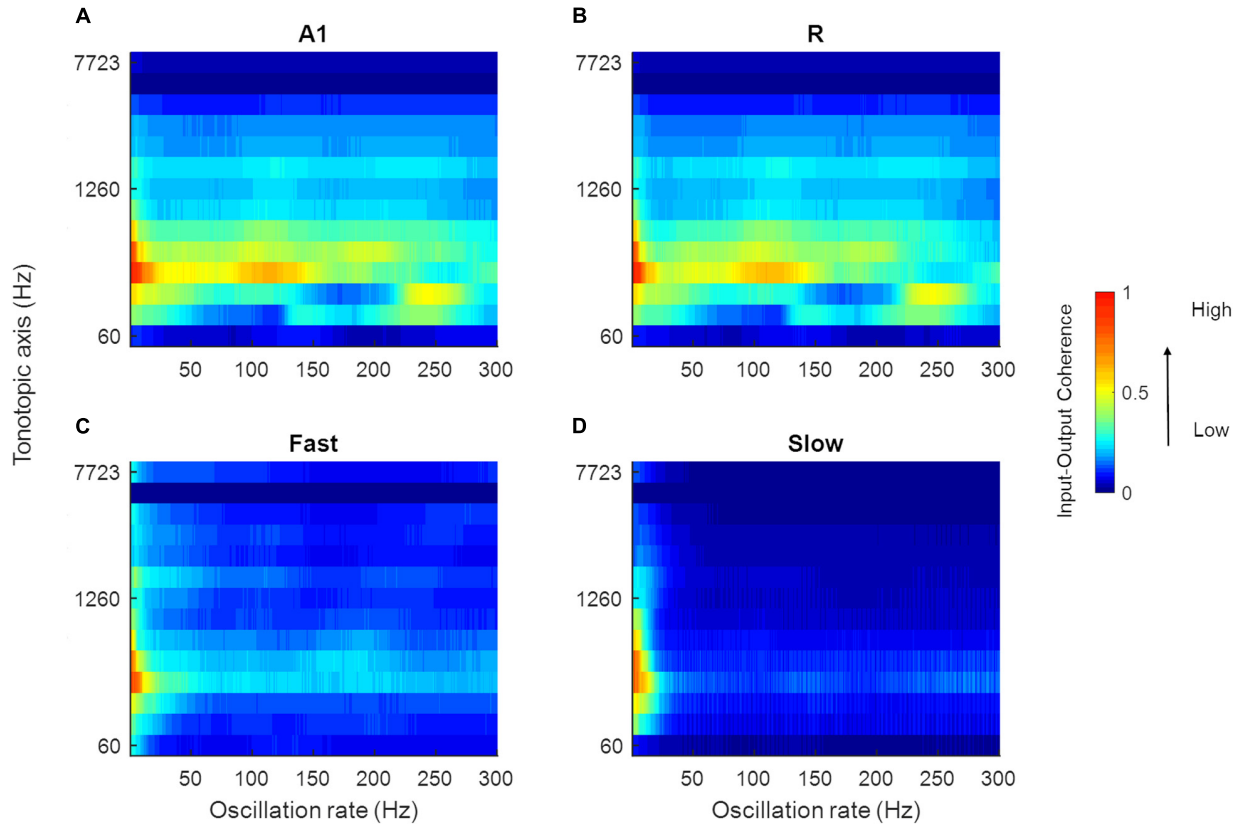


**FIGURE 9 |** Mean magnitude spectrum coherence between speech sounds and model output. The coherence values in A1, R, *Fast*, and *Slow* areas are shown in **(A–D)**, respectively (scaled by the normalized mean spatial response of the model to 630 speech sounds). All areas show high coherence with the slow oscillations present in the input signal (indicated by red and yellow colors).

measured by YIN algorithm is shown as the white boxes). The simulated *belt* regions show functional specialization to represent pitch spectrally (in the *Slow* area) and temporally (in the *Fast* area) in parallel streams.

Overall, the model responses to speech sounds highlight the presence of a distributed code for representing different temporal features of speech signals at the level of *belt* regions, but not for the *core* regions. Each *belt* area showed a functionally relevant specialization, as the temporal features highlighted by *Slow* and *Fast* areas are key structures of speech signals.

## DISCUSSION

In this study, we presented a computational model of the AC that consists of information processing streams optimized for processing either fine-grained temporal or spectral information. The model is employed to investigate the contribution of the different cortical streams in the representation and processing of basic acoustic features (i.e., temporal modulation, pitch) in the context of artificial and natural (speech) stimuli.

We started by simulating responses to artificial AM sounds. Electrophysiological studies have characterized AM coding by a dual mechanism of temporal (synchronization) and rate coding (Joris et al., 2004). In comparison with the phase-locking in the auditory nerve (reported up to 1.5–8 kHz in humans; Verschooten et al., 2019), the synchronization code has been measured to be comparatively diminished at the level of the cortex for human and non-human primates. The preferred AM rates have been reported as ranging from 1 to 50 Hz in monkeys (Steinschneider et al., 1980; Bieser and Müller-Preuss, 1996; Lu et al., 2001), despite neurons have been shown to synchronize as high as 200 Hz in monkeys (Steinschneider et al., 1980) and similar weak synchronization could be detected in humans with electrocorticography (Nourski et al., 2013). In agreement with these electrophysiology studies, our model exhibited a dual coding mechanism. While the contribution of a temporal code (synchronization) was strong up to a maximum of 50 Hz, synchronizations became weaker for higher modulation rates and were complemented with a rate code mechanism.

Furthermore, in electrophysiology, the maximum AM rate for which a temporal code is present has been reported to differ across fields of the AC (Liang et al., 2002). Caudal fields (i.e., regions belonging to the dorsal processing stream) are reported to be as fast as or even faster than the primary AC and synchronize with the stimulus envelope up to high AM rates. Instead the rostral field (i.e., part of the ventral processing stream) does not show a temporal code for AM sounds but

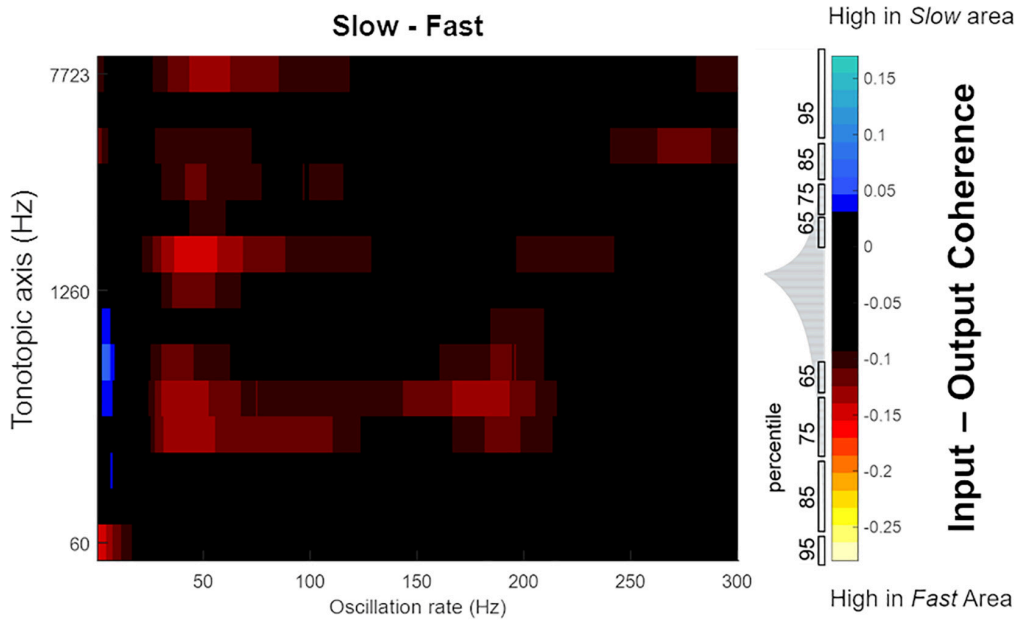


**FIGURE 10 |** Mean difference in magnitude spectrum coherence between belt regions. The *Slow* area showed higher coherence with slow oscillations in speech (4–8 Hz, indicated by blue colors). Instead, the *Fast* area showed greater coherence to faster oscillations of speech (30–70 Hz, around 100 and 150–200 Hz, indicated by the warmer colors). The distribution of difference in magnitude spectrum coherence between *Slow* and *Fast* area for all 630 sounds is shown in gray, adjacent to the color bar, with percentiles marked to indicate the statistical significance.

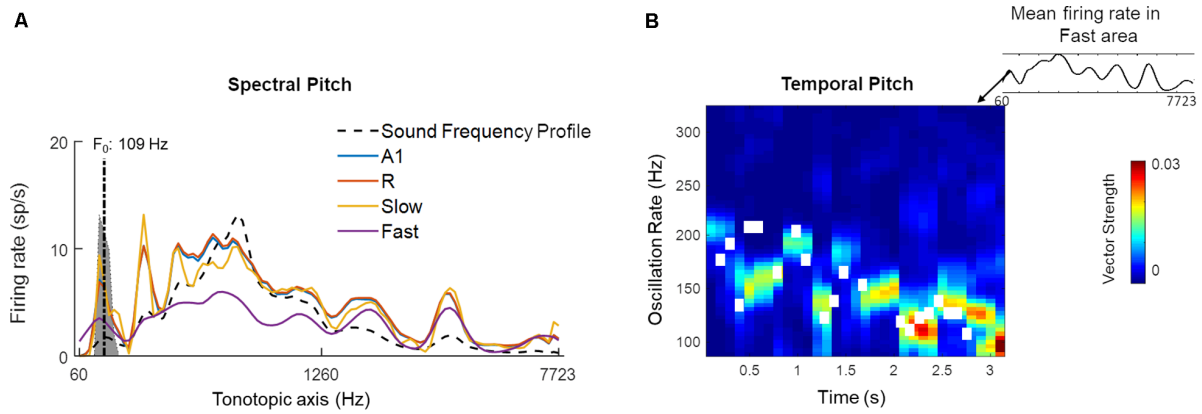


**FIGURE 11 |** A dual code for pitch estimation. For a sample sound, **(A)** Mean firing rate of all units in the four simulated areas (A1, R, *Slow*, and *Fast*, colored lines) is shown. Sound frequency profile (scaled) is plotted in black dashed line for reference. The gray highlighted portion of the plot indicates estimates of pitch by YIN algorithm (distribution over time, with best estimate of $F_0$ plotted with dash-dotted line; de Cheveigné and Kawahara, 2002). A spectral code is observed in model outputs with firing rate peaks overlapping with YIN estimates. **(B)** Temporal code for pitch is observed as weak synchronization to pitch contour in oscillatory activity (measured as Vector Strength) of the *Fast* area unit corresponding to spectral peak corresponding to best pitch estimate by YIN algorithm. The pitch contour estimates over time computed by YIN algorithm are depicted by white boxes. The correlation between YIN estimates and the temporal profile of the *Fast* area is 0.74.

instead codes AM with changes in firing rate (i.e., a rate code) (Bieser and Müller-Preuss, 1996). In the simulated responses, the relative contribution of the temporal and rate coding mechanisms also varied across the simulated cortical areas, depending upon the areas' temporal and spectral processing properties. While the temporal code displayed a low-pass filter profile, the shape of the rate code varied from low-pass to band-pass and band-stop patterns. Evidence for such variation in rate coding pattern has been reported in electrophysiological studies as well with sAM stimuli (Schreiner and Urbas, 1988; Bieser and Müller-Preuss, 1996; Liang et al., 2002; Bendor and Wang, 2008). In our model, this observation was highlighted when the firing rate was examined within carrier-matched frequency channels. The interaction of spectral and temporal response properties underlies these observations.

In order to assess the relationship between neural population activity (i.e., synchronization and firing rate) with human behavior, we next used the model to simulate

psychoacoustic experiments. We were able to successfully predict psychoacoustically-determined modulation detection thresholds (i.e., modulation detection transfer functions, tMTFs) for AM noise and tones (Bacon and Viemeister, 1985; Kohlrausch et al., 2000). The model suggested a role for auditory *core* areas, rather than *belt* areas, in coding modulation detection with simple AM stimuli. The tMTF for AM noise was replicated by computing temporal synchronization. However, for AM tones, we observed the best prediction of the psychoacoustical tMTF by using a combination of synchronization and spatial (sideband detection) code. Additionally, we observed that compared to low-frequency carriers, high carriers allowed modulation detection up to faster rates. This replicated psychoacoustic observations of detection up to faster modulation rates with a higher carrier frequency (Sek and Moore, 1995; Kohlrausch et al., 2000; Moore and Glasberg, 2001; Simpson et al., 2013). Our simulations indicate that these frequency-specific responses, which arise at the periphery, are inherited by the cortex, especially in the *core* areas.

We further evaluated the contribution of temporal coding mechanisms to psycho-acoustical phenomena. While current views on pitch perception suggest that the role of synchronization is limited to auditory periphery and cortex might use information from individual harmonics (Plack et al., 2014), there is evidence of temporal cues being used especially for unresolved harmonics for low pitch sounds (Bendor et al., 2012). The model successfully decoded the low frequency missing fundamentals of complex tones and showed dependence of strength of synchronization on the order of harmonics. By simulating a psychoacoustic task employing missing fundamental complex tones with varying unresolved harmonics, we further investigated the role of synchronization and its dependence on number and order of harmonics. The model output matched the previously reported human behavior performance through synchronization in the simulated neural responses, but not by a rate coding mechanism. That is, we could successfully replicate three key findings from Houtsma and Smurzynski (1990). First, the synchronization to the missing $F_0$ was stronger for the lower compared to higher-order harmonic sounds and second, it improved with an increasing number of components of complex tone. Third, only for the lower order harmonic sounds, the improvement in model performance was sharp when the number of components was increased from two to four and displayed a plateau when further components were added. Interestingly, the match between psychoacoustics and the model output was limited to the *Fast* area, suggesting a role for this fine-grained temporal processing stream in the extraction of the pitch using temporal cues. Additionally, using speech sounds, we further observed a strong spatial (spectral) pitch correlate (observed in all areas, strongest in *Slow* area) along with weaker oscillations tracking pitch contour (only in *Fast* area). However, the spatial code is not observable in model output for pitch with missing fundamental complex tones and suggests need for a more complex network to effectively detect pitch just from harmonic information in space. Moreover, the temporal code for pitch can benefit from feedback connectivity (Balaguer-Ballester et al., 2009) while precise interspike intervals can shed light on phase sensitivity of pitch perception (Huang and Rinzel, 2016). Thus, future model modifications can move from general (current) to more specific hypotheses of auditory processing.

Coding of pitch in the AC has been extensively investigated with fMRI, resulting in somewhat conflicting findings. While some studies pointed to lateral Heschl's Gyrus (HG) as a pitch center (Griffiths and Hall, 2012; Norman-Haignere et al., 2013; De Angelis et al., 2018), other studies showed that pitch-evoking sounds produced the strongest response in human planum temporale (PT) (Hall and Plack, 2009; Garcia et al., 2010). This disagreement may be due to differences between studies in experimental methods and stimuli. Our computational model provides an opportunity to merge these fMRI-based findings, as it allows for the efficient and extensive testing of model responses to a broad range of sounds. Based on the sounds we tested, observations of a pitch center in PT, part of the *Fast* stream, may be dominated by temporal pitch. Instead, human fMRI studies reporting a pitch area in lateral HG (Griffiths and Hall, 2012; Norman-Haignere et al., 2013; De Angelis et al., 2018), which is part of the *Slow* stream), maybe reflecting the spectral rather than the temporal processing of pitch. Our simulations suggest a functional relevance for temporal representations albeit through weak synchronization. These predictions are in line with evidence of synchronization in the AC contributing to the percept of pitch (up to 100 Hz) observed with MEG (Coffey et al., 2016) and require future studies with both high spectral and temporal precision data from the AC.

The distributed coding pattern shown by the different regions (i.e., coding of modulation detection thresholds by the *core* regions, coding of temporal pitch by the *Fast* area and spectral acuity by the *Slow* area of the *belt* stream) reflected a hierarchical processing scheme based on varying spectro-temporal properties of the neural populations. We then applied this modeling framework to the analysis of (continuous) speech with the aim of exploring the influence of basic neural processing properties on the representation and coding of speech. All modeled areas represented the slow oscillations present in speech (<20 Hz). In the belt areas, an additional distributed coding of temporal information was observed. That is, the optimization for coding slow temporal changes with high spectral precision in the *Slow* stream resulted in the coding of temporal oscillations in the lower 4–8 Hz frequency range. Processing properties similar to those of the *Slow* stream may thus be suited for coding spectral pitch and prosody in speech signals. Instead, optimization for processing fast temporal changes with low spectral precision in the *Fast* stream resulted in coding of temporal oscillations in the higher 30–70 and 100–200 Hz frequency ranges. Processing properties similar to those of the *Fast* stream may therefore instead be optimal for coding phonemes (consonants), and temporal pitch. In sum, we showed that the hierarchical temporal structure of speech may be reflected in parallel and through distributed mechanisms by the modeled areas, especially by simulated belt areas. This is in line with the idea that the temporal response properties of auditory fields contribute to distinct functional pathways (Jasmin et al., 2019).

The "division of labor" observed between the simulated processing streams provides predictions regarding cortical speech processing mechanisms. Specifically, the slowest oscillations,

representing the speech envelope, were coded in parallel across regions with different processing properties and may serve to time stamp the traces of different speech aspects belonging to the same speech utterance across streams. This may serve as a distributed clock: A binding mechanism that ensures the unified processing of different components of speech (Giraud and Poeppel, 2012; Yi et al., 2019) that are instead coded in a distributed fashion. Such a temporal code can also underlie binding of auditory sources in stream segregation (Elhilali et al., 2009). While in the current implementation of the model the responses are driven by stimuli, the model could be extended to include stimulus-independent oscillatory cortical activity. As the oscillations inherent to AC processing that occur on multiple timescales are known to decode complimentary informational structures in speech processing (Overath et al., 2015) and auditory scene analysis, such a model extension may in the future be used to study the effects on these 'inherent' oscillations on responses to speech and other structured inputs.

To summarize, we have presented a recurrent neural model built on simple and established assumptions on general mechanisms of neuronal processing and on the auditory cortical hierarchy. Despite its simplicity, the model was able to mimic results from (animal) electrophysiology and was useful to link these results to those of psychophysics and neuroimaging studies in humans. As the response properties of the AC (tonotopic organization, phase-locking, etc.) are inherited from the periphery, it remains possible that the model actually depicts earlier stages in the auditory pathway rather than AC. In future implementations of the model, the distinction between peripheral and cortical stages can benefit from a more detailed peripheral model (Meddis et al., 2013; Zilany et al., 2014). Ultimately, establishing a clear distinction between peripheral and cortical contribution would require simultaneous high-resolution (spatial and temporal) recordings across multiple locations of the auditory pathway and cortex. Furthermore, how the model dynamics shape up in presence of intrinsic noise in the system can also provide interesting insights into sound processing.

Nonetheless, the model is valuable for generating hypotheses on how the different cortical areas/streams may contribute toward behaviorally relevant aspects of acoustic signals. The presented model may be extended to include a physiological model of neurovascular coupling (Havlicek et al., 2017) and thus generate predictions that can be directly verified using functional MRI. Such a combination of modeling and imaging approaches is relevant for linking the spatially resolved but temporally slow hemodynamic signals to dynamic mechanisms of neuronal processing and interaction.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

## AUTHOR CONTRIBUTIONS

IZ and EF designed the model. IZ wrote the manuscript. All authors analyzed the model output. The manuscript was reviewed and edited by all authors.

## FUNDING

This work was supported by the Netherlands Organization for Scientific Research (NWO VICI Grant No. 453-12-002 to EF, and VENI Grant No. 451-15-012 to MM) and the Dutch Province of Limburg.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2019.00095/full#supplementary-material

## REFERENCES

Arnott, S. R., Binns, M. A., Grady, C. L., and Alain, C. (2004). Assessing the auditory dual-pathway model in humans. *Neuroimage* 22, 401–408. doi: 10.1016/j.neuroimage.2004.01.014

Bacon, S. P., and Viemeister, N. F. (1985). Temporal modulation transfer functions in normal-hearing and hearing-impaired listeners. *Audiology* 24, 117–134. doi: 10.3109/00206098509081545

Balaguer-Ballester, E., Clark, N. R., Coath, M., Krumbholz, K., and Denham, S. L. (2009). Understanding pitch perception as a hierarchical process with top-down modulation. *PLoS Comput. Biol.* 5:e1000301. doi: 10.1371/journal.pcbi.1000301

Bartlett, E. L., Sadagopan, S., and Wang, X. (2011). Fine frequency tuning in monkey auditory cortex and thalamus. *J. Neurophysiol.* 106, 849–859. doi: 10.1152/jn.00559.2010

Belin, P., and Zatorre, R. J. (2000). "What", "where" and "how" in auditory cortex. *Nat. Neurosci.* 3, 965–966. doi: 10.1038/79890

Bendor, D., Osmanski, M. S., and Wang, X. (2012). Dual-pitch processing mechanisms in primate auditory cortex. *J. Neurosci.* 32, 16149–16161. doi: 10.1523/JNEUROSCI.2563-12.2012

Bendor, D., and Wang, X. (2008). Neural response properties of primary, rostral, and rostrotemporal core fields in the auditory cortex of marmoset monkeys. *J. Neurophysiol.* 100, 888–906. doi: 10.1152/jn.00884.2007

Bieser, A., and Müller-Preuss, P. (1996). Auditory responsive cortex in the squirrel monkey: neural responses to amplitude-modulated sounds. *Exp. Brain Res.* 108, 273–284.

Camalier, C. R., D'Angelo, W. R., Sterbing-D'Angelo, S. J., de la Mothe, L. A., and Hackett, T. A. (2012). Neural latencies across auditory cortex of macaque support a dorsal stream supramodal timing advantage in primates. *Proc. Natl. Acad. Sci. U.S.A.* 109, 18168–18173. doi: 10.1073/pnas.1206387109

Chi, T., Ru, P., and Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* 118, 887–906. doi: 10.1121/1.1945807

Chrostowski, M., Yang, L., Wilson, H. R., Bruce, I. C., and Becker, S. (2011). Can homeostatic plasticity in deafferented primary auditory cortex lead to travelling waves of excitation? *J. Comput. Neurosci.* 30, 279–299. doi: 10.1007/s10827-010-0256-1

Coffey, E. B. J., Herholz, S. C., Chepesiuk, A. M. P., Baillet, S., and Zatorre, R. J. (2016). Cortical contributions to the auditory frequency-following response revealed by MEG. *Nat. Commun.* 7:11070. doi: 10.1038/ncomms11070

Cowan, J. D., Neuman, J., and van Drongelen, W. (2016). Wilson–cowan equations for neocortical dynamics. *J. Math. Neurosci.* 6:1. doi: 10.1186/s13408-015-0034-5

De Angelis, V., De Martino, F., Moerel, M., Santoro, R., Hausfeld, L., and Formisano, E. (2018). Cortical processing of pitch: model-based encoding and decoding of auditory fMRI responses to real-life sounds. *Neuroimage* 180(Pt A), 291–300. doi: 10.1016/j.neuroimage.2017.11.020

de Cheveigné, A., and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *J. Acoust. Soc. Am.* 111, 1917–1930. doi: 10.1121/1.1458024

Eggermont, J. J. (1991). Rate and synchronization measures of periodicity coding in cat primary auditory cortex. *Hear. Res.* 56, 153–167. doi: 10.1016/0378-5955(91)90165-6

Eggermont, J. J. (1998). Representation of spectral and temporal sound features in three cortical fields of the cat. Similarities outweigh differences. *J. Neurophysiol.* 80, 2743–2764. doi: 10.1152/jn.1998.80.5.2743

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., and Shamma, S. A. (2009). Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron.* 61, 317–329. doi: 10.1016/j.neuron.2008.12.005

Ermentrout, G. B., and Cowan, J. D. (1979). A mathematical theory of visual hallucination patterns. *Biol. Cybern.* 34, 137–150. doi: 10.1007/bf00336965

Galaburda, A., and Sanides, F. (1980). Cytoarchitectonic organization of the human auditory cortex. *J. Comp. Neurol.* 190, 597–610. doi: 10.1002/cne.901900312

Garcia, D., Hall, D. A., and Plack, C. J. (2010). The effect of stimulus context on pitch representations in the human auditory cortex. *Neuroimage* 51, 808–816. doi: 10.1016/j.neuroimage.2010.02.079

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., et al. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download*. Philadelphia: Linguistic Data Consortium.

Giraud, A. L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. *Nat. Neurosci.* 15, 511–517. doi: 10.1038/nn.3063

Glasberg, B. R., and Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hear. Res.* 47, 103–138. doi: 10.1016/0378-5955(90)90170-t

Goldberg, J. M., and Brown, P. B. (1969). Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization. *J. Neurophysiol.* 32, 613–636. doi: 10.1152/jn.1969.32.4.613

Griffiths, T. D., and Hall, D. A. (2012). Mapping pitch representation in neural ensembles with fMRI. *J. Neurosci.* 32, 13343–13347. doi: 10.1523/jneurosci.3813-12.2012

Hackett, T. A., Stepniewska, I., and Kaas, J. H. (1998). Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. *J. Comp. Neurol.* 394, 475–495. doi: 10.1002/(sici)1096-9861(19980518)394:4<475::aid-cne6>3.0.co;2-z

Hall, D. A., and Plack, C. J. (2009). Pitch processing sites in the human auditory brain. *Cereb. Cortex* 19, 576–585. doi: 10.1093/cercor/bhn108

Havlicek, M., Ivanov, D., Roebroeck, A., and Uludağ, K. (2017). Determining excitatory and inhibitory neuronal activity from multimodal fMRI data using a generative hemodynamic model. *Front. Neurosci.* 11:616. doi: 10.3389/fnins.2017.00616

Heil, P., and Irvine, D. R. F. (2017). First-spike timing of auditory-nerve fibers and comparison with auditory cortex. *J. Neurophysiol.* 78, 2438–2454. doi: 10.1152/jn.1997.78.5.2438

Houtsma, A. J., and Smurzynski, J. (1990). Pitch identification and discrimination for complex tones with many harmonics. *J. Acoust. Soc. Am.* 87, 304–310. doi: 10.1121/1.399297

Huang, C., and Rinzel, J. (2016). A neuronal network model for pitch selectivity and representation. *Front. Comput. Neurosci.* 10:57. doi: 10.3389/fncom.2016.00057

Jasmin, K., Lima, C. F., and Scott, S. K. (2019). Understanding rostral–caudal auditory cortex contributions to auditory perception. *Nat. Rev. Neurosci.* 20, 425–434. doi: 10.1038/s41583-019-0160-2

Joris, P. X., Schriener, C. E., and Rees, A. (2004). Neural processing of amplitude-modulated sounds. *Physiol. Rev.* 84, 541–577. doi: 10.1152/physrev.00029.2003

Kaas, J. H., and Hackett, T. A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11793–11799. doi: 10.1073/pnas.97.22.11793

Kaas, J. H., Hackett, T. A., and Tramo, M. J. (1999). Auditory processing in primate cerebral cortex. *Curr. Opin. Neurobiol.* 9, 164–170.

Kohlrausch, A., Fassel, R., and Dau, T. (2000). The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers. *J. Acoust. Soc. Am.* 108, 723–734. doi: 10.1121/1.429605

Kuśmierek, P., and Rauschecker, J. P. (2009). Functional specialization of medial auditory belt cortex in the alert rhesus monkey. *J. Neurophysiol.* 102, 1606–1622. doi: 10.1152/jn.00167.2009

Kuśmierek, P., and Rauschecker, J. P. (2014). Selectivity for space and time in early areas of the auditory dorsal stream in the rhesus monkey. *J. Neurophysiol.* 111, 1671–1685. doi: 10.1152/jn.00436.2013

Liang, L., Lu, T., and Wang, X. (2002). Neural representations of sinusoidal amplitude and frequency modulations in the primary auditory cortex of awake primates. *J. Neurophysiol.* 87, 2237–2261. doi: 10.1152/jn.2002.87.5.2237

Loebel, A., Nelken, I., and Tsodyks, M. (2007). Processing of sounds by population spikes in a model of primary auditory cortex. *Front. Neurosci.* 1, 197–209. doi: 10.3389/neuro.01.1.1.015.2007

Lu, T., Liang, L., and Wang, X. (2001). Temporal and rate representations of time-varying signals in the auditory cortex of awake primates. *Nat. Neurosci.* 4, 1131–1138. doi: 10.1038/nn737

Ma, N., Green, P., Barker, J., and Coy, A. (2007). Exploiting correlogram structure for robust speech recognition with multiple speech sources. *Speech Commun.* 49, 874–891. doi: 10.1016/j.specom.2007.05.003

May, P. J. C., Westö, J., and Tiitinen, H. (2015). Computational modelling suggests that temporal integration results from synaptic adaptation in auditory cortex. *Eur. J. Neurosci.* 41, 615–630. doi: 10.1111/ejn.12820

Meddis, R., Lecluyse, W., Clark, N. R., Jürgens, T., Tan, C. M., Panda, M. R., et al. (2013). A computer model of the auditory periphery and its application to the study of hearing. *Adv. Exp. Med. Biol.* 787, 11–20. doi: 10.1007/978-1-4614-1590-9_2

Moore, B. C. (2003). *An Introduction to the Psychology of Hearing*. Cambridge: Academic Press.

Moore, B. C., and Glasberg, B. R. (2001). Temporal modulation transfer functions obtained using sinusoidal carriers with normally hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.* 110, 1067–1073. doi: 10.1121/1.1385177

Norman-Haignere, S., Kanwisher, N., and McDermott, J. H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *J. Neurosci.* 33, 19451–19469. doi: 10.1523/JNEUROSCI.2880-13.2013

Nourski, K. V., Brugge, J. F., Reale, R. A., Kovach, C. K., Oya, H., Kawasaki, H., et al. (2013). Coding of repetitive transients by auditory cortex on posterolateral superior temporal gyrus in humans: an intracranial electrophysiology study. *J. Neurophysiol.* 109, 1283–1295. doi: 10.1152/jn.00718.2012

Nourski, K. V., Steinschneider, M., McMurray, B., Kovach, C. K., Oya, H., Kawasaki, H., et al. (2014). Functional organization of human auditory cortex: investigation of response latencies through direct recordings. *Neuroimage* 101, 598–609. doi: 10.1016/j.neuroimage.2014.07.004

Oshurkova, E., Scheich, H., and Brosch, M. (2008). Click train encoding in primary and non-primary auditory cortex of anesthetized macaque monkeys. *Neuroscience* 153, 1289–1299. doi: 10.1016/j.neuroscience.2008.03.030

Overath, T., McDermott, J. H., Zarate, J. M., and Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat. Neurosci.* 18, 903–911. doi: 10.1038/nn.4021

Oxenham, A. J. (2012). Pitch perception. *J. Neurosci.* 32, 13335–13338.

Patterson, R. D. (1986). "Auditory filters and excitation patterns as representations of frequency resolution," in *Frequency Selectivity in Hearing*, ed. B. C. J. Moore, (London: Academic), 123–177.

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., and Allerhand, M. (1992). "Complex sounds and auditory images," in *Proceedings of the 9th International Symposium Hearing Audit., Physiol. Perception*, Carcens, 429–446. doi: 10.1016/b978-0-08-041847-6.50054-x

Plack, C. J., Barker, D., and Hall, D. A. (2014). Pitch coding and pitch processing in the human brain. *Hear. Res.* 307, 53–64. doi: 10.1016/j.heares.2013.07.020

Rauschecker, J. P., and Tian, B. (2000). Mechanisms and streams for processing of "what" and "where" in auditory cortex. *Proc. Natl. Acad. Sci. U.S.A.* 97, 11800–11806. doi: 10.1073/pnas.97.22.11800

Rauschecker, J. P., Tian, B., Pons, T., and Mishkin, M. (1996). Serial and parallel processing in macaque auditory cortex. *J. Comp. Neurol.* 382, 89–103. doi: 10.1002/(sici)1096-9861(19970526)382:1<89::aid-cne6>3.3.co;2-y

Read, H. L., Winer, J. A., and Schreiner, C. E. (2002). Functional architecture of auditory cortex. *Curr. Opin. Neurobiol.* 12, 433–440. doi: 10.1016/s0959-4388(02)00342-2

Recanzone, G. H., Guard, D. C., and Phan, M. L. (2000). Frequency and intensity response properties of single neurons in the auditory cortex of the behaving macaque monkey. *J. Neurophysiol.* 83, 2315–2331. doi: 10.1152/jn.2000.83.4.2315

Rivier, F., and Clarke, S. (1997). Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: evidence for multiple auditory areas. *Neuroimage* 6, 288–304. doi: 10.1006/nimg.1997.0304

Romanski, L. M., Tian, B., Fritz, J., Mishkin, M., Goldman-Rakic, P. S., and Rauschecker, J. P. (1999). Dual streams of auditory afferents target multiple domains in the primate prefrontal cortex. *Nat. Neurosci.* 2, 1131–1136. doi: 10.1038/16056

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., et al. (2014). Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. *PLoS Comput. Biol.* 10:e1003412. doi: 10.1371/journal.pcbi.1003412

Schönwiesner, M., and Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proc. Natl. Acad. Sci. U.S.A.* 106, 14611–14616. doi: 10.1073/pnas.0907682106

Schreiner, C. E., and Urbas, J. V. (1988). Representation of amplitude modulation in the auditory cortex of the cat. II. Comparison between cortical fields. *Hear. Res.* 32, 49–63. doi: 10.1016/0378-5955(88)90146-3

Schreiner, C. E., Read, H. L., and Sutter, M. L. (2000). Modular organization of frequency integration in primary auditory cortex. *Annu. Rev. Neurosci.* 23, 501–529. doi: 10.1146/annurev.neuro.23.1.501

Sclar, G., Maunsell, J. H., and Lennie, P. (1990). Coding of image contrast in central visual pathways of the macaque monkey. *Vis. Res.* 30, 1–10. doi: 10.1016/0042-6989(90)90123-3

Scott, B. H., Malone, B. J., and Semple, M. N. (2011). Transformation of temporal processing across auditory cortex of awake macaques. *J. Neurophysiol.* 105, 712–730. doi: 10.1152/jn.01120.2009

Sek, A., and Moore, B. C. (1995). Frequency discrimination as a function of frequency, measured in several ways. *J. Acoust. Soc. Am.* 97, 2479–2486. doi: 10.1121/1.411968

Simpson, A. J. R., Reiss, J. D., and McAlpine, D. (2013). Tuning of human modulation filters is carrier-frequency dependent. *PLoS One* 8:e73590. doi: 10.1371/journal.pone.0073590

Steinschneider, M., Arezzo, J., and Vaughan, H. G. (1980). Phase-locked cortical responses to a human speech sound and low-frequency tones in the monkey. *Brain Res.* 198, 75–84. doi: 10.1016/0006-8993(80)90345-5

Tabas, A., Andermann, M., Schuberth, V., Riedel, H., Balaguer-Ballester, E., and Rupp, A. (2019). Modeling and MEG evidence of early consonance processing in auditory cortex. *PLoS Comput Biol.* 15:e1006820. doi: 10.1371/journal.pcbi.1006820

Tian, B., Reser, D., Durham, A., Kustov, A., and Rauschecker, J. P. (2001). Functional specialization in rhesus monkey auditory cortex. *Science* 292, 290–293. doi: 10.1126/science.1058911

Verschooten, E., Shamma, S., Oxenham, A. J., Moore, B. C. J., Joris, P. X., Heinz, M. G., et al. (2019). The upper frequency limit for the use of phase locking to code temporal fine structure in humans: a compilation of viewpoints. *Hear. Res.* 377, 109–121. doi: 10.1016/j.heares.2019.03.011

Wallace, M. N., Johnston, P. W., and Palmer, A. R. (2002). Histochemical identification of cortical areas in the auditory region of the human brain. *Exp. Brain Res.* 143, 499–508. doi: 10.1007/s00221-002-1014-z

Wilson, H. R. (1997). A neural model of foveal light adaptation and afterimage formation. *Vis. Neurosci.* 14, 403–423. doi: 10.1017/s0952523800012098

Wilson, H. R. (1999). *Computation by Excitatory and Inhibitory Networks in Spikes, Decisions & Actions: Dynamical Foundations of Neuroscience* (Oxford: Oxford University Press), 88–115.

Wilson, H. R., and Cowan, J. D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.* 12, 1–24. doi: 10.1016/s0006-3495(72)86068-5

Wilson, H. R., and Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetik* 13, 55–80. doi: 10.1007/bf00288786

Wilson, H. R., and Kim, J. (1994). Perceived motion in the vector sum direction. *Vis. Res.* 34, 1835–1842. doi: 10.1016/0042-6989(94)90308-5

Yarden, T. S., and Nelken, I. (2017). Stimulus-specific adaptation in a recurrent network model of primary auditory cortex. *PLoS Comput. Biol.* 13:e1005437. doi: 10.1371/journal.pcbi.1005437

Yi, H. G., Leonard, M. K., and Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron* 102, 1096–1110. doi: 10.1016/j.neuron.2019.04.023

Yost, W. A. (2009). Pitch perception. *Atten. Percept. Psychophys.* 71, 1701–1715. doi: 10.3758/APP.71.8.1701

Yost, W. A. (2010). Pitch perception. *Senses A Compr. Ref.* 3, 807–828.

Zilany, M. S., Bruce, I. C., and Carney, L. H. (2014). Updated parameters and expanded simulation options for a model of the auditory periphery. *J. Acoust. Soc. Am.* 135, 283–286. doi: 10.1121/1.4837815

**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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 22 January 2020
doi: 10.3389/fncom.2019.00092

# A Computational Model of Interactions Between Neuronal and Astrocytic Networks: The Role of Astrocytes in the Stability of the Neuronal Firing Rate

*Kerstin Lenk[1*†], Eero Satuvuori[1,2,3,4†], Jules Lallouette[5,6], Antonio Ladrón-de-Guevara[1], Hugues Berry[5,6] and Jari A. K. Hyttinen[1]*

[1] BioMediTech, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland, [2] Institute for Complex Systems (ISC), National Research Council (CNR), Sesto Fiorentino, Italy, [3] Department of Physics and Astronomy, University of Florence, Sesto Fiorentino, Italy, [4] Department of Human Movement Sciences, MOVE Research Institute Amsterdam, Vrije Universiteit Amsterdam, Amsterdam, Netherlands, [5] INRIA, Villeurbanne, France, [6] LIRIS UMR5205, University of Lyon, Villeurbanne, France

OPEN ACCESS

**Edited by:**
*Yu-Guo Yu,*
*Fudan University, China*

**Reviewed by:**
*Maurizio De Pittà,*
*Basque Center for Applied Mathematics, Spain*
*Xiaojuan Sun,*
*Beijing University of Posts and Telecommunications (BUPT), China*

***Correspondence:**
*Kerstin Lenk*
*lenk.kerstin@gmail.com*

*†These authors have contributed equally to this work*

**Received:** *24 June 2019*
**Accepted:** *20 December 2019*
**Published:** *22 January 2020*

**Citation:**
*Lenk K, Satuvuori E, Lallouette J, Ladrón-de-Guevara A, Berry H and Hyttinen JAK (2020) A Computational Model of Interactions Between Neuronal and Astrocytic Networks: The Role of Astrocytes in the Stability of the Neuronal Firing Rate. Front. Comput. Neurosci. 13:92. doi: 10.3389/fncom.2019.00092*

Recent research in neuroscience indicates the importance of tripartite synapses and gliotransmission mediated by astrocytes in neuronal system modulation. Although the astrocyte and neuronal network functions are interrelated, they are fundamentally different in their signaling patterns and, possibly, the time scales at which they operate. However, the exact nature of gliotransmission and the effect of the tripartite synapse function at the network level are currently elusive. In this paper, we propose a computational model of interactions between an astrocyte network and a neuron network, starting from tripartite synapses and spanning to a joint network level. Our model focuses on a two-dimensional setup emulating a mixed *in vitro* neuron-astrocyte cell culture. The model depicts astrocyte-released gliotransmitters exerting opposing effects on the neurons: increasing the release probability of the presynaptic neuron while hyperpolarizing the post-synaptic one at a longer time scale. We simulated the joint networks with various levels of astrocyte contributions and neuronal activity levels. Our results indicate that astrocytes prolong the burst duration of neurons, while restricting hyperactivity. Thus, in our model, the effect of astrocytes is homeostatic; the firing rate of the network stabilizes to an intermediate level independently of neuronal base activity. Our computational model highlights the plausible roles of astrocytes in interconnected astrocytic and neuronal networks. Our simulations support recent findings in neurons and astrocytes *in vivo* and *in vitro* suggesting that astrocytic networks provide a modulatory role in the bursting of the neuronal network.

**Keywords: simulation, neuron, astrocyte, network, calcium signaling, gliotransmission**

## INTRODUCTION

Neuroscience research has focused for long on neurons and their interacting networks. However, the brain also consists of a large number of other different cell types, among which glial cells represent roughly 50% of the brain cells (Kettenmann and Verkhratsky, 2008; Azevedo et al., 2009). Among glial cells, astrocytes offer metabolic support to neurons, regulate the extracellular ions like potassium and calcium released upon neuronal activity (Dallérac et al., 2013; Hertz et al., 2015) and uptake neurotransmitters (Bezzi et al., 1998; Araque et al., 2001; Perea and Araque, 2007; Volterra et al., 2014). Indeed, some of the synapses of the central nervous system are contacted by astrocytes that wrap around them, thus forming a structural ensemble called the tripartite synapse: presynaptic neuron, post-synaptic neuron and the ensheathing astrocyte (Araque et al., 1999).

Intracellular calcium ($Ca^{2+}$) transients are a prominent readout signal of astrocyte activity, and happens at different time scales (Kastanenka et al., 2019). They may be triggered by neuronal activity (Di Castro et al., 2011; Dallérac et al., 2013). At glutamatergic synapses, inositol 1,4,5-trisphosphate ($IP_3$) is released in the astrocyte cytoplasm after some of the presynaptically released glutamate binds to metabotropic glutamate receptors in the astrocytic plasma membrane. The released $IP_3$ binds to $IP_3$- and $Ca^{2+}$-gated $Ca^{2+}$ channels in the membrane of the endoplasmic reticulum, thus leading to a $Ca^{2+}$ elevation in the astrocyte cytosol. In return, these transient changes in the level of free cytoplasmic $Ca^{2+}$ lead to the opening of further $IP_3$ channels in a $Ca^{2+}$-induced $Ca^{2+}$ release (CICR) mechanism that further amplifies $Ca^{2+}$ release from the endoplasmic reticulum. The internal calcium pathways may also be linked to the release by the astrocyte of so-called gliotransmitters—like glutamate, D-serine, adenosine triphosphate (ATP), and GABA (γ-aminobutyric acid)—that influence the activity of the contacted neurons (Pasti et al., 2001; Henneberger et al., 2010; Zorec et al., 2012; Araque et al., 2014; Sahlender et al., 2014).

Neuron-astrocyte interactions are thought to occur—or be initiated—at the thinnest astrocytic processes/branchlets (Bazargani and Attwell, 2016; Bindocci et al., 2017). Furthermore, astrocytes themselves form interconnected networks via gap junctions. Gap junctions formed by connexins build a pore through the cell membranes of two adjacent astrocytes, joining their cytosols and letting through certain sized molecules, including $IP_3$ and potassium ions (Fellin, 2009; Giaume et al., 2010). The modulating effect of astrocytes on neuronal network activity has been shown in several *in vitro* experiments. Tukker et al. (2018) showed that the spike and burst rates were reduced in matured networks with glutamatergic neurons and astrocytes compared to glutamatergic neurons only. Co-cultured human stem cell-derived neurons and astrocytes exhibited a marginal decrease in the spike rate and an increase in the burst rate and duration, while the number of spikes per bursts was constant when more astrocyte were present in the network (Paavilainen et al., 2018).

Dedicated computational models of the cross-talk between neuron networks and astrocytes have been successfully employed to explore specific issues related to neuron-astrocyte interactions (for a review, see Oschmann et al., 2018). For example, Amiri et al. (2013) combined two coupled Morris-Lecar neuron models and the dynamic astrocyte model of Postnov et al. (2009). They simulated 50 pyramidal neurons, 50 interneurons, and 50 astrocytes, connected in a chain-like manner, with each astrocyte connected to one pyramidal cell, one interneuron, and one neighboring astrocyte via gap junctions. This study suggested that increasing the influence of the astrocytes toward the neurons leads to a reduction of the synchronized neuronal oscillations. Valenza et al. (2013) developed a transistor-like description of the tripartite synapse and also included short-term synaptic plasticity for excitatory synapses. They simulated a network containing 1,000 neurons and 1,500 astrocytes where at least one astrocyte was linked to each neuron. This model was able to produce spontaneous polychronous activity—i.e., reproducible time-locked but not synchronous firing—in neural groups.

More recently, Aleksin et al. (2017) presented neural network simulation software called ARACHNE, which is partially based on the NEURON environment. This model includes a chain-like structure in ring form, basic equations for the internal astrocytic dynamics and extracellular diffusion of gliotransmitters (volume transmission). Additionally, Stimberg et al. (2019) recently presented how the Brian 2 simulator can be used to model networks of interacting neurons and astrocytes. The authors notably showed how, after a period of high external stimulation of the neurons, gliotransmission can maintain a high level of neuronal activity and firing synchrony for several seconds after the end of the external stimulation. Although those modeling studies clearly advanced our understanding of the interaction between neuron networks and astrocyte networks, few of them included all three of the following significant ingredients of astrocyte networks: (i) Astrocytes form gap junction-based networks that convey calcium-based signals as waves (Charles et al., 1996; Fellin, 2009); (ii) each astrocyte contacts a large number of synapses, estimated to be up to 100,000 synapses per astrocyte in rat hippocampus (Bushong et al., 2002); and (iii) astrocytes can release distinct types of gliotransmitters (Di Castro et al., 2011; Sahlender et al., 2014; Schwarz et al., 2017), for instance, a single hippocampal astrocyte can co-release both excitatory (glutamate) and depressing gliotransmitters (adenosine), thus exerting a biphasic control of the synapse (Covelo and Araque, 2018).

In this work, we develop a mathematical model of combined astrocyte-neuron networks to study the role of astrocyte networks on the modulation of the neuronal firing rate. In our model, which we call INEXA, astrocytes regulate neuronal communication through the tripartite synaptic function, and they can release both excitatory and depressing gliotransmitters in response to synaptic activity. We moreover introduce the biological property that each astrocyte is connected to hundreds of synapses. In a two-dimensional spatial setup emulating neuron-astrocyte co-cultures, we study how astrocytes control the homeostasis in neuronal networks by increasing the ratio of astrocytes. Further, we assess how the level of neuronal

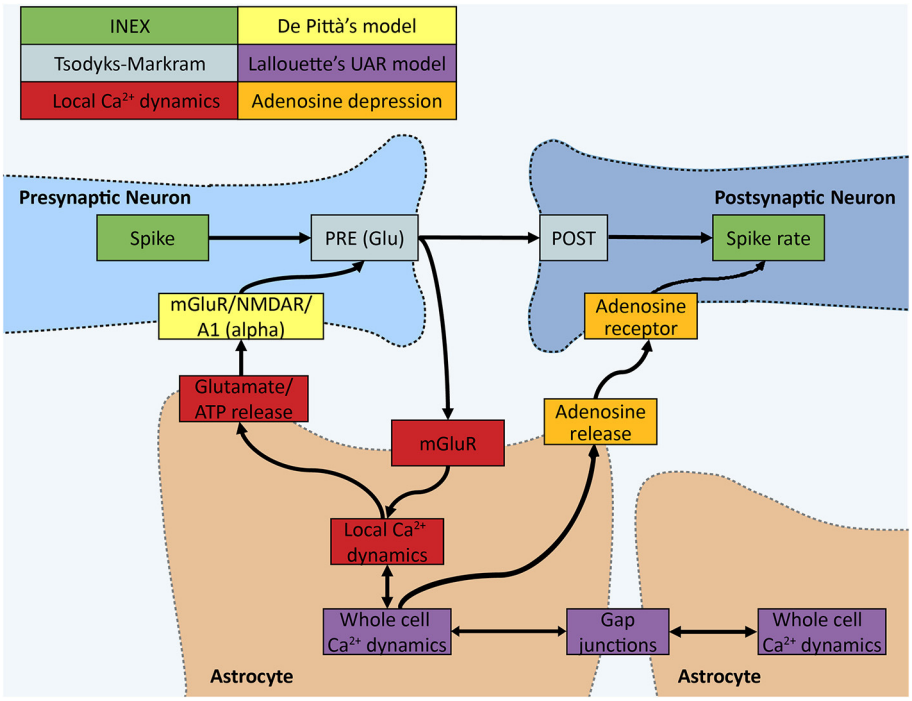


**FIGURE 1 |** Schematic of the INEXA model. The colors represent different parts of the simulator. In the INEX model by Lenk et al. (green), the spike has an effect on the spiking rate of the post-synaptic neuron through the synaptic weight. We added the Tsodyks-Markram (gray) synapse model together with De Pittà's astrocyte gliotransmitter interface (yellow). To monitor the synapse activity, a local calcium dynamics simulator (red) was added to each synapse, which is controlled by an astrocyte. Local astrocyte dynamics control gliotransmission to the synapse. All the local calcium simulators can have an effect on the whole cell calcium signaling modeled in the UAR model (purple) by Lallouette et al. In the UAR model, the calcium activity can spread across cells, mimicking calcium wave propagation through gap junction-mediated $IP_3$ diffusion. A whole cell calcium signal sets the local calcium dynamics to a high calcium state and ATP (quickly degraded into adenosine, orange part) is released into the extracellular space by the astrocyte to restrict the spiking of neurons nearby.

input can alter both the neuronal firing rate and the astrocytic calcium activity.

## METHODS

We developed a computational model that integrates the key components of astrocyte-neuron modulation (**Figure 1**). In section INEXA: A Computational Framework to Model Neuron-Astrocyte Networks, we describe the full INEXA model including the neuronal and astrocytic components and the manner in which they are coupled with each other. In section Numerical and Analysis Methods, we describe the numerical methods for analyzing the simulated neuronal and astrocytic activity. The outline of the simulations is specified at the end of section Numerical and Analysis Methods.

## INEXA: A Computational Framework to Model Neuron-Astrocyte Networks

### Neuronal Components

#### *Neuronal activity*

Our goal was to develop a model of neuronal spiking in primary mixed cultures (i.e., containing neurons and astrocytes) grown on multielectrode arrays (MEAs). We based our model on the phenomenological INEX model (Lenk, 2011), since it was initially built for *in vitro* neuronal networks. INEX is a stochastic cellular automaton in which inhibitory and excitatory neurons are connected to each other via synapses. Moreover, noise is applied to each neuron to reproduce background activity. In this fashion, INEX is a computationally-light model that has also been shown of well-reproducing neuronal dynamics of neuronal cultures plated on MEAs (Lenk, 2011; Lenk et al., 2016). For all these reasons, we adopted it as a starting platform for neuronal networks to be complemented by astrocytic coupling.

Briefly, INEX is a discrete-time model with a time step $t_k = \Delta t$. The instantaneous firing rate $\lambda_i$ of neuron $i$ in time slice $t_k$ is calculated as (Lenk, 2011):

$$\lambda_i(\mathbf{t_k}) = \max\left(\mathbf{0},\ \mathbf{c_i} + \sum_j \mathbf{y_{ij}} \mathbf{s_j}(\mathbf{t_{k-1}})\right) \quad (1)$$

where $c_i$ is the noise of neuron $i$ and $y_{ij}$ the synaptic strength from presynaptic neuron $j$ to post-synaptic neuron $i$. For each neuron, the value of $c_i$ was set independently by sampling from a triangular distribution between 0 and an upper bound, $C_{max}$. The value of $C_{max}$ depends on the simulation, in order to explore the effects of the noise level (see **Table 1**). The term $s_j$ indicates whether a spike has been emitted by neuron $j$ in the previous time step ($s_j = 1$ if a spike has been emitted, else $s_j = 0$).

**TABLE 1 |** Basic simulation parameters.

| Parameter | Value | Unit | Definition |
|---|---|---|---|
| $C_{max}$ | 0.01; 0.02; 0.03 | – | Upper boundaries for the three noise levels |
| $Y^{+}_{max}$ | 0.7 | – | Upper boundary for excitatory synaptic weights |
| $Y^{-}_{max}$ | −0.7 | – | Upper boundary for inhibitory synaptic weights |
| $\Omega_d$ | 4.0405 | $s^{-1}$ | Recovery rate of synaptic vesicles |
| $\Omega_f$ | 2.0 | $s^{-1}$ | Rate of synaptic facilitation |
| α | 0.7 | – | Effect parameter of astrocyte regulation of synaptic release |
| $\Omega_g$ | 0.077 | $s^{-1}$ | Recovery rate of gliotransmitter receptors |
| $g_r$ | 0.3 | – | Fraction of unbound receptors recruited by gliotransmission |
| $Ca_{th}$ | 0.1 | – | Calcium threshold for gliotransmitter release |
| $\Omega_{acc}$ | 0.05 | – | Accumulation rate between $IP_3$ and $Ca^{2+}$ |
| $\Omega_{IP_3}$ | 152.3 | $s^{-1}$ | $IP_3$ degradation rate |
| Astrocytes | 28; 63; 107 | – | Number of astrocytes for NN+A(10%), NN+A(20%) and NN+A(30%), respectively |
| M | 5 | – | Multiplier between astrocyte near synapse and whole astrocyte self-induced $IP_3$ flux |
| Connection distance | 100 | $\mu m$ | Maximum distance between two connected astrocytes |
| $\tau_A$ | 1.5 | $s$ | Average activation time of an astrocyte |
| $\tau_R$ | 7.0 | $s$ | Average refractory time of an astrocyte |
| $\tau_U$ | 5.0 | $s$ | Average time needed to activate an astrocyte |
| $b_0$ | 0.02 | – | Slope of the activation threshold |
| $b_1$ | 0.205 | – | Intercept of the activation threshold |
| $y_{Astro}$ | 0.01 | – | Depressing signal applied by astrocytes |
| Culture area | [750 750 10] | $\mu m$ | Resamples MEA electrode area for each dimension |
| Min. neuron distance | 10 | $\mu m$ | Minimum distance between randomly placed neurons |
| Min. astrocyte distance | 30 | $\mu m$ | Minimum distance between randomly placed astrocytes |
| $\sigma_N$ | 200 | $\mu m$ | Standard deviation of neuronal connections |
| $\sigma_A$ | 150 | $\mu m$ | Standard deviation of astrocyte-neuron connections without limiter |
| $d_A$ | 70 | $\mu m$ | Limiter cutting the Gaussian standard deviation connection probability set by standard deviation |
| T | 300 | $s$ | Simulation time |

*To keep the model as computationally light as possible and to maintain biological plausibility, the previously introduced models are combined using relatively simple components that are not accurate descriptions of the processes, but rather descriptive. The parameters in INEX are phenomenological and were fixed using brute force to find sets of parameters that produced results in reasonable ranges (Lenk et al., 2016). By adding the Tsodyks-Markram presynapse model, we introduced short-term memory at the level of individual synapses. The parameters are adapted from the model of De Pittà et al. (2011), which uses approximations of the local astrocytic calcium and $IP_3$. For the implementation of the UAR model, the parameters described in the supplementary part of the paper by Lallouette et al. (2014) are used. The values of the adenosine depression are chosen in such a way, that the astrocyte can reduce the probability of the neuronal spiking but cannot shut it down completely (Yoon and Lee, 2014). The basic principle of building our neuronal and astrocytic network topologies is that it reasonably represents a cultured network on an in vitro multielectrode array (Wallach et al., 2014; Paavilainen et al., 2018; Tukker et al., 2018). The figure of 250 neurons was found to be computationally fast enough, since several runs are needed to optimize parameters and produce comparable statistics. Astrocytes are set randomly but at least 30 μm apart. The simulation does not take into account the exact microdomains (Bushong et al., 2002; Agarwal et al., 2017) occupied by astrocytes, but assumes that the shape of the astrocytes allows them to occupy spaces that are non-uniformly spread around the cell soma.*

Note that, in our model, each excitatory presynapse is connected to an astrocyte with a probability that decreases with the distance between the synapse and the soma of the astrocyte (see Neuron and Astrocyte Network Spatial Topologies). We thus have thus adapted Equation (1) to account for the effect of astrocytes on the synapse (see Glial Components).

The probability $P_i(t_k)$ for neuron $i$ to emit a spike during time step $k$—i.e., between $t_k$ and $t_k + \Delta t$—is then modeled as an inhomogeneous Poisson process with rate $\lambda_i(t_k)$:

$$\mathbf{P_i}\,(\mathbf{t_k}) = \mathbf{e}^{-\lambda_i(\mathbf{t_k})\Delta \mathbf{t}}.\lambda_{\boldsymbol{i}}\,(\mathbf{t_k})\,\boldsymbol{\Delta t}. \tag{2}$$

Here, we used $\Delta t = 5\,\text{ms}$ to cover the typical duration of an action potential and the subsequent refractory period. Thus, we neglected the probability that more than one spike may be emitted by a given neuron during a single time step. At the benefit of computational efficiency, a time step as large as $\Delta t = 5\,\text{ms}$ can be adopted and the INEX network model can still reliably simulate neuronal activity recorded in MEA cultures (Lenk, 2011; Lenk et al., 2016).

### *Presynaptic dynamics*

For the dynamics of presynaptic neuronal release, we used the Tsodyks-Markram (TM) presynapse model (Tsodyks et al., 1998). The TM model consists of two variables, $x$ and $u$, describing the fraction of neurotransmitters available in the presynaptic terminal and the fraction of these available neurotransmitters that are ready for release (which can be seen as the release probability), respectively. We have discretized the original TM equations and thus, for each synapse $ij$ applied:

$$\mathbf{x}_{ij}(t_k) = \left(\mathbf{x}_{ij}(\mathbf{t}_{k-1}) - \mathbf{RR}_{\mathbf{ij}}(t_k)\right) + \left[1 - \left(\mathbf{x}_{\mathbf{ij}}(\mathbf{t}_{k-1}) - \mathbf{RR}_{\mathbf{ij}}(t_k)\right)\right]\left(1 - \mathbf{e}^{-\Omega_d \mathbf{t}}\right), \quad (3)$$

$$\mathbf{u}_{ij}(t_k) = \left[\left(1 - \mathbf{u}_{ij}(\mathbf{t}_{k-1})\right) U^*_{ij}(t_k)\, s_j(t_k) + \mathbf{u}_{ij}(t_{k-1})\right] \mathbf{e}^{-\Omega_f \Delta \mathbf{t}}, \quad (4)$$

$$RR_{ij}(t_k) = \mathbf{x}_{\mathbf{ij}}(\mathbf{t}_{k-1}) \left[\left(1 - u_{ij}(t_{k-1})\right) U^*_{ij}(t_k) s_j(t_k) + u_{ij}(t_{k-1})\right] s_j(t_k), \quad (5)$$

where $\Omega_d$ represents the rate of reintegration of neurotransmitters in the presynaptic terminal, $\Omega_f$ the rate of decrease of release probability, $RR_{ij}$ the fraction of released neurotransmitters, and $U^*_{ij}$ denotes the maximal increment of the ready-for-release fraction triggered by the arrival of a presynaptic spike.

The discretization of the TM equations was achieved by assuming that neuronal spikes happen at the very start of the 5 ms time steps. Just after a spike at the start of time step $t_k$, the release probability $u$ takes the value $\left(1 - u_{ij}\left(t_{k-1}\right)\right) U^*_{ij}(t_k) s_j(t_k) + u_{ij}\left(t_{k-1}\right)$: the sum of its previous values at the end of time slice $t_{k-1}$ and the additional recruitment of a fraction $U^*_{ij}$ of the previously non-recruited available resources. This temporary value $u$ just after a spike is used to compute: (1) the value of $u$ at the end of the time step $t_k$ (Equation 4) by applying a simple exponential decay term, and (2) the released resources for this time slice (Equation 5) by simply multiplying it by the fraction of available resources $x$ at the end of time step $t_{k-1}$. The available resources at the end of time step $t_k$ are then computed (Equation 3) by subtracting the released resources from the available resources at the end of time step $t_{k-1}$ and then applying an exponential term accounting for the reintegration of resources. In our model, the value of $U^*_{ij}$ in turn varies with time depending on gliotransmitter release by the astrocyte that enwraps the synapse (see Glial Components).

The strength of the synapse $y_{ij}$ was chosen to be directly proportional to the fraction of released resources $RR_{ij}$:

$$\mathbf{y}_{ij}(\mathbf{t}_k) = \mathbf{Y}_{\max} \cdot \mathbf{RR}_{\mathbf{ij}}(\mathbf{t}_k), \quad (6)$$

where $Y_{max}$ represents the largest value that the inhibitory ($Y^-_{max}$) or excitatory ($Y^+_{max}$) strength of a synapse can take.

## Glial Components

### *Regulation of synaptic dynamics by gliotransmission*

The questions of whether gliotransmitters are actually released by astrocytes and whether released gliotransmitters do contribute to the modulation of neuronal activity are still debated (see e.g., the two main perspectives expressed in Fiacco and McCarthy, 2018; Savtchouk and Volterra, 2018). In particular, the mechanisms by which gliotransmitters can be released are unclear, although both calcium-dependent vesicular release and channel-based release have been evidenced (Sahlender et al., 2014). However, an increasing number of experiments confirm that astrocytes are not just passive read-out units; they are heavily involved in the modulation of neuronal synapses and their activity (Fellin et al., 2004; Perea et al., 2009; Clarke and Barres, 2013). These results show that depending on the type of receptors expressed by the presynaptic and post-synaptic neurons, astrocyte-released glutamate can either potentiate (via presynaptic or extrasynaptic NMDAR) or depress the synapse (via presynaptic mGluR; Jourdain et al., 2007; Fellin, 2009; Bonansco et al., 2011; Min et al., 2012; Papouin and Oliet, 2014).

In addition to glutamate, astrocytes can also release purines such as ATP and adenosine (Newman, 2003; Bowser and Khakh, 2007; Lorincz et al., 2009; Hines and Haydon, 2014). Moreover, extracellular ATP of astrocytic origin could also be hydrolyzed into adenosine. By binding to A1 receptors on the presynaptic terminal, adenosine has been shown to reduce synaptic strength (Boddum et al., 2016; Savtchouk and Volterra, 2018). In a very similar way, astrocytes have also been reported to release GABA, a phenomenon involved in tonic inhibition (McIver et al., 2013), probably via calcium-regulated channels (Lee et al., 2010). Therefore, converging experimental evidence suggests that astrocytes release gliotransmitters that can either increase or decrease synaptic activity. In neurons, segregation between inhibitory and excitatory transmission is the rule. Excitatory neurons usually release glutamate, whereas inhibitory neurons release GABA, although exceptions exist, including the co-release of GABA and glutamate by the same presynaptic synapse (Shrivastava et al., 2011). However, the only available related experimental report on astrocytes concluded against segregation: in hippocampal slices, it was shown that a single astrocyte can release both glutamate and adenosine, thus mediating an initial potentiation of the synapse, followed by longer-lasting depression (Covelo and Araque, 2018). Lorincz et al. (2009) and Newman (2003) suggested in their studies that adenosine could also bind to A1 receptors post-synaptically and trigger neuronal inhibition through G protein-coupled inwardly rectifying $K^+$ channels.

In the present work, we explore the effects of such a non-segregated gliotransmitter release, assuming that a single astrocyte can release both potentiating and depressing gliotransmitters. Therefore, we assumed that gliotransmitter release is not segregated in astrocytes—i.e., a single astrocyte can release both potentiating and depressing gliotransmitters at the same synapse. To model the effect of depressing gliotransmitters, we added to each excitatory synapse contacted by an astrocyte an additional depressing signal from the astrocyte that could be mediated by adenosine (Newman, 2003; Lorincz et al., 2009). This was accounted for in the model by a term modulating the synaptic weights $y_{Astro}$, that modified Equation (1) to:

$$\lambda_i(\mathbf{t}_k) = \max\left(\mathbf{0}, c_i + \sum_j \mathbf{y}_{ij} \cdot \mathbf{s}_j(\mathbf{t}_{k-1}) - \sum_j \mathbf{y}_{Astro} \cdot \mathbf{A}_{ija}(\mathbf{t}_{k-1})\right), \quad (7)$$

where $A_{ija} = 1$ if synapse $ij$ is enwrapped by astrocyte "$a$" and if astrocyte "$a$" was in the active state at the previous time-step, else $A_{ija} = 0$ (the conditions for astrocyte activation are detailed in section Astrocytic network dynamics). Therefore, if an astrocyte is close enough to synapse $ij$ to enwrap it, the astrocyte exerts a depressing effect, $y_{Astro}$, on the synapse as long as the astrocyte is in the active state. Note that the duration of the resulting depression is set by the time spent by the astrocyte in the active state. In our simulations, this activation time is usually large (seconds, **Figure 5D**).

To model the effects of potentiating gliotransmitter release on the presynaptic part, we followed a paper by De Pittà et al. (2011), wherein a single parameter, $\alpha$, is used to describe the effects of the co-operation of multiple receptors. We considered that ATP and glutamate are released in a single release event and that their binding kinetics to their receptors are fairly similar. As in De Pittà et al. (2011) and De Pittà (2019), $\alpha$ modifies the value of $U^*_{ij}(t_k)$, which describes the effect of gliotransmission on the synaptic release probability (see section "Presynaptic dynamics"):

$$\mathbf{U^*_{ij}}(\mathbf{t_k}) = \frac{\mathbf{y_{ij_{base}}}}{\mathbf{Y_{max}}} \cdot \left(\mathbf{1} - \mathbf{g_{ij}}(\mathbf{t_k})\right) + \alpha \cdot \mathbf{g_{ij}}(\mathbf{t_k}), \quad (8)$$

where $g_{ij}(t_k)$ is the fraction of bound presynaptic gliotransmitter receptors (see section Astrocyte response to presynaptic stimulations). In the absence of gliotransmission, i.e., for the synapses that are not connected by an astrocyte, $g_{ij}(t_k) = 0$ for all time steps $t_k$, so that $U^*_{ij}$ is set to a constant value ($U^*_{ij} = \frac{y_{ij_{base}}}{Y_{max}}$). The value of α sets the influence of gliotransmission on presynaptic release: depending on its value, $\alpha$ can account for depressing gliotransmission ($0 < \alpha < \frac{y_{ij_{base}}}{Y_{max}}$) or potentiating gliotransmission ($\frac{y_{ij_{base}}}{Y_{max}} < \alpha < 1$). Here, our focus is on the non-segregated gliotransmitter release as reported by Covelo and Araque (2018), where a single astrocyte can sequentially elicit sequentially a potentiation of the synaptic weights followed by a longer-lasting depression. The latter phase is accounted for by the term $y_{\text{Astro}} A_{ija}$ in Equation (7). We thus emulate the initial potentiation phase by setting $\alpha$ to a potentiating value ($\alpha = 0.7$ while $\frac{y_{ij_{base}}}{Y_{max}} < 0.7$; see below and **Table 1**). The parameter $y_{ij_{base}}$ is the basal synaptic strength of synapse $ij$ in the absence of gliotransmission: a spike arriving at the presynaptic terminal of synapse without an adjacent astrocyte that has fully recovered from its previous activity (i.e., $x_{ij}(t_{k-1}) = 1$ and $u_{ij}(t_{k-1}) = 0$), yields $y_{ij}(t_k) = y_{ij_{base}}$ from Equations (5–7) above. In our model, $y_{ij_{base}}$ was sampled randomly from a triangular distribution ($0 \leq y_{ij_{base}} \leq 0.7$). The triangular distribution was a simplification of the Gaussian distribution, which guaranteed the positivity of the values.

### *Astrocyte response to presynaptic stimulations*

Calcium transients in astrocytes can be classified into at least two main types. Transient calcium elevations can happen independently of neuronal activity (spontaneous transients) or they can be triggered by the activity of nearby presynaptic neurons (activity-driven transients) (Perea et al., 2009; Wallach et al., 2014). Although astrocytic calcium signals can invade the whole cell (Volterra et al., 2014; Bindocci et al., 2017) and even be transmitted to coupled astrocytes (Parri et al., 2001), some calcium signals are restricted to the neighborhood of their origin. Thus, they cause calcium elevation locally, at a range of only one or a few synapses (Perea et al., 2009; Di Castro et al., 2011; Bindocci et al., 2017).

To account for the response of the astrocyte to glutamate release by the presynaptic element of the tripartite synapse, we modeled each astrocyte as a multi-compartment cell with local areas and a soma. Local area *ija* of astrocyte "*a*" represents the subpart of the astrocyte that is in direct contact with synapse *ij* and is associated to its own local $IP_3$ and calcium dynamics. Here, we expressed those local $IP_3$ and calcium transients using a simplified version of the astrocyte $IP_3$/calcium dynamics described by De Pittà and co-workers (De Pittà et al., 2008, 2019). The variables $[IP_3]$ and $[Ca^{2+}]$ denote the concentrations of $IP_3$ and $Ca^{2+}$, respectively in local area *ija* of astrocyte "*a*". Upon emission of a presynaptic spike by neuron *j*, $[IP_3]_{ija}(t_k)$ is incremented by a value that depends on the amount of resources released into the synaptic cleft, $RR_{ija}(t_k)$. $[IP_3]_{ija}(t_k)$ then decreases exponentially fast at rate $\Omega_{IP_3}$:

$$\begin{aligned}\mathbf{[IP_3]_{ija}}(\mathbf{t_k}) &= \mathbf{[IP_3]_{ija}}(\mathbf{t_{k-1}}) \cdot \mathbf{e}^{-\Omega_{\mathbf{IP_3}}\Delta\mathbf{t}} \\ &+ \left(\mathbf{1} - \mathbf{[IP_3]_{ija}}(\mathbf{t_{k-1}}) \cdot \mathbf{e}^{-\Omega_{\mathbf{IP_3}}\Delta\mathbf{t}}\right) \cdot \mathbf{RR_{ij}}(\mathbf{t_k}). \quad (9)\end{aligned}$$

To express the local calcium dynamics, we simplified the dynamics further and chose to focus on amplitude-modulated (AM) astrocyte responses to stimulation (De Pittà et al., 2008). Thus, larger $IP_3$ concentrations translate into larger calcium concentrations and not larger oscillation frequencies (De Pittà et al., 2008). To account for the expected slow time scale of the calcium-release machinery (up to seconds), we made the local calcium dynamics $[Ca^{2+}]_{ija}(t_k)$ converge to $[IP_3]_{ija}(t_k)$ with time scale $\Omega_{acc}$:

$$\begin{aligned}\mathbf{[Ca^{2+}]_{ija}}(\mathbf{t_k}) &= \mathbf{[Ca^{2+}]_{ija}}(\mathbf{t_{k-1}}) + \Omega_{acc} \cdot (\mathbf{[IP_3]_{ija}}(\mathbf{t_k}) \\ &- \mathbf{[Ca^{2+}]_{ija}}(\mathbf{t_{k-1}})). \quad (10)\end{aligned}$$

Gliotransmission occurs when the local calcium concentration exceeds the threshold $[Ca^{2+}]_{th}$:

$$\mathbf{g_{ij}}(\mathbf{t_k}) = \begin{cases} \left(\mathbf{g_{ij}}(\mathbf{t_{k-1}}) + \left(\mathbf{1} - \mathbf{g_{ij}}(\mathbf{t_{k-1}})\right) \cdot \mathbf{g_r}\right) \cdot \mathbf{e}^{-\Omega_{\mathbf{g}}\Delta\mathbf{t}} & \text{if } \mathbf{[Ca^{2+}]_{ija}}(\mathbf{t_{k-1}}) < \mathbf{[Ca^{2+}]_{th}} < \mathbf{[Ca^{2+}]_{ija}}(\mathbf{t_k}) \\ \mathbf{g_{ij}}(\mathbf{t_{k-1}}) \cdot \mathbf{e}^{-\Omega_{\mathbf{g}}\Delta\mathbf{t}} & \text{otherwise} \end{cases}, \quad (11)$$

where the condition for $[Ca^{2+}]_{ija}$ ensures the absence of a new gliotransmission event when calcium drops back below the threshold. In this equation, $g_{ij}(t_k)$ is the fraction of bound presynaptic gliotransmitter receptors, $g_r$ the fraction of unbound receptors recruited, and $\Omega_g$ the recovery rate of gliotransmitter receptors. For simplicity, and unlike in De Pittà et al. (2008), we consider a constant gliotransmission recruiting fraction.

### *Astrocytic network dynamics*

To model astrocyte-astrocyte calcium signaling, we used the UAR model introduced by Lallouette et al. (2014, 2019). In the network model, each astrocyte is a node, and gap junctions are links between the nodes. In the UAR model, an astrocyte "*a*" can have three possible states $S_a$: active state (A), inactive dormant state (U), and refractory (R), during which the cell cannot transmit calcium signals. At any time, the cell will be in one of these states. Transitions between states are probabilistic and depend on the propagation efficiency of coupled astrocytes. The propagation efficiency of an active astrocyte "*a*" is (Lallouette et al., 2014, 2019):

$$\beta_{\boldsymbol{a}}\,(\mathbf{t_k}) = \begin{cases} \frac{1}{\mathbf{I_a}(t_k)} \text{ if } \mathbf{S_a}\,(\mathbf{t_k}) = \mathbf{A} \\ \mathbf{0} \text{ else} \end{cases}, \tag{12}$$

where $I_a(t_k)$ is the number of astrocytes that are gap junction-coupled to "$a$" and are not in the active state $A$. The activation propensity of "$a$" is then obtained with:

$$\gamma_{\mathbf{a}}\,(\mathbf{t_k}) \;=\; \theta_{\mathbf{a}} \sum\nolimits_{\mathbf{b}\in\mathcal{N}(\mathbf{a})} \beta_{\boldsymbol{b}}\,(\mathbf{t_k}) + \frac{\sum [\mathbf{Ca^{2+}}]_{\boldsymbol{ija}}}{\boldsymbol{N}} \cdot \boldsymbol{M}, \tag{13}$$

where $\mathcal{N}\,(a)$ is the set of astrocytes that are gap-junction-coupled to "$a$" and $\theta_a$ is the astrocyte activation threshold. The sum in the second term of the right-hand side of Equation (13) runs over all local areas *ija* composing astrocyte "*a,*" thus effectively adding up the calcium $[Ca^{2+}]_{ija}$ of each of the astrocyte's regions. These local responses are averaged over the whole astrocyte ($N$ is the number of excitatory connections to astrocyte "$a$") and scaled by a factor $M$ to arrive at their contribution to the activation propensity. If the activation propensity of an astrocyte is larger than the threshold $\theta_a$, this astrocyte can activate. Following Lallouette et al. (2014), this threshold changes with the number of astrocyte neighbors $n_a$ as:

$$\theta_{\boldsymbol{a}}(\mathbf{n}_{\boldsymbol{a}}) \;=\; \mathbf{b_0 n_a + b_1}, \tag{14}$$

where $b_0$ denotes the slope of the activation threshold and $b_1$ as the intercept of the activation threshold. The probability for astrocyte "$a$" to become active (U → A) at time step $t_k$ is finally calculated as:

$$\mathbf{P(U\rightarrow A)}_{\boldsymbol{a}}\,(\mathbf{t_k}) \;=\; \begin{cases} \frac{\Delta \mathbf{t}}{\tau_A} \textit{if } \gamma_{\boldsymbol{a}}\,(\mathbf{t_k}) > \theta_{\boldsymbol{a}}\,(\mathbf{n_a}) \\ \mathbf{0 \text{ else}} \end{cases}, \tag{15}$$

where $\tau_A$ is a parameter that sets the time scale of the activation transition. Moreover, the activation of astrocyte "$a$" is signaled back to all its local areas by the following additional rule: The $IP_3$ concentration $[IP_3]_{ija}$ of every local area *ij* composing "$a$" is forced to its maximum value ($[IP_3]_{ija} = 1$) for the entire duration of the active state of "*a.*" Note that, as described by Equation (3), activated astrocytes also release adenosine during the entire duration of the active state.

Finally, transitions from the active to refractory (A → R) and from the refractory to inactive state (R → U) happen spontaneously:

$$\mathbf{P(A\rightarrow R)} \;=\; \boldsymbol{\Delta t}/\tau_{\boldsymbol{R}}, \tag{16}$$

$$\mathbf{P(R\rightarrow U)} \;=\; \boldsymbol{\Delta t}/\tau_{\boldsymbol{U}}. \tag{17}$$

### Neuron and Astrocyte Network Spatial Topologies

Astrocytes were randomly placed on a virtual 2D MEA culture surface area of 750 × 750 $\mu m^2$ (with uniform distribution). If the distance between two astrocyte somas was smaller than 30 μm, one of the two astrocytes was randomly relocated until all inter-soma distances were larger than 30 μm. Each astrocyte was connected by gap junctions to every neighboring astrocyte whose inter-soma distance was smaller than 100 μm. Hence, the diameter of one astrocyte is ∼100 μm in our model (**Figure 2A**).

The spatial distribution of the neurons on the virtual MEA was chosen the same way as for astrocytes. However, the method for connecting the neurons differed. Since neurons form long distance connections, we used a connection probability set by a scaled Gaussian distribution:

$$\boldsymbol{P_{NN}}\,(\mathbf{d}) \;=\; \boldsymbol{e}^{-\frac{\mathbf{d^2}}{\mathbf{2}\sigma_{\boldsymbol{N}}{}^{\mathbf{2}}}}, \tag{18}$$

where $d$ is the (inter-soma) distance between two neurons. Each synapse was connected to the nearest astrocyte in a similar probabilistic way, except that a synapse cannot connect to an astrocyte that is farther than a certain cut-off:

$$\boldsymbol{P_{AN}}\,(\boldsymbol{d}) \;=\; \boldsymbol{e}^{-\frac{\boldsymbol{d^2}}{\mathbf{2}\sigma_{\boldsymbol{A}}{}^{\mathbf{2}}}} \cdot \mathbf{H}\,(\boldsymbol{d_A} - \mathbf{d}), \tag{19}$$

where $d$ is the distance between the cell body of the nearest astrocyte and the synapse. $H()$ denotes the Heaviside function ($H(x) = 1$ if $x > 0$, otherwise $H(x) = 0$) and $d_A$ is the cutoff distance, which we set to 70 μm (**Figure 2A**). If the synapse does not connect to the nearest astrocyte, the next-nearest astrocyte is tried and so forth. Note that, in our model, an excitatory synapse can end up without an astrocyte.

## Numerical and Analysis Methods

### Spike and Burst Detection

In this paper, we analyzed neuronal activity in the form of spikes and bursts which are cascades of spikes. Synchronous population bursts are characteristics of matured and well-connected networks (Giugliano et al., 2004; Wagenaar et al., 2006; Lenk et al., 2016). Spike and burst features were calculated using a modified version of the cumulative moving average (CMA) algorithm (Kapucu et al., 2012; Välkki et al., 2017). The threshold used to decide whether a spike belongs to a burst was set by the skewness of the cumulative moving average of the interspike interval distribution. Using the CMA algorithm, we calculated the spike rate in spikes/minute, the burst rate in bursts/minute, the average burst duration in milliseconds, and the average spikes per burst at the post-synapse. **Figure 3** depicts an example spike train from our simulations with detected bursts. For each spike/burst feature and noise level, we performed a one-way ANOVA (GraphPad Prism v8.2.1, GraphPad Software Inc., California, USA) to confirm that the features were statistically different for each model scenario.

### Frequency and Activity Analysis

We constructed multiple parameter sets describing different neuron or neuron-astrocyte networks. The total spike count of the neuronal network was calculated for each run. The resulting signal was then centered by subtracting its mean, and a discrete Fourier transform (DFT) was applied. We only considered the modulus of the Fourier transform coefficients. For each simulation, we applied the DFT to each of the five conducted runs (see section Simulations) and calculated the corresponding

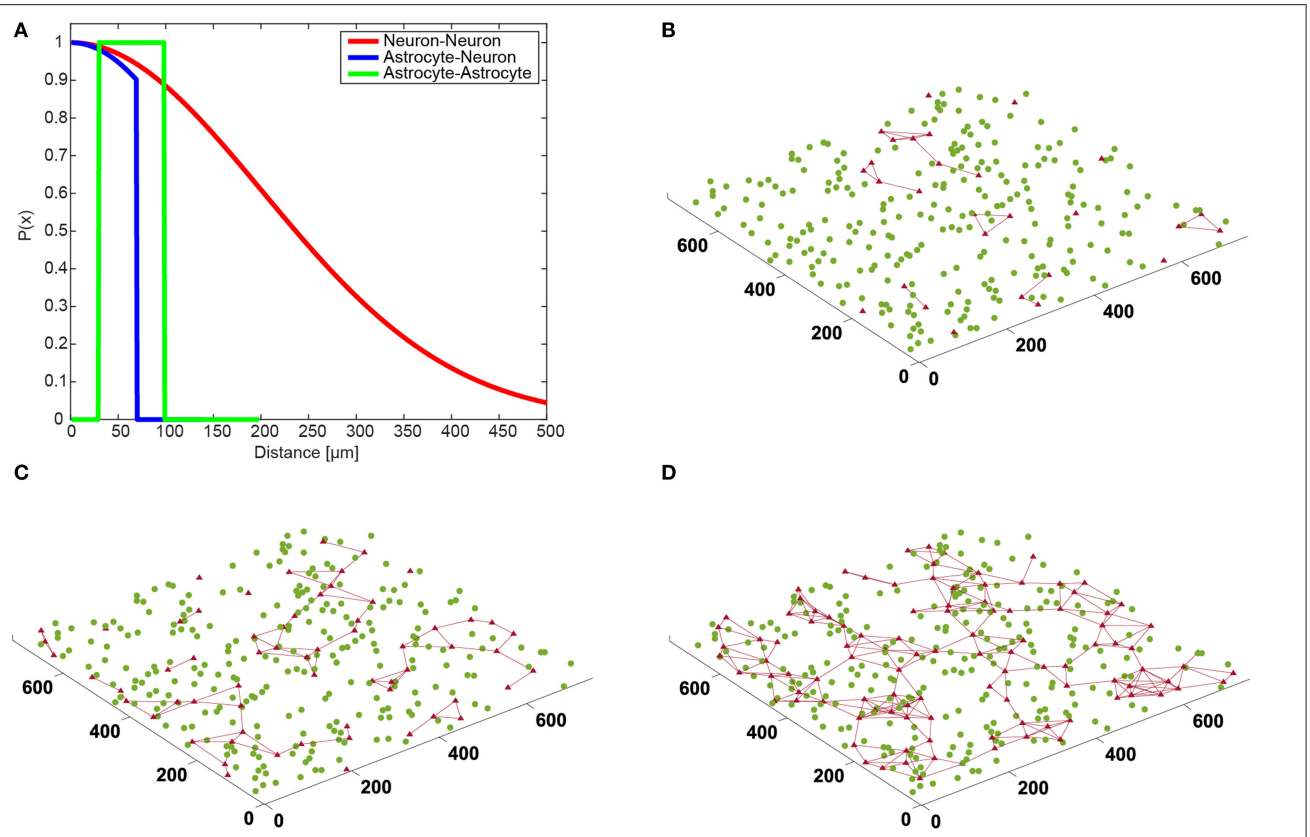


**FIGURE 2 |** Connection distances between cells and spatial neural network topology. **(A)** Neuron-neuron connections (red) are Gaussian based on distance. Astrocyte-neuron connection probability (blue) follows Gaussian until it reaches a limiter. Astrocyte-astrocyte connections (green) form as long as the two cells are closer than a set limiter. **(B–D)** The graphics show the neuronal network with **(B)** 10%, **(C)** 20%, and **(D)** 30% astrocytes on a "virtual" multielectrode array (units are in μm). The neuronal network is represented by the position of the neurons (green circles), and the astrocytic network includes the cells (red triangles) and the connections between them (red lines).

average frequency spectra. The average frequency spectrum was then smoothed by convolution with a Gaussian kernel:

$$\zeta_s(\mathbf{f}) = \int_{-\infty}^{+\infty} \frac{\zeta(\mathbf{x})\, e^{-\frac{(\mathbf{f}-\mathbf{x})^2}{2\sigma^2}}}{\int_{-\infty}^{+\infty} \boldsymbol{H(y)} e^{-\frac{(\mathbf{x}-\mathbf{y})^2}{2\sigma^2}}\, dy} dx \qquad (20)$$

with $\zeta(f)$ the DFT coefficients and $H(y) = 1$ if $y$ is between the minimum and maximum frequencies obtained from the DFTs, and 0 otherwise. This allows a correction of border effects. For all frequency spectra shown in this paper, we used $\sigma = 0.025$ Hz.

Cross-correlation between neuronal and astrocytic activities was computed by smoothing the neuronal (respectively, astrocytic) activities by

$$\mathbf{L_s}(\mathbf{t}) = \int_{-\infty}^{+\infty} \frac{\mathbf{L}(\tau)\, e^{-\frac{(\mathbf{t}-\tau)^2}{2\rho^2}}}{\int_{-\infty}^{+\infty} \mathbf{F(y)} e^{-\frac{(\tau-\mathbf{y})^2}{2\rho^2}}\, dy} \mathbf{d}\tau, \qquad (21)$$

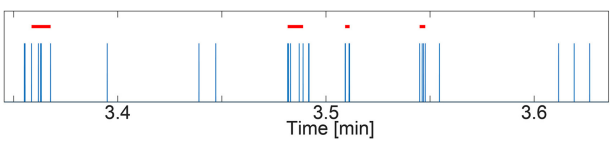


**FIGURE 3 |** Example spike train (spikes in blue) with detected bursts (red bars) using the cumulative moving average (CMA) algorithm. The simulated spike train stems from a data set with NN+A(30%) and Noise = 0.01. The y-axis shows time in minutes.

with $L$ the original pooled neuronal or astrocytic activity signal, and $L_s$ the smoothed signal. $F(y)$ is equal to 1 if $y$ is between 0 and the maximum time of simulation (usually 300 s), and 0 otherwise. We used $\rho = 3$ s. For each run, we computed the cross-correlation using the `crosscorr` function in Matlab (version R2017b, MathWorks, USA). The cross-correlation was then averaged across the five runs for each relevant scenario.

Average astrocyte activation ratios were computed for simulations in which astrocytic networks were used. As for neuronal activity, the astrocyte activity was pooled in 5 ms bins; at each time step, the total number of currently active astrocytes

in the simulation was recorded. The average astrocyte activation ratios $AR$ were then computed by:

$$\mathbf{AR} = \frac{\langle \mathbf{B} \rangle}{\mathbf{n_A}} \frac{\tau_{R\mathbf{t}} + \tau_U + \tau_A}{\tau_R} \quad (22)$$

with $\langle B \rangle$ the average number of astrocytes activated at any given time and $n_A$ the total number of astrocytes. $\frac{\langle B \rangle}{n_A}$ was thus the average fraction of astrocytes that were activated at any given time. The average transition times between astrocyte states were used to scale the activity such that a value of 1 corresponded to the highest average activity possible (when astrocytes continuously changed from inactivated ($\tau_U$), to activated ($\tau_A$) to refractory ($\tau_R$) states). When applicable, Spearman's rank correlation coefficients and associated $p$-values were computed using the `corr` function in Matlab.

The homeostatic effects of astrocytes can further be investigated by looking at how the average neuronal spike rate changes when astrocytes activate faster (represented by parameter $\tau_A$), or when the strength of their presynaptic effect is changed (represented by $\Omega_g$). Low values for $\tau_A$ lead to high activation while high values prevent activation [see Equation (15)]. On the other hand, parameter $\Omega_g$ controls the presynaptic effect of astrocyte processes: high values lead to fast recovery of glutamate receptor (and thus low presynaptic effects) while low values lead to slow recovery (and thus high presynaptic potentiation). Therefore, we ran NN+A(30%) simulations with noise $c_i = 0.02$ and varied $\tau_A$ between 1.0 and 4.5 s and $\Omega_g$ between 0.077 and 51.29 $s^{-1}$.

### Simulations

To illustrate how the INEXA network model and what the astrocyte contribution to its dynamics is, astrocytic signaling was progressively added, starting from the original INEX model in four sequential stages:

- **Noise only:** we only included the neuronal background noises $c_i$ (Equation 7), i.e., all synaptic weights and the astrocytic depressing terms were set to zero ($y_{ij} = y_{\text{Astro}} = 0$ in Equation 7). This scenario therefore is to be considered as a reference where the neurons are connected neither to each other nor to the astrocytes.
- **NN only:** we set the synaptic weights to constant values (i.e., $-0.7 \leq y_{ij} \leq 0.7$), keeping $y_{\text{Astro}} = 0$. This stage thus corresponds to a pure neuronal network response with no influence of the astrocytes on the neurons.
- **NN + PSA:** each excitatory presynapse was connected to an astrocyte (PSA). In this scenario, however, the astrocytes themselves did not form a network (i.e., the term $\beta_a$ of Equation 12 was set to zero for all astrocytes at all times) and no adenosine was released into the extracellular space (i.e., we keep $y_{\text{Astro}} = 0$ in Equation 7).
- **NN+A(x%):** the complete INEXA model was tested and compared to the second and third phase (i.e., $\beta_a$ was computed according to Equation 12 and $y_{\text{Astro}}$ was set to the value found in **Table 1**). Furthermore, to test the effect of the number of astrocytes on the network activity, we simulated cultures composed of roughly 10% [called "NN+A(10%)"], 20% ["NN+A(20%)"], and 30% ["NN+A(30%)"] astrocytes.

**TABLE 2 |** Statistics of the neuronal network.

| Measure | Value |
|---|---|
| Maximum amount of neuronal network connections | 62,250 |
| Average number of connections to other neurons | 72.12 |
| Network connectivity in % | 28.96 |
| Average length of connections in micrometer | 211.57 |
| Number of bidirectional connections | 5,284 |

In all simulations, the network consisted of 250 neurons, of which 200 were excitatory (80%) and 50 inhibitory (20%). Each of the above described simulation phases was run five times with three different noise levels (the upper boundaries of $c_i$ were set to $C_{max}$ = 0.01, 0.02, or 0.03). The same neuronal network was used in all simulations. However, if present, the astrocytic network was resampled at each run. In total, these four phases produced 18 scenarios. A total simulated time of 5 min was chosen. The values of the parameters used in the simulations are given in **Table 1**.

### Topology

**Table 2** summarizes the statistics of the simulated neuronal and astrocyte networks. The connectivity within the neuronal network was 29%. Each astrocyte was to connected to between 130 and 250 excitatory synapses depending on the ratio of astrocytes in the network ["NN+A(10%)," "NN+A(20%)," and "NN+A(30%)," more astrocytes yielding less synapses per astrocyte, see **Table 3**]. Likewise, each astrocyte was connected to one to five neighboring astrocytes through gap junctions depending on the astrocyte ratio (more astrocytes yielding more gap junction couplings per astrocyte).

**Figures 2B–D** shows the spatial topology of neurons and the astrocytic network resulting from the spatial rules described in section Neuron and Astrocyte Network Spatial Topologies. In the case of "NN+A(10%)" (**Figure 2B**), only a few astrocytes formed connections, and half of the excitatory synapses (51.1%) were not controlled by an astrocyte. In "NN+A(20%)" (**Figure 2C**), almost all astrocytes were connected to at least one neighboring astrocyte. However, the number of astrocytes used was not enough to reach all synapses, and 15.2% of the excitatory synapses were left without any astrocyte. Finally in "NN+A(30%)" (**Figure 2D**), a widely interconnected astrocytic network spread all over the entire neuronal network, and only 3.8% of the excitatory synapses were not connected to an astrocyte.

# RESULTS

## Single Synapse-Astrocyte Interaction

We first use simulation results to illustrate how communication between neurons and astrocytes shapes the dynamics of our INEXA model. **Figure 4** shows three time series from a simulation with 30% astrocytes ["NN+A(30%)" scenario]. The release of resources (**Figure 4B**) was induced by the activity

**TABLE 3 |** Statistics of the astrocytic network: mean value and standard deviation over the five runs for NN+A(10%), NN+A(20%), and NN+A(30%), respectively.

| Measure | NN+A(10%) | NN+A(20%) | NN+A(30%) |
|---|---|---|---|
| Connections of an astrocyte to nearby excitatory synapses | 252.05 ± 13.16 | 194.22 ± 6.15 | 129.68 ± 1.88 |
| Gap junction connections between astrocytes | 1.42 ± 0.56 | 2.55 ± 0.27 | 4.86 ±0.31 |
| Lowest and highest gap junction amount (rounded) | 0 ± 0–4 ± 1 | 0 ± 0–5 ± 1 | 0 ± 1–9 ± 1 |
| Distance between connected astrocytes in $\mu$m | 68.65 ± 4.78 | 70.92 ± 1.35 | 70.14 ± 0.87 |
| Number of excitatory synapses without an astrocyte (rounded) | 7363 ± 368 | 2185 ± 387 | 544 ± 201 |
| Percent of "naked" (without astrocyte) excitatory synapses | 51.06 ± 2.55 | 15.15 ± 2.68 | 3.77 ± 1.40 |

of the presynaptic terminal (**Figure 4A**), but the amount of neurotransmitters released into the synaptic cleft varied, depending on the fraction of available vesicles (Equation 3) and the fraction of these vesicles that were ready for release (Equation 4). The amount of neurotransmitter in the cleft was directly linked to the post-synaptic activity as described by Equations (5) and (7). Accordingly, more frequent post-synaptic spikes were elicited when larger amounts of neurotransmitters were released (compare **Figures 4B,F**).

In our model, spike-induced neurotransmitter release had an impact not only on the neuronal network, but also on the astrocytic network. The astrocytes were able to detect synaptic activity through the resources released by the presynaptic terminal in the synaptic cleft. Hence, in response to presynaptic activity, the local astrocyte $IP_3$ level increased, which led to the release of calcium from the astrocytic ER (**Figures 4C,D**). When the astrocyte local calcium concentration exceeded a threshold (the red line in **Figure 4D**), gliotransmission took place (as indicated by the black diamonds) and a sudden increase in the gliotransmitter concentration was detected (**Figure 4E**). Gliotransmission signaled back to the synapse, affecting the internal dynamics of the presynaptic terminal: the amount of resources released into the synaptic cleft was therefore higher on average when the gliotransmitter concentration was large (compare **Figures 4B,E**). Therefore, gliotransmission was release-increasing or potentiating for this particular synapse (see Glial Components). Upon activation of the whole astrocyte, both $IP_3$ and calcium levels switched to a high state (**Figures 4C,D**; the local $IP_3$ level is set to 1 upon astrocyte activation). Once activated, the astrocyte released adenosine into the extracellular space, reducing the activity of the post-synaptic neuron, which progressively decreases the spike rate (**Figure 4F**). In addition, the presynaptic neuron was also indirectly affected by astrocyte activation. The level of local calcium was maintained above the release threshold while the astrocyte was active, which prevented new releases of gliotransmitter. Thus, temporarily canceling the potentiating effect of gliotransmission on the presynaptic terminal [see Equations (5–7)].

As described in the Methods section, the dynamics of astrocyte activation is governed by two variables in our model: the local $Ca^{2+}$ activity from the enwrapped synapses and the contribution to this activity by intercellular $Ca^{2+}$ wave propagation (Equations 12–17). **Figure 5** shows the excitation dynamics of the astrocyte connected to the synapse shown in **Figure 4**. **Figures 5A,B** demonstrate how the global calcium signal generally increased upon periods of high presynaptic activity. However, the global calcium signal could reach high values even when the presynaptic activity in this particular neuron was weak. This is due to calcium release triggered by other synapses to which the astrocyte was connected. Moreover, the activation propensity of the astrocyte (**Figure 5C**) depended on the number of its neighboring astrocytes [see Equations (12–13)]. Most of the time, both signals were needed to activate the astrocyte. That means, to activate the astrocyte usually demanded that both the amount of global calcium becomes larger than its threshold and that the activation propensity of the coupled astrocytes crosses over its own threshold. This is for example the case slightly after $t = 20$ in **Figure 5**, where activation occured when both the calcium trace (panel B) and the propensity trace (panel D) overcame their respective thresholds (red lines). However, having both signals crossing over their thresholds was not mandatory to activate the astrocyte, since astrocyte activation could also be triggered by only one of them. For instance, the activation occurring around $t = 55$ in **Figure 5** was triggered when the global astrocyte $Ca^{2+}$ crossed over its threshold, at a time step where the propensity trace was still well below its own threshold.

**Figure 5D** shows the astrocyte state [inactive (U), active (A), or refractory (R)] along the simulation time. When the astrocyte became activated, the global calcium signal switched to a high state. Those active periods also corresponded to the high state periods observed in the local $IP_3$ and calcium signals in **Figure 4**. The post-synaptic activity was clearly reduced as a consequence of the depression exerted during astrocyte active periods regardless of the activity at the synapse (**Figure 5E**).

## Spike and Burst Detection

To understand how the local dynamics of the tripartite synapses in the models impacted the dynamics of the whole network, we next quantified the bursting behavior of the neuronal network for each simulation scenario (see section simulations above), especially when presynaptic astrocyte signaling and the formation of astrocytic networks were added to the model. **Figure 6** shows the burst and the spike rates as well as the number of spikes per burst and the burst duration in each of the studied simulation scenarios (except for the "noise only" scenario that, as expected, exhibited no remarkable bursting).

When the neuronal network was formed via synaptic connections that did not depend on astrocyte activity ("NN only," the blue bars in **Figure 6**), the spike rate increased with the noise level, since the noise level determined basal firing activity. Those spikes proportionally contributed to the burst

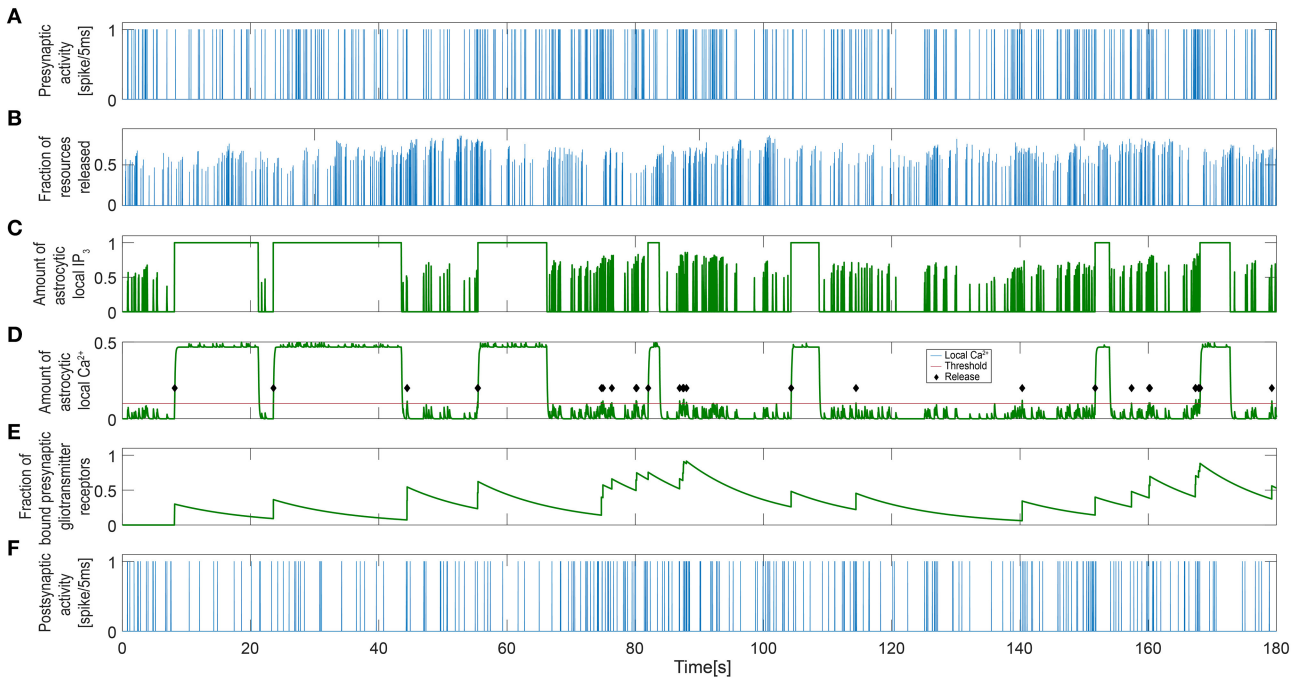


**FIGURE 4 |** Signaling process governing single tripartite synapse activity. **(A)** Single presynaptic neuronal activity. **(B)** Resources released to the synaptic cleft. The correlation between the release process and the presynaptic activity indicates that this is a spike-induced process. **(C)** Astrocytic local $IP_3$ and **(D)** calcium concentration. The levels of local calcium follow those of $IP_3$ with a small delay. The different timescales for the neuronal and astrocytic networks are detectable (Equation 9). The red line represents the threshold level for the gliotransmitter release and the green diamonds indicate that gliotransmission has occurred. **(E)** Gliotransmitter glutamate released from the astrocyte controlled by local calcium dynamics. **(F)** Single post-synaptic neuronal activity.

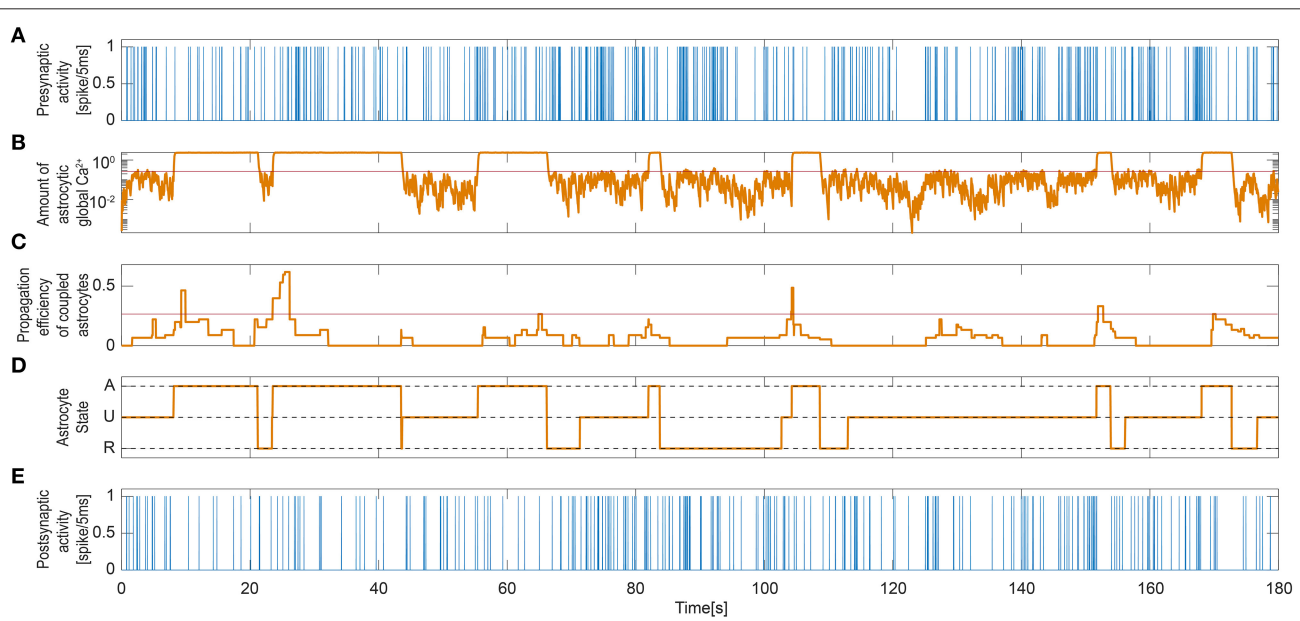


**FIGURE 5 |** Signaling process governing single astrocyte activity. **(A)** Single presynaptic neuronal activity. **(B)** Astrocytic global calcium dynamics corresponding to the averaged and scaled local responses from all enwrapped synapses. The red line indicates the threshold set for the activation of the astrocyte. **(C)** $IP_3$ influx that the current astrocyte receives from all its active neighbors. Again, the red line is the threshold for the activation of the astrocyte. **(D)** State signal of the UAR model astrocyte: inactive dormant state (U), active signaling state (A), and refractory period (R). **(E)** Single Post-synaptic neuronal activity.

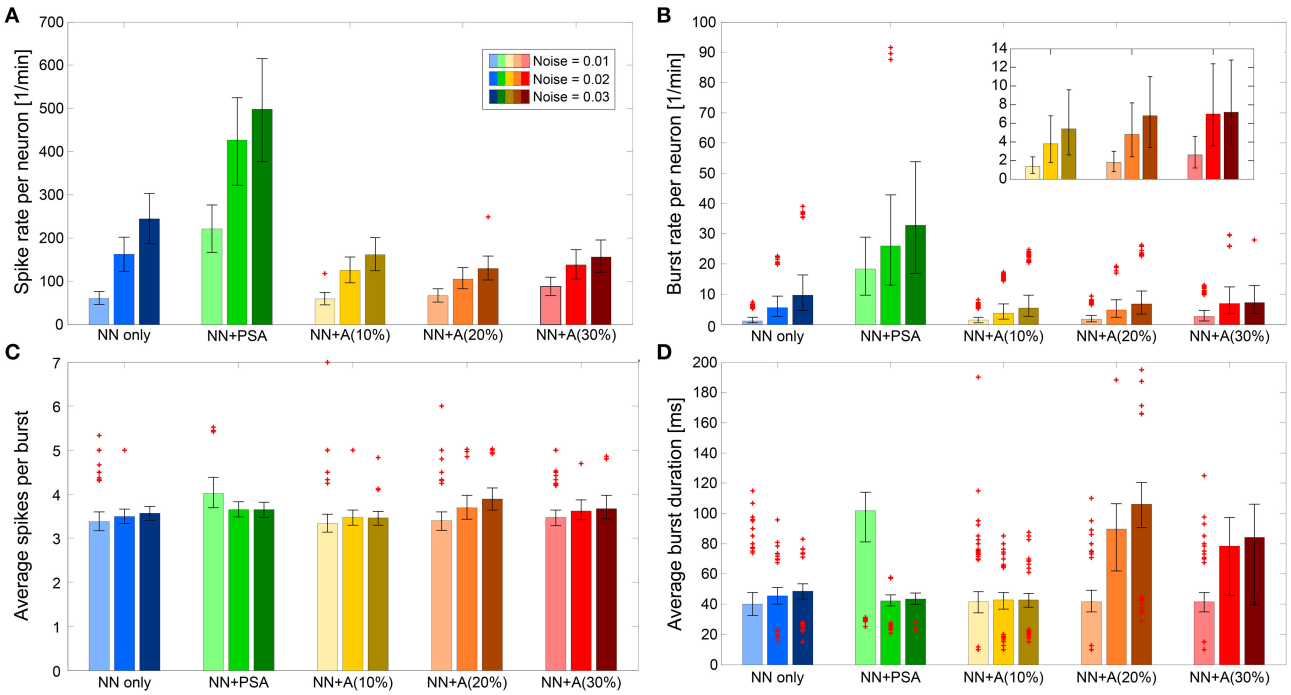


**FIGURE 6 |** Spiking and bursting behavior features. The features are displayed for the three noise levels in cases of "NN only" (blue), "NN + PSA" (green), "NN + A(x%)" (yellow, orange, red) averaged across five runs, respectively. Error bars plot the 25th and 75th percentiles. Individual plotted points represent extreme data points not considered within the percentiles. **(A)** Average spike rate per neuron per minute. **(B)** Burst rate per neuron per minute. **(C)** Average spikes per burst. **(D)** Average burst duration in milliseconds.

development as indicated also by the higher burst rate. However, **Figures 6C,D** shows that the characteristics of the bursts (number of spikes per burst, burst duration) were not affected by noise level.

In the "NN+PSA" case, where the astrocytes were connected to the presynaptic terminals of the neuronal network but not to each other (the green bars in **Figure 6**), the network as a whole became more active as a result of the potentiating effect of the astrocytes on the excitatory synapses. As one might expect, the spike rate increased with the noise level/basal rate (**Figure 6A**). Moreover, the burst duration decreased since the number of spikes per burst was constant, but the burst rate increased. These changes were the consequences of the gliotransmitters released from the astrocytes. On average, gliotransmission increased the presynaptic release probability [see Equation (7)], which led to a larger amount of resources released into the synaptic cleft [see Equation (5)], and thus a larger firing rate of the post-synaptic neuron compared to the "NN only" scenario.

The addition of the astrocytic network to the model strongly changed the bursting behavior of the neuronal network. In those "NN+A(x%)" scenarios, we both introduced astrocyte to astrocyte coupling via gap junction, but also the depressing impact of astrocytes on the post-synaptic firing rate. The immediate effect of the addition of the astrocytic network was that both the spike rate and the burst rate were much lower than those obtained in the "NN+PSA" case (**Figures 6A,B**) while the mean number of spikes per burst was not altered (**Figure 6C**). Interestingly, the spike rate was almost constant regardless of the number of astrocytes [compare the different "NN+A(x%)" scenarios] because of the trade-off between the effect of glutamate transmission and adenosine depression. However, as can be seen in the inset of **Figure 6B**, the burst rate slightly increased with the number of astrocytes, which suggested that one of the consequences of the astrocytic network might be the introduction of bursting behavior.

Analyzing the effects on burst duration was more complex. In the case of "NN+A(10%)" (the yellow bars in **Figure 6**), the average burst duration did not significantly change with the introduced noise levels. However, the high number of outliers for the average burst duration revealed the existence of two types of behaviors within the neural network for intermediate-to-high noise levels (**Figure 6D**). This might result from an astrocytic network that was too sparse to compensate for the high activity of the neural network with high noise. Indeed in "NN+A(20%)" and "NN+A(30%)," the burst duration increased with increasing noise and with respect to "NN+A(10%)." These results support our above interpretation: as the number of astrocytes increased, the astrocytic network was also strengthened. Thus, it was able to control the whole neuronal network by preventing it from overexcitation, even at high noise levels.

One-way ANOVA confirmed that the spike and burst features were significantly different for each model scenario ($p < 0.0001$). We performed the test for each feature and noise level separately. Taken together, **Figure 6** shows that the astrocyte network

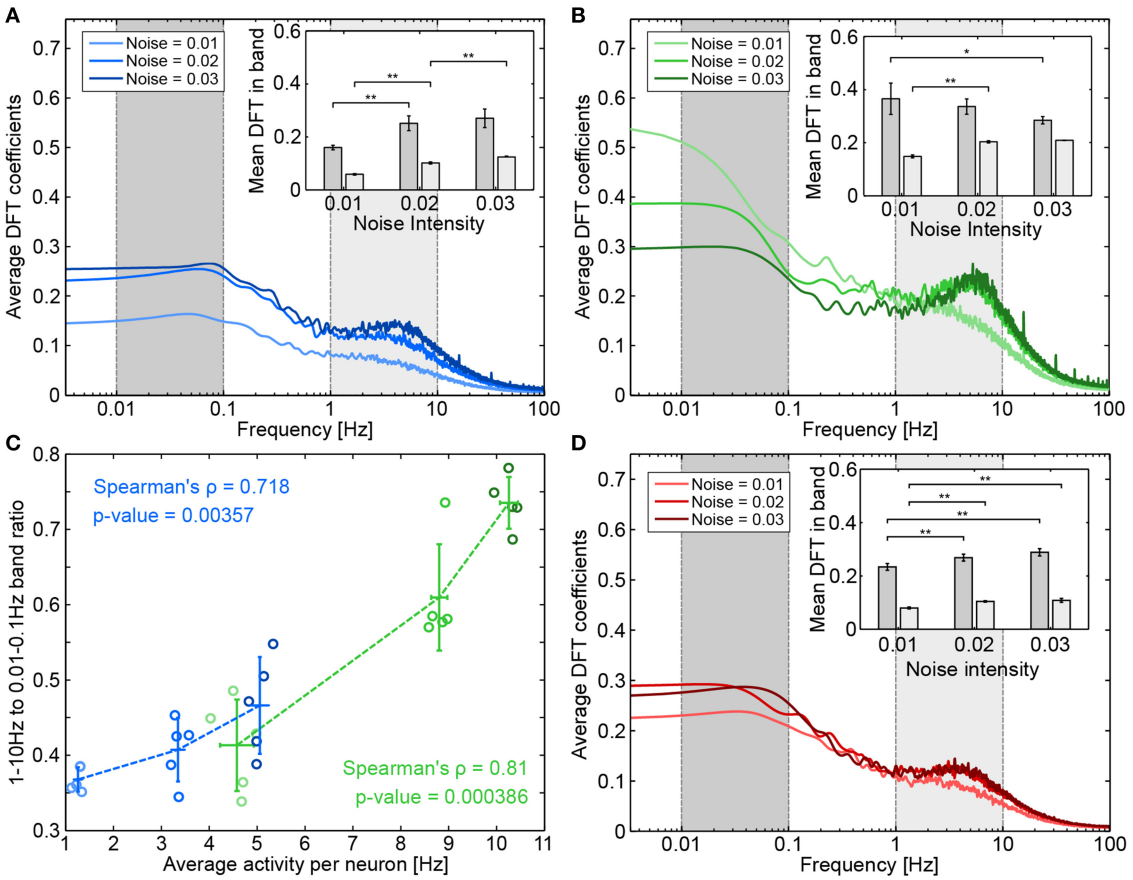


**FIGURE 7 |** Effect of presynapse-astrocyte processes on neuronal activity. Raw frequency spectrums for **(A)** "NN-only" **(B)** "NN + PSA" and **(D)** "NN+A(30%)" were averaged across five runs and smoothed as described in the Methods section. The inset shows the average DFT coefficients for different noise intensities and for two frequency bands: 0.01–0.1 Hz (dark gray) and 1–10 Hz (light gray). Error bars plot the standard deviation of band averages across runs. Significance was assessed by double-sided Mann-Whitney tests (comparing distributions of band averages). $^{*}p < 0.05$, $^{**}p < 0.01$. **(C)** Relationship between average activity per neuron and the ratio between band averages. Each circle represents a run, a darker circle denotes a higher noise intensity; blue data corresponds to "NN only" and green data corresponds to "NN + PSA".

downregulated the activity of the neural network by decreasing its burst and spike rates while increasing burst duration.

## Activity and Frequency Analysis

To further analyze how the addition of presynaptic astrocyte signaling and full astrocytic networks affected neuronal activity, we next quantified the changes in the overall activity levels and in specific frequency bands of the neuronal network activity. Therefore, we applied discrete Fourier transforms (DFT) on the pooled neuronal activity signals (details in the Methods section).

### Effect of Presynapse-Astrocyte Processes

The "NN only" scenario is a natural comparison point for understanding the effect of astrocytes on neuronal activity. **Figure 7A** shows the frequency spectra corresponding to the "NN only" scenario for different levels of noise. The frequency spectra display a slight increase for two frequency decades: very low frequencies, between 0.01 and 0.1 Hz (the dark gray band); and medium frequencies between 1 and 10 Hz (the light gray band). As noise intensity increased (the light to dark blue curves), the amplitude of both frequency bands increased. However, as can be seen in the inset of **Figure 7A**, in which both frequency bands were averaged, the gap between them seemed to decrease as the noise intensity increased.

When presynaptic astrocytes were added ("NN+PSA"), the average intensity of both bands strongly increased (see the green bars in **Figure 6A**). Gliotransmitter release from the astrocyte increased the value of the basal release probability $U_{ij}^{*}$ of TM synapses (De Pittà et al., 2011), which thus increased the amount of released resources. The corresponding frequency spectrums can be seen on **Figure 7B**. While the power in the 1–10 Hz band seemed to increase with noise intensity, the power in the 0.01–0.1 Hz band actually decreased. The increase of the average neuronal activity evidenced by **Figure 6** is thus not uniformly distributed across frequencies.

Since noise intensity was linked to increased average activity, we checked whether the changes in medium and low frequency bands could be linked to average activity in both the "NN

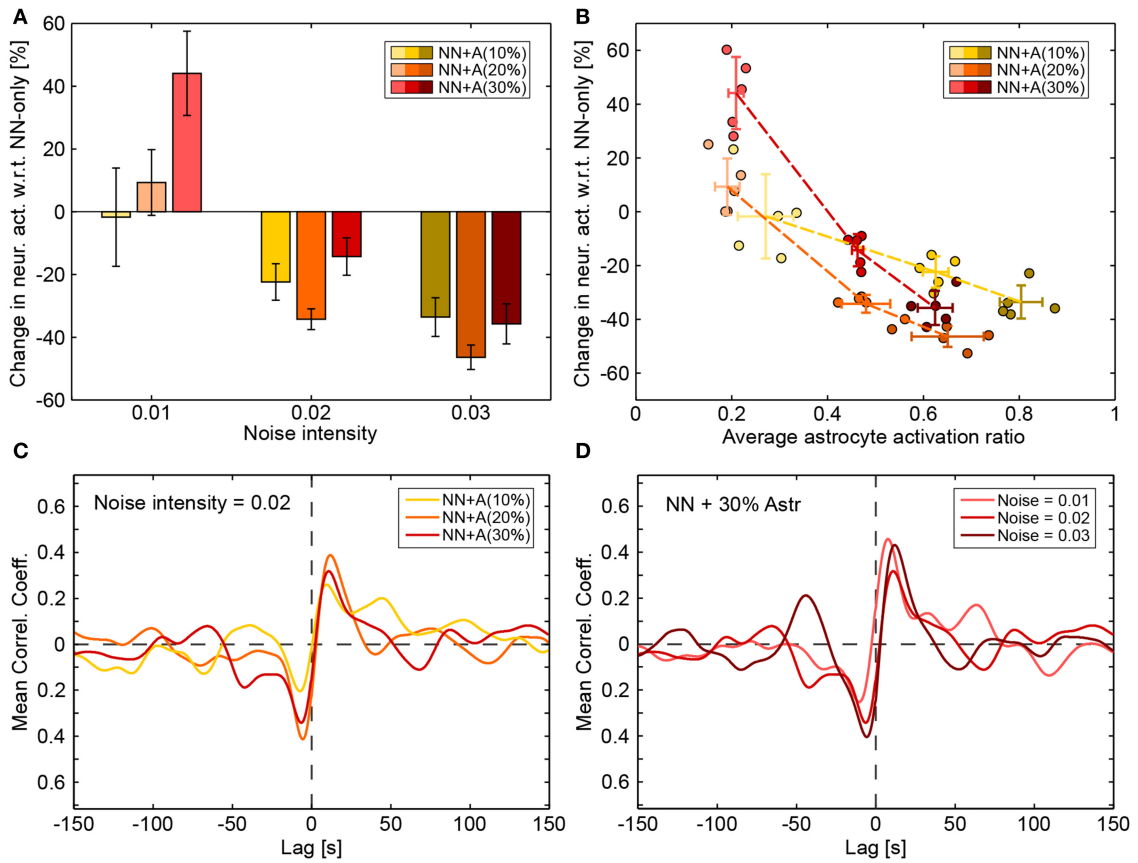


**FIGURE 8 |** Effect of astrocytic networks. **(A)** Changes in neuronal activity introduced by the addition of astrocytes (compared with "NN only" simulations). Values were averaged across runs and error bars plot the standard deviation across runs. **(B)** Changes in neuronal activity introduced by the addition of astrocytes (compared with "NN only" simulations) as a function of the average astrocyte activation ratio (a value of 1 denotes the highest possible activity in the astrocytic network). Each circle represents a run. Darker circles denote higher noise intensity and the hue (yellow to red) denotes the amount of astrocytes in the simulation (10–30%). Crossed error bars indicate averages and the standard deviation across the runs. **(C)** Average cross correlations between neuronal and astrocytic smoothed activities for a constant noise intensity of 0.02 and for "NN+A(10%)" (yellow), "NN+A(20%)" (orange), and "NN+A(30%)" (red). **(D)** Cross correlation between neuronal and astrocytic smoothed activities for varying noise intensity (light to dark red) in the "NN+A(30%)" scenario. Cross correlation values were computed as described in the Methods section.

only" and "NN+PSA" scenarios. We thus examined how the ratio between the 1–10 Hz and the 0.01–0.1 Hz bands changed as a function of average activity. **Figure 7C** shows these values for both "NN only" (blue) and presynaptic astrocyte signaling ("NN+PSA," green) scenarios. In both cases, increases in average activity were significantly correlated with increased band amplitude ratios, meaning that increased spiking activity mostly influenced the higher medium frequencies as opposed to low frequencies. This agreed with the spike and bursts analysis since in the "NN only" and "NN+PSA" scenarios, the increase in the burst rate per neuron with the noise seen in **Figure 6** could be associated with the increase in the amplitude of the 1–10 Hz band.

## Effect of Astrocytic Networks

The addition of a full astrocytic network—which could potentially synchronize distant synapses and depress the whole neuronal network through adenosine release—changed how the neuronal network behaved. With respect to "NN only" simulations (the blue bars in **Figure 6A**), the average activity of the neural network (the yellow to dark red bars) was slightly increased by the astrocyte network for low noise intensity (the left-most bars of each group), but it was strongly decreased for high noise intensities.

**Figure 7D** shows the average frequency spectra obtained when 30% of astrocytes were present (corresponding figures for 10 and 20% show similar results). In contrast to the above results, when the noise intensity increased, the frequency spectrums did not change greatly and stayed close to the frequency spectrums of "NN only" simulations (**Figure 8A**). As the average band intensity increased with the noise intensity, as shown in the inset, the strength of both low (dark gray) and medium (light gray) frequency bands slightly increased as well. In contrast to the "NN+PSA" scenario, the 1–10 Hz frequency band did not increase much with increasing noise. **Figure 8A** shows how astrocytic networks affected neuronal activity by displaying the change (in %) between "NN only" and "NN+Astr(x%)" simulations (yellow to red corresponds to 10

to 30% astrocytes) for increasing noise intensities. Increasing the number of astrocytes in the networks had two opposing effects: (1) It introduced more enwrapped synapses, which, as already mentioned, increased the average neuronal activity. (2) It decreased neuronal activity by releasing ATP/adenosine upon astrocyte activation. In case astrocytes were not stimulated enough to be consistently activated, adenosine release was rare and effect (2) was weak compared to (1). With low noise, the low average neuronal activity therefore explains the increase of activity seen in **Figure 8A**, because effect (1) was greater than (2). On the other hand, when the noise increased (noises 0.02 and 0.03), adenosine signaling was more frequently activated, and the overall effect of the astrocyte network was to decrease activity when compared to the "NN only" scenario.

The interplay between astrocyte activation and changes in neuronal activity can clearly be seen in **Figure 8B**: high astrocyte activity clearly correlated with decreased neuronal activity while low astrocyte activity correlated with increased neuronal activity. The higher the number of astrocytes, the steeper this relationship became (yellow to red curves). With enough astrocytes, the interplay between neuronal and astrocytic networks even impacted the cross-correlation between average neuronal activity and average astrocyte activity. **Figure 8C** shows the average cross-correlation between neuronal and astrocytic activities for increasing number of astrocytes (yellow to red) at a constant noise intensity. **Figure 8D** shows the same cross-correlation but only for the "NN+A(30%)" scenario and for increasing noise intensities (light to dark red). In all cases, neuronal and astrocytic activities were negatively correlated with lags around −5 s (global minimum of the mean correlation coefficient) and positively correlated with lags around 10 s (global maximum of the mean correlation coefficient). This means that high astrocyte activity was followed by low neuronal activity ∼5 s later, while high neuronal activity was followed by high astrocyte activity ∼10 s later (which is of the order of the time needed by an astrocyte to activate).

To explore if astrocytes contribute to network firing stability as a homeostatic modulator, we varied the recovery rate of the gliotransmitters, $\Omega_g$, and the average activation time of an astrocyte, $\tau_A$ (**Figure 9**). As expected, increasing $\tau_A$ led to a decreased neuronal activity across the whole range of $\Omega_g$ values. Increasing $\Omega_g$ resulted in a decreased presynaptic potentiation, and thus in a decreased average spike rate. No further changes could be seen for $\Omega_g > 1\ s^{-1}$, since presynaptic glutamate receptors recover very fast and prevent any presynaptic potentiation. The resulting average spike rate thus resulted from a trade-off between local astrocyte processes (whose potentiating effect is controlled by $\Omega_g$) and global astrocyte activations (whose depressing effect is controlled by $\tau_A$).

To summarize, our simulations revealed that astrocytes exerted two opposite effects on neuronal activity. The activation of presynaptic astrocyte processes increased the neuronal activity through the release of potentiating gliotransmitters like glutamate. When neuronal activity became high enough to elicit significant astrocyte activation, depressing gliotransmitters like ATP/adenosine were released, leading to a decrease of the neuronal activity. Overall, these results show that astrocytic networks promoted stabilization of the average neuronal activity, boosting low average neuronal activity through the effect of presynaptic astrocyte processes while reducing high activity levels through adenosine release.

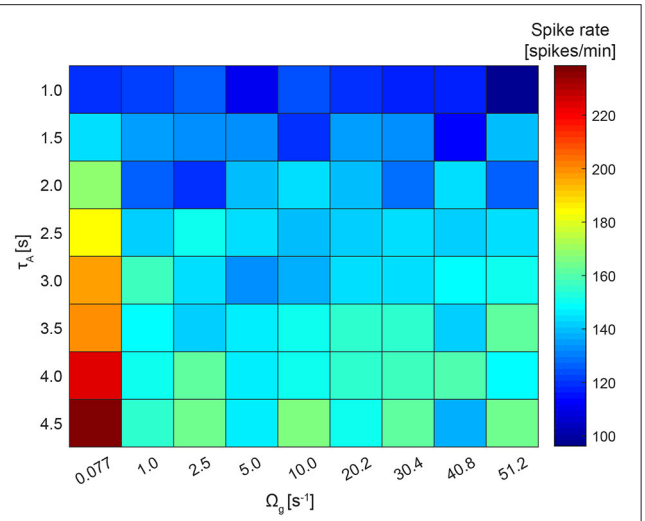


**FIGURE 9 |** Network firing stability. The recovery rate of the gliotransmitters, $\Omega_g$, varies between 0.077 and 51.2 $s^{-1}$ and the average activation time of an astrocyte, $\tau_A$, between 1.0 and 4.5 s. For this simulation, the NN+A(30%) model and $c_i$ was fixed to 0.02 was used. For each run, the average across the resulting spike rates of all 250 neurons was calculated.

## DISCUSSION

We developed an *in silico* description of connected neuronal and astrocytic networks and assessed their interactions combining in a biologically plausible fashion previously introduced models for different parts of those networks (De Pittà et al., 2011; Lenk, 2011; Lallouette et al., 2014). Our goal was to study the role of astrocyte networks when coupled to neuronal networks. To assess the effects of the astrocyte networks on the neuron network, we quantified spike and burst features and used pooled spike trains as indicators of frequency based activity at the network level. The frequency analysis of the pooled spike trains allowed us to identify changes in the signaling patterns of the network.

Astrocytes may play a role on short-term and long-term synaptic plasticity (De Pittà et al., 2016). Short-term plasticity includes the potentiation or depression of neurotransmitter release, which occurs in the milliseconds to minutes range. Astrocytes were also connected to influence long-term potentiation or depression (Turrigiano, 2008; De Pittà et al., 2016). Memory and learning related changes of the global synaptic strengths could be a result of adjustments to an increasing or decreasing firing rate. However, they could also be related to more local homeostatic effects (Turrigiano, 2008).

With our model, we have mainly investigated short-term effects. Comparing the spike and burst features between the pure neuronal network ("NN only") and the neuronal network where each excitatory presynapse was connected to an astrocyte ("NN+PSA"), our simulations show that more noise means

more activity, because of the absence of depression mechanisms stronger than the short-term depression introduced by the Tsodyks-Markram synapses. When astrocytes are introduced to the model, we can observe two types of responses from the network as compared to "NN only." On the one hand, when the average activity is low (noise = 0.01), the astrocytes promote neuronal activity, since the presynaptic effect of the astrocytes prevails over adenosine depression. On the other hand, when the average activity is higher (noise = 0.02 and 0.03), neuronal activity decreases due to astrocyte effects, meaning that the depression effect prevails over the presynaptic signaling. Additionally, the longest bursts are obtained in simulations where the astrocytes form a significantly coupled network [especially "NN+A(20%)" and "NN+A(30%)"].

Our results therefore suggest that astrocytes may stabilize the activity of the neuronal network on a short-term (De Pittà et al., 2016): the astrocyte network would decrease neuronal activity through adenosine release when it is high or increase it through release-increasing presynaptic signaling when it is low. This homeostatic mechanism is based on the competition between two short-term synaptic plasticities regulated by gliotransmission: (1) gliotransmitter-based short-term increase of glutamate release by the presynaptic element and (2) short-term depression of the synapse via depressing gliotransmitters like adenosine. The system is homeostatic because (1) dominates (2) when neuronal activity is low, whereas (2) dominates (1) when neuronal activity is very large. That astrocytes could act as homeostatic regulators of the neuronal network activity has already been suggested based on the experimental observation that astrocytes release TNFα in response to prolonged periods of neuronal inactivity (De Pittà et al., 2016). At long time scales (hours to days) the released TNFα is expected to strengthen excitatory synapses while depressing inhibitory ones, thus contributing to the restoration of activity in the neuronal network (De Pittà et al., 2016). Our model adds to this possibility suggesting that astrocytes could also bring forth a further homeostatic mechanism based on competing processes of synaptic plasticity that could occur on fast time scales of the order of second or minutes. Consequently, future studies are required to better understand how astrocyte-mediated homeostasis on different time scales could ultimately mold neuronal network activity.

To investigate further if astrocytes contribute to network firing stability, we altered the recovery rate of the gliotransmitters and the average activation time of an astrocyte in case of "NN+A(30%)." As expected, the firing rate increased when the astrocytic activation time was increasing. Thus, the inhibiting effect of astrocytes—that dominates over the potentiating one—was diminished. For a longer recovery rate of the gliotransmitters, the astrocytes did not seem to have a clear effect on the network firing. The reason might be that the recovery/degradation was much faster than the time scale of neuronal activity.

Savtchenko and Rusakov (2014) presented a ring-like network model including pyramidal neurons and fast-spiking interneurons as well as volume-limited regulation of the synaptic efficacy. They used this latter mechanism as a way to emulate the spatially constrained effects of gliotransmission. The depression, e.g., upon astrocytic adenosine release, of the excitatory signals to the interneurons resulted in a decreased firing rate and network synchronization. In contrast, the facilitation. e.g., upon glutamate release, increased the firing rate while not altering much the network synchronization. In our simulations, the synaptic regulation from each astrocyte was also volume-limited but the astrocytes were inter-connected, allowing sequential activation of neighboring astrocytes. In addition, Savtchenko and Rusakov (2014) decoupled the potentiation or depression of synapses from the actual neuronal activity. In contrast, our simulations implemented a feedback loop between neuronal and astrocytic activity. Taken together, these differences make it unclear whether the same effects on network synchronization could be observed once the feedback loop is closed.

Recently, Paavilainen et al. (2018) compared hiPSC co-cultures aged 8+ weeks with hiPSC co-cultures aged 15+ weeks containing neuron and astrocyte networks. They observed a slight decrease in the spike rate for the hiPSC co-cultures aged 15+ weeks, together with an increase of the burst rate and duration, while the number of spikes per bursts was constant. Importantly, the hiPSC co-cultures aged 8+ weeks contained about 5% astrocytes and the hiPSC co-cultures aged 15+ weeks contained about 25 % astrocytes. Comparing our simulation results with 30% astrocytes to those with 10% astrocytes produces similar results (increased burst rate and duration, no change in spike count per burst), although the spike rates are similar in our case. Therefore, our model predicts that the change in activity observed in Paavilainen et al. (2018) could be due to the change in the astrocyte/neuron ratio. Currently, our computational model is established in 2D to resemble experimental *in vitro* data. However, it can be easily extended to 3D, and thus can give more insights on *in vivo* data.

While all of the mechanisms, pathways, and released gliotransmitters described in this paper have been adapted from astrocyte studies, the biological evidence that they co-exist in a single astrocyte is still sparse (Covelo and Araque, 2018). It is thus possible that the effects are a result of separate astrocyte populations or even astrocytes in different brain regions, just as neurons differ from one area to another. However, our model can simulate many of the subsets of astrocytic and neuronal mechanisms. Predictions about the functional role of astrocytes in neural networks are conceivable. In the future, it will be possible to adjust the model to specific combinations or even brain areas with differently functioning neurons and astrocytes.

To conclude, we have developed a neural network model in order to study the effect of astrocytes on neuronal network behavior. Our simulations show that astrocyte networks can act as homeostatic controllers with release-increasing and depressing effects on the synapse. These effects act on two different time scales for astrocytes and neurons. Our simulations suggest that tripartite synapses alone are not enough to produce these effects, and thus, the astrocytic network dynamics based on $IP_3$-controlled calcium waves are essential for understanding how astrocytes modify neuronal communication. The model presented here provides a basis for further studies

of neural interaction and the relevance of this interaction for brain function.

## DATA AVAILABILITY STATEMENT

Both the code and the raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

## AUTHOR CONTRIBUTIONS

KL, ES, AL-G, and JH designed and performed research. ES, KL, AL-G, and JL wrote analysis tools and analyzed the data. ES and KL wrote the first draft of the manuscript. KL, ES, JL, AL-G, HB, and JH contributed to the manuscript writing and revision. In addition, they have read and approved the submitted version.

## FUNDING

The research of KL, ES, and JH was supported by the 3DNeuroN project in the European Union's Seventh Framework Programme, Future and Emerging Technologies (grant agreement no 296590) and TEKES—the Finnish funding agency for innovation (Human Spare Part 2 Project). KL was funded by the Academy of Finland (decision nos. 314647, 326452). ES's project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement (No. 642563).

## ACKNOWLEDGMENTS

The authors thank Barbara Genocchi for her valuable comments on the manuscript.

## REFERENCES

Agarwal, A., Wu, P. H., Hughes, E. G., Fukaya, M., Tischfield, M. A., Langseth, A. J., et al. (2017). Transient opening of the mitochondrial permeability transition pore induces microdomain calcium transients in astrocyte processes. *Neuron* 93, 587–605.e7. doi: 10.1016/j.neuron.2016.12.034

Aleksin, S. G., Zheng, K., Rusakov, D. A., and Savtchenko, L. P. (2017). ARACHNE: a neural-neuroglial network builder with remotely controlled parallel computing. *PLoS Comput. Biol.* 13:e1005467. doi: 10.1371/journal.pcbi.1005467

Amiri, M., Hosseinmardi, N., Bahrami, F., and Janahmadi, M. (2013). Astrocyte-neuron interaction as a mechanism responsible for generation of neural synchrony: a study based on modeling and experiments. *J. Comput. Neurosci.* 34, 489–504. doi: 10.1007/s10827-012-0432-6

Araque, A., Carmignoto, G., and Haydon, P. G. (2001). Dynamic signaling between astrocytes and neurons. *Annu. Rev. Physiol.* 63, 795–813. doi: 10.1146/annurev.physiol.63.1.795

Araque, A., Carmignoto, G., Haydon, P. G., Oliet, S. H. R., Robitaille, R., and Volterra, A. (2014). Gliotransmitters travel in time and space. *Neuron* 81, 728–739. doi: 10.1016/j.neuron.2014.02.007

Araque, A., Parpura, V., Sanzgiri, R., and Haydon, P. (1999). Tripartite synapses: glia, the unacknowledged partner. *Trends Neurosci.* 22, 208–215. doi: 10.1016/S0166-2236(98)01349-6

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., et al. (2009). Equal numbers of neuronal and non-neuronal cells make the human brain an isometrically scaled-up primate brain. *J. Comp. Neurol.* 513, 532–541. doi: 10.1002/cne.21974

Bazargani, N., and Attwell, D. (2016). Astrocyte calcium signaling: the third wave. *Nat. Neurosci.* 19, 182–189. doi: 10.1038/nn.4201

Bezzi, P., Carmignoto, G., Pasti, L., Vesce, S., Rossi, D., Rizzini, B. L., et al. (1998). Prostaglandins stimulate calcium-dependent glutamate release in astrocytes. *Nature* 391, 281–285. doi: 10.1038/34651

Bindocci, E., Savtchouk, I., Liaudet, N., Becker, D., Carriero, G., and Volterra, A. (2017). Three-dimensional Ca2+ imaging advances understanding of astrocyte biology. *Science* 356:eaai8185. doi: 10.1126/science.aai8185

Boddum, K., Jensen, T. P., Magloire, V., Kristiansen, U., Rusakov, D. A., Pavlov, I., et al. (2016). Astrocytic GABA transporter activity modulates excitatory neurotransmission. *Nat. Commun.* 7:13572. doi: 10.1038/ncomms13572

Bonansco, C., Couve, A., Perea, G., Ferradas, C. Á., Roncagliolo, M., and Fuenzalida, M. (2011). Glutamate released spontaneously from astrocytes sets the threshold for synaptic plasticity. *Eur. J. Neurosci.* 33, 1483–1492. doi: 10.1111/j.1460-9568.2011.07631.x

Bowser, D. N., and Khakh, B. S. (2007). Vesicular ATP is the predominant cause of intercellular calcium waves in astrocytes. *J. Gen. Physiol.* 129, 485–491. doi: 10.1085/jgp.200709780

Bushong, E., a, Martone, M. E., Jones, Y. Z., and Ellisman, M. H. (2002). Protoplasmic astrocytes in CA1 stratum radiatum occupy separate anatomical domains. *J. Neurosci.* 22, 183–192. doi: 10.1523/JNEUROSCI.22-01-00183.2002

Charles, A. C., Kodali, S. K., and Tyndale, R. F. (1996). Intercellular calcium waves in neurons. *Mol. Cell. Neurosci.* 7, 337–353. doi: 10.1006/mcne.1996.0025

Clarke, L. E., and Barres, B. A. (2013). Emerging roles of astrocytes in neural circuit development. *Nat. Rev. Neurosci.* 14, 311–321. doi: 10.1038/nrn3484

Covelo, A., and Araque, A. (2018). Neuronal activity determines distinct gliotransmitter release from a single astrocyte. *Elife* 7:e32237. doi: 10.7554/eLife.32237

Dallérac, G., Chever, O., and Rouach, N. (2013). How do astrocytes shape synaptic transmission? Insights from electrophysiology. *Front. Cell. Neurosci.* 7:159. doi: 10.3389/fncel.2013.00159

De Pittà, M. (2019). "Gliotransmitter exocytosis and its consequences on synaptic transmission," in *Computational Glioscience*, eds. M. De Pittà and H. Berry (Cham: Springer International Publishing), 245–287. doi: 10.1007/978-3-030-00817-8_10

De Pittà, M., Ben-Jacob, E., and Berry, H. (2019). "G Protein-Coupled Receptor-Mediated Calcium Signaling in Astrocytes," in *Computational Glioscience*, eds M. De Pittà and H. Berry (Cham: Springer International Publishing), 115–150. doi: 10.1007/978-3-030-00817-8_5

De Pittà, M., Brunel, N., and Volterra, A. A. (2016). Astrocytes: orchestrating synaptic plasticity? *Neuroscience* 323, 43–61. doi: 10.1016/j.neuroscience.2015.04.001

De Pittà, M., Volman, V., Berry, H., and Ben-Jacob, E. (2011). A tale of two stories: astrocyte regulation of synaptic depression and facilitation. *PLoS Comput. Biol.* 7:e1002293. doi: 10.1371/journal.pcbi.1002293

De Pittà, M., Volman, V., Levine, H., Pioggia, G., De Rossi, D., and Ben-Jacob, E. (2008). Coexistence of amplitude and frequency modulations in intracellular calcium dynamics. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 77:030903. doi: 10.1103/PhysRevE.77.030903

Di Castro, M. A., Chuquet, J., Liaudet, N., Bhaukaurally, K., Santello, M., Bouvier, D., et al. (2011). Local Ca2+ detection and modulation of synaptic release by astrocytes. *Nat. Neurosci.* 14, 1276–1284. doi: 10.1038/nn.2929

Fellin, T. (2009). Communication between neurons and astrocytes: relevance to the modulation of synaptic and network activity. *J. Neurochem.* 108, 533–544. doi: 10.1111/j.1471-4159.2008.05830.x

Fellin, T., Pascual, O., Gobbo, S., Pozzan, T., Haydon, P. G., and Carmignoto, G. (2004). Neuronal synchrony mediated by astrocytic glutamate through activation of extrasynaptic NMDA receptors. *Neuron* 43, 729–743. doi: 10.1016/j.neuron.2004.08.011

Fiacco, T. A., and McCarthy, K. D. (2018). Multiple lines of evidence indicate that gliotransmission does not occur under physiological conditions. *J. Neurosci.* 38, 3–13. doi: 10.1523/JNEUROSCI.0016-17.2017

Giaume, C., Koulakoff, A., Roux, L., Holcman, D., and Rouach, N. (2010). Astroglial networks: a step further in neuroglial and gliovascular interactions. *Nat. Rev. Neurosci.* 11, 87–99. doi: 10.1038/nrn2757

Giugliano, M., Darbon, P., Arsiero, M., Lüscher, H.-R., and Streit, J. (2004). Single-neuron discharge properties and network activity in dissociated cultures of neocortex. *J. Neurophysiol.* 92, 977–996. doi: 10.1152/jn.00067.2004

Henneberger, C., Papouin, T., Oliet, S., and Rusakov, D. (2010). Long-term potentiation depends on release of D -serine from astrocytes. *Nature* 463, 232–236. doi: 10.1038/nature08673

Hertz, L., Gerkau, N. J., Xu, J., Durry, S., Song, D., Rose, C. R., et al. (2015). Roles of astrocytic Na+, K+-ATPase and glycogenolysis for K+ homeostasis in mammalian brain. *J. Neurosci. Res.* 93, 1019–1030. doi: 10.1002/jnr.23499

Hines, D. J., and Haydon, P. G. (2014). Astrocytic adenosine: from synapses to psychiatric disorders. *Philos. Trans. R. Soc. B Biol. Sci.* 369:20130594. doi: 10.1098/rstb.2013.0594

Jourdain, P., Bergersen, L. H., Bhaukaurally, K., Bezzi, P., Santello, M., Domercq, M., et al. (2007). Glutamate exocytosis from astrocytes controls synaptic strength. *Nat. Neurosci.* 10, 331–339. doi: 10.1038/nn1849

Kapucu, F. E, Tanskanen, J. M, Mikkonen, J. E., Ylä-Outinen, L., Narkilahti, S., and Hyttinen, J. A. (2012). Burst analysis tool for developing neuronal networks exhibiting highly varying action potential dynamics. *Front. Comput. Neurosci.* 6:38. doi: 10.3389/fncom.2012.00038

Kastanenka, K. V., Moreno-Bote, R., De Pittà, M., Perea, G., Eraso-Pichot, A., Masgrau, R., et al. (2019). A roadmap to integrate astrocytes into Systems Neuroscience. *Glia* 68, 5–26. doi: 10.1002/glia.23632

Kettenmann, H., and Verkhratsky, A. (2008). Neuroglia: the 150 years after. *Trends Neurosci.* 31, 653–659. doi: 10.1016/j.tins.2008.09.003

Lallouette, J., De Pittà, M., Ben-Jacob, E., and Berry, H. (2014). Sparse short-distance connections enhance calcium wave propagation in a 3D model of astrocyte networks. *Front. Comput. Neurosci.* 8:45. doi: 10.3389/fncom.2014.00045

Lallouette, J., De Pittà, M., and Berry, H. (2019). "Astrocyte Networks and Intercellular Calcium Propagation," in *Computational Glioscience*, eds M. De Pittà and H. Berry (Cham: Springer International Publishing), 177–210. doi: 10.1007/978-3-030-00817-8_7

Lee, S., Yoon, B.-E., Berglund, K., Oh, S.-J., Park, H., Shin, H.-S., et al. (2010). Channel-mediated tonic GABA release from Glia. *Science* 330, 790–796. doi: 10.1126/science.1184334

Lenk, K. (2011). "A simple phenomenological neuronal model with inhibitory and excitatory synapses," in *Proceedings of the 5th International Conference on Advances in Nonlinear Speech Processing* NOLISP'11, eds C. M. Travieso-González and J. B. Alonso-Hernández (Berlin: Springer-Verlag), 232–238. doi: 10.1007/978-3-642-25020-0_30

Lenk, K., Priwitzer, B., Ylä-Outinen, L., Tietz, L. H. B., Narkilahti, S., and Hyttinen, J. A. K. (2016). Simulation of developing human neuronal cell networks. *Biomed. Eng. Online* 15:105. doi: 10.1186/s12938-016-0226-6

Lorincz, M. L., Geall, F., Bao, Y., Crunelli, V., and Hughes, S. W. (2009). ATP-dependent infra-slow (<0.1 Hz) oscillations in thalamic networks. *PLoS ONE* 4:e4447. doi: 10.1371/journal.pone.0004447

McIver, S., Faideau, M., and Haydon, P. G. (2013). "Astrocyte-neuron communications," in *Neural-Immune Interactions in Brain Function and Alcohol Related Disorders*, eds C. Cui, L. Grandison, and A. Noronha (New York, NY: Springer Science & Business Media), 587. doi: 10.1007/978-1-4614-4729-0_2

Min, R., Santello, M., and Nevian, T. (2012). The computational power of astrocyte mediated synaptic plasticity. *Front. Comput. Neurosci.* 6:93. doi: 10.3389/fncom.2012.00093

Newman, E. A. (2003). Glial cell inhibition of neurons by release of ATP. *J. Neurosci.* 23, 1659–66. doi: 10.1523/JNEUROSCI.23-05-01659.2003

Oschmann, F., Berry, H., Obermayer, K., and Lenk, K. (2018). From *in silico* astrocyte cell models to neuron-astrocyte network models: a review. *Brain Res. Bull.* 136, 76–84. doi: 10.1016/j.brainresbull.2017.01.027

Paavilainen, T., Pelkonen, A., Mäkinen, M. E. L., Peltola, M., Huhtala, H., Fayuk, D., et al. (2018). Effect of prolonged differentiation on functional maturation of human pluripotent stem cell-derived neuronal cultures. *Stem Cell Res.* 27, 151–161. doi: 10.1016/j.scr.2018.01.018

Papouin, T., and Oliet, S. H. (2014). Organization, control and function of extrasynaptic NMDA receptors. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369:20130601. doi: 10.1098/rstb.2013.0601

Parri, H. R., Gould, T. M., and Crunelli, V. (2001). Spontaneous astrocytic Ca2+ oscillations *in situ* drive NMDAR-mediated neuronal excitation. *Nat. Neurosci.* 4, 803–812. doi: 10.1038/90507

Pasti, L., Zonta, M., Pozzan, T., Vicini, S., and Carmignoto, G. (2001). Cytosolic calcium oscillations in astrocytes may regulate exocytotic release of glutamate. *J. Neurosci.* 21, 477–484. doi: 10.1523/JNEUROSCI.21-02-00477.2001

Perea, G., and Araque, A. (2007). Astrocytes potentiate transmitter release at single hippocampal synapses. *Science* 317, 1083–1086. doi: 10.1126/science.1144640

Perea, G., Navarrete, M., and Araque, A. (2009). Tripartite synapses: astrocytes process and control synaptic information. *Trends Neurosci.* 32, 421–431. doi: 10.1016/j.tins.2009.05.001

Postnov, D. E., Koreshkov, R. N., Brazhe, N. A., Brazhe, A. R., and Sosnovtseva, O. V. (2009). Dynamical patterns of calcium signaling in a functional model of neuron-astrocyte networks. *J. Biol. Phys.* 35, 425–445. doi: 10.1007/s10867-009-9156-x

Sahlender, D. A., Savtchouk, I., and Volterra, A. (2014). What do we know about gliotransmitter release from astrocytes? *Philos. Trans. R. Soc. B Biol. Sci.* 369:20130592. doi: 10.1098/rstb.2013.0592

Savtchenko, L. P., and Rusakov, D. A. (2014). Regulation of rhythm genesis by volume-limited, astroglia-like signals in neural networks. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 369:20130614. doi: 10.1098/rstb.2013.0614

Savtchouk, I., and Volterra, A. (2018). Gliotransmission: beyond black-and-white. *J. Neurosci.* 38, 14–25. doi: 10.1523/JNEUROSCI.0017-17.2017

Schwarz, Y., Zhao, N., Kirchhoff, F., and Bruns, D. (2017). Astrocytes control synaptic strength by two distinct v-SNARE-dependent release pathways. *Nat. Neurosci.* 20, 1529–1539. doi: 10.1038/nn.4647

Shrivastava, A. N., Triller, A., and Sieghart, W. (2011). GABAA receptors: post-synaptic co-localization and cross-talk with other receptors. *Front. Cell. Neurosci.* 5:7. doi: 10.3389/fncel.2011.00007

Stimberg, M., Goodman, D. F. M., Brette, R., and De Pittà, M. (2019). "Modeling neuron–glia interactions with the brian 2 simulator," in *Computational Glioscience*, eds. M. De Pittà and H. Berry (Cham: Springer International Publishing), 471–505. doi: 10.1007/978-3-030-00817-8_18

Tsodyks, M., Pawelzik, K., and Markram, H. (1998). Neural networks with dynamic synapses. *Neural Networks* 835, 821–835. doi: 10.1162/089976698300017502

Tukker, A. M., Wijnolts, F. M. J., de Groot, A., and Westerink, R. H. S. (2018). Human iPSC-derived neuronal models for *in vitro* neurotoxicity assessment. *Neurotoxicology* 67, 215–225. doi: 10.1016/j.neuro.2018.06.007

Turrigiano, G. G. (2008). The self-tuning neuron: synaptic scaling of excitatory synapses. *Cell* 135, 422–435. doi: 10.1016/j.cell.2008.10.008

Valenza, G., Tedesco, L., Lanata, A., De Rossi, D., and Scilingo, E. P. (2013). Novel Spiking Neuron-Astrocyte Networks based on nonlinear transistor-like models of tripartite synapses. *Conf. Proc. Annu. Int. Conf. IEEE EMBS* 2013, 6559–6562. doi: 10.1109/EMBC.2013.6611058

Välkki, I. A., Lenk, K., Mikkonen, J. E., Kapucu, F. E., and Hyttinen, J. A. K. (2017). Network-wide adaptive burst detection depicts neuronal activity with improved accuracy. *Front. Comput. Neurosci.* 11:40. doi: 10.3389/fncom.2017.00040

Volterra, A., Liaudet, N., and Savtchouk, I. (2014). Astrocyte Ca2+ signalling: an unexpected complexity. *Nat. Rev. Neurosci.* 15, 327–335. doi: 10.1038/nrn3725

Wagenaar, D., Pine, J., and Potter, S. (2006). An extremely rich repertoire of bursting patterns during the development of cortical cultures. *BMC Neurosci.* 7:11. doi: 10.1186/1471-2202-7-11

Wallach, G., Lallouette, J., Herzog, N., De Pittà, M., Ben Jacob, E., Berry, H., et al. (2014). Glutamate mediated astrocytic filtering of neuronal activity. *PLoS Comput. Biol.* 10:e1003964. doi: 10.1371/journal.pcbi.1003964

Yoon, B.-E., and Lee, C. J. (2014). GABA as a rising gliotransmitter. *Front. Neural Circuits* 8:141. doi: 10.3389/fncir.2014.00141

Zorec, R., Araque, A., Carmignoto, G., Haydon, P. G., Verkhratsky, A., and Parpura, V. (2012). Astroglial excitability and gliotransmission: an appraisal of Ca2+ as a signalling route. *ASN Neuro.* 4, 103–119. doi: 10.1042/AN20110061

**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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 29 January 2020
doi: 10.3389/fncom.2020.00002

# A Machine Learning Approach to the Differentiation of Functional Magnetic Resonance Imaging Data of Chronic Fatigue Syndrome (CFS) From a Sedentary Control

*Destie Provenzano, Stuart D. Washington and James N. Baraniuk**

*Baraniuk Lab, Department of Medicine, Georgetown University Medical Center, Washington, DC, United States*

Chronic Fatigue Syndrome (CFS) is a debilitating condition estimated to impact at least 1 million individuals in the United States, however there persists controversy about its existence. Machine learning algorithms have become a powerful methodology for evaluating multi-regional areas of fMRI activation that can classify disease phenotype from sedentary control. Uncovering objective biomarkers such as an fMRI pattern is important for lending credibility to diagnosis of CFS. fMRI scans were evaluated for 69 patients (38 CFS and 31 Control) taken before (Day 1) and after (Day 2) a submaximal exercise test while undergoing the n-back memory paradigm. A predictive model was created by grouping fMRI voxels into the Automated Anatomical Labeling (AAL) atlas, splitting the data into a training and testing dataset, and feeding these inputs into a logistic regression to evaluate differences between CFS and control. Model results were cross-validated 10 times to ensure accuracy. Model results were able to differentiate CFS from sedentary controls at a 80% accuracy on Day 1 and 76% accuracy on Day 2 (**Table 3**). Recursive features selection identified 29 ROI's that significantly distinguished CFS from control on Day 1 and 28 ROI's on Day 2 with 10 regions of overlap shared with Day 1 (**Figure 3**). These 10 shared regions included the putamen, inferior frontal gyrus, orbital (F3O), supramarginal gyrus (SMG), temporal pole; superior temporal gyrus (T1P) and caudate ROIs. This study was able to uncover a pattern of activated neurological regions that differentiated CFS from Control. This pattern provides a first step toward developing fMRI as a diagnostic biomarker and suggests this methodology could be emulated for other disorders. We concluded that a logistic regression model performed on fMRI data significantly differentiated CFS from Control.

**Keywords: functional magnetic resonance imaging (fMRI), Chronic Fatigue Syndrome (CFS), logistic regression, machine learning, recursive feature elimination (RFE)**

OPEN ACCESS

***Edited by:***
*Ilan Goldberg,*
*Wolfson Medical Center, Israel*

***Reviewed by:***
*Michael Peer,*
*Hebrew University of Jerusalem, Israel*
*Roberto Santana,*
*University of the Basque Country, Spain*

****Correspondence:***
*James N. Baraniuk*
*baraniuj@georgetown.edu*

***Received:*** *08 October 2019*
***Accepted:*** *08 January 2020*
***Published:*** *29 January 2020*

***Citation:***
*Provenzano D, Washington SD and Baraniuk JN (2020) A Machine Learning Approach to the Differentiation of Functional Magnetic Resonance Imaging Data of Chronic Fatigue Syndrome (CFS) From a Sedentary Control. Front. Comput. Neurosci. 14:2. doi: 10.3389/fncom.2020.00002*

# INTRODUCTION

Chronic Fatigue Syndrome (CFS) is a debilitating condition estimated to affect at least 1 million individuals in the United States that causes $9.1 billion in annual losses in productivity (Centers for Disease Control Prevention, 2006). CFS is characterized by chronic persistent fatigue that is not alleviated with rest as well as pain, cognitive dysfunction, sleep abnormalities, and symptom relapse after minimal exertion (post-exertional malaise) (Fukuda et al., 1994; Carruthers et al., 2003; Centers for Disease Control Prevention, 2006; Committee on the Diagnostic Criteria for Myalgic Encephalomyelitis/Chronic Illness, 2015).

Controversy persists about the underlying etiology and pathophysiology of CFS, and there remains a need for objective measures of dysfunction to distinguish CFS from psychosocial etiologies like neurasthenia and depression (Pichot, 1994; Pearce, 2006; Committee on the Diagnostic Criteria for Myalgic Encephalomyelitis/Chronic Illness, 2015). Functional magnetic resonance imaging (fMRI) of the brain has shown to be a promising diagnostic tool because CFS subjects may have reduced gray matter thickness and cortical volume losses compared to age-matched controls (Okada et al., 2004), utilize more frontal and parietal regions during cognitive tasks than age-matched controls (cognitive compensation) (Lange et al., 2005), and have different activation patterns when making mistakes (De Lange et al., 2004). CFS subjects are less responsive than age-matched controls on tasks of auditory responsiveness (Tanaka et al., 2006) and demonstrate additional dysfunction following a light exercise task that may provide evidence for post-exertional malaise (Cook et al., 2017). Performance on n-back tasks indicated dysfunction on working memory (Caseras et al., 2006). The results of these studies provided a rationale to investigate if fMRI and the post-exertional malaise experienced by subjects with CFS could differentiate CFS subjects from a sedentary control.

Standard fMRI analysis seeks to compare univariate regions of brain activation at rest or during a task between CFS and control groups. However, multivariate classification methods have become an increasingly popular tool for identifying patterns of brain activity that can differentiate disease physiology (Cox and Savoy, 2003; Kriegeskorte et al., 2006; Haynes et al., 2007; De Martino et al., 2008; Ryali et al., 2010). Machine learning algorithms such as logistic regression, support vector machines, and random forests combined with feature selection can be applied to clustered voxel data to determine patterns of brain regions that may characterize a disorder (Mourão-Miranda et al., 2005; Pereira et al., 2009). We hypothesized that an acute physiology stressor such as a light exercise task combined with the implementation of a machine learning algorithm would allow us to identify a pattern of predominant behavior during fMRI scanning of CFS subjects while performing the n-back memory paradigm.

Subjects underwent fMRI scans on consecutive days while performing the continuous version of the n-back working memory test before (Day 1) and after (Day 2) a bicycle exercise stress test (Rayhan et al., 2013). Blood oxygenation level dependent (BOLD) signals were compared between groups on both days. We followed a standard approach of predictive model building for fMRI data involving feature extraction, model build, validation, and evaluation of performance (Sen et al., 2018). Voxel maps of activations from each subject were mapped to the Automatic Anatomical Labeling (AAL) atlas$^{AAL}$ using SPM12$^{SPM}$. Predictive model features were created from the number of significantly activated voxels for each AAL region for each subject run through a recursive features selection algorithm to identify importance. Data points were iteratively split into training and testing sets to create a logistic regression model (training set) and then validate the results (testing set). Model results were cross-validated to ensure performance. The output of this model was a multivariate pattern of activation that signified the cognitive differences between groups of CFS and sedentary control subjects. This strategy differs from the traditional fMRI analysis technique that quantify significant BOLD differences on voxel-by-voxel and regional basis. The outcomes provide a proof of concept for the implementation of a machine learning algorithm on fMRI data to create a diagnostic tool for CFS.

# METHODS

## Ethics

Subjects gave written informed consent for participation and use of all data for publication purposes. Studies were approved by the Georgetown University Institutional Review Board (IRB 2009-229, 2013-0943, 2015-0579) and U.S. Army Medical Research and Material Command (USAMRC) Human Research Protection Office (HRPO A-155547.0, A-18749), and registered on clinicaltrials.gov as NCT01291758, NCT03560830, and NCT03567811. All clinical investigations were conducted according to the principles expressed in the Declaration of Helsinki.

## Subjects

Data was collected from candidates who responded online or by phone or personal contact. Telephone screening after verbal informed consent was performed with 216 subjects, but 105 declined to participate or were excluded from participation after protocol explanation and assessment of chronic medical and psychiatric disease (Jones et al., 2009; Nater et al., 2009). Chronic Fatigue Syndrome was assessed by 1994 Fukuda CDC criteria by having 6 months of debilitating fatigue without medical or psychiatric cause plus at least for of the following eight criteria: problems with memory or concentration, sore throat, sore lymph nodes, myalgia, arthralgia, headache, sleep disturbance, and post-exertional malaise (Fukuda et al., 1994). Veterans with Gulf War Illness were examined by the same process and were excluded (Steele, 2000; Haynes et al., 2007).

Subjects were admitted to the Georgetown Howard Universities Clinical Translation Science Clinical Research Unit and were tested for the N-back working memory task in a 3T MRI scanner on two separate days. They underwent their first fMRI scan and N-back working memory task after overnight rest and then performed a submaximal exercise stress test. Subjects cycled at 70% of age-predicted maximum heart rate (220-age)

for 25 min, the ramped up their effort to reach 85% of predicted heart rate. On the next day they had that same submaximal exercise test followed by the second fMRI scan with n-back testing. This study reports on 38 subjects with Chronic Fatigue Syndrome and 31 sedentary controls.

## N-Back Task

Subjects practiced the complete n-back task with blocks of 0-back and 2-back loads in a mock scanner until they felt satisfied with their performance. fMRI data were collected on the non-exercise day (Day 1) and about 1 h after the second submaximal bicycle exercise stress test (Day 2).

The continuous version of the verbal N-back task is a challenging test of subject attention, memory, retrieval, and updating (Owen et al., 2005; Rayhan et al., 2013). Each 1 min long block had three components: 0-back task, 2-back task, and fixation between tasks. Subjects began each block with fixation by viewing a blank screen for 8 s. They proceeded to 0-back testing by viewing a string of nine letters (A, B, C, D) presented in random order for 2 s per letter. Subjects used both hands to press the button on a fiber-optic button box (ePrime software) that corresponded to the letter being viewed[1]. After another fixation period, they viewed a second string of nine letters for the 2-back task. Subjects had to remember the 1st and 2nd letters. When the 3rd letter was presented, they had to press the button corresponding to the letter seen "2-back" (the 1st letter seen 4 s before). The task was designed such that subjects orient, reorder, and engage their working memory to focus their attention in preparation for the next letter. Subjects used individual strategies to remember single letters in series (e.g., A-B-C-D) or through "chunks" (AB-BC-CD, or ABC-BCD). The 1-min blocks were repeated five times which produced time-series scans for 45 letters for 0-back stimulus response measurements (five blocks × nine responses) and 35 responses for the 2-back task (five blocks × seven responses each).

## Functional Magnetic Resonance Imaging (fMRI) Data Acquisition

fMRI acquisition was performed in a Siemens 3T Tim Trio scanner equipped with a transmit-receive body coil and a commercial 12-channel head coil array. Structural 3D T1-weighted Magnetization Prepared Rapid Acquisition Gradient Echo (MPRAGE) image parameters were: TR/TE = 1,900/2.52 ms, TI = 900 ms, field-of-view(FoV) = 250 mm, 176 slices, slice resolution = 1.0 mm, and voxel size 1 × 1 × 1 mm. Functional T2*-weighted gradient-echo planar imaging (EPI) parameters were: number of slices = 47, TR/TE = 2,000/30 ms, flip angle = 90°, matrix size = 64 × 64, FoV = 205 $mm^2$, and voxel size = 3.2 $mm^2$ (isotropic).

## Data Pre-processing

BOLD data was pre-processed through the default pipeline of the CONN version 17 toolbox (Whitfield-Gabrieli and Nieto-Castanon, 2012). Data underwent processing and spatial smoothing with a spatially stationary Gaussian filter of 6 mm full-width half maximum (FWHM) size through the SPM12 software (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/). SPM12 was used to account for movement artifacts between scans and functional anatomic differences not otherwise already compensated for. Spatially normalized images were converted into the Montreal Neurological Institute (MNI) standard stereotactic space (Mazziotta et al., 1995). Preprocessing included a slice-timing correction, outlier detection for Framewise Displacement based on Artifact Detection Tools, and realignment and unwarping of functional images. Spatial normalization resulted in a voxel size of 2.0 $mm^3$ (isotropic).

Preprocessed EPI data from individuals were modeled with the following events: instruction, fixation, 0-back, and 2-back. The 2-back > 0-back contrast was analyzed by one-sample *t*-test with motion parameters as covariates of no-interest. The residual 2-back > 0-back condition identified voxels that were significantly more activated during the high cognitive load 2-back than the low cognitive load 0-back periods. The optimal threshold *t*-value to identify significantly activated voxels was determined by plotting the number of significant voxels per subject as a function of *T*-values. The *T* value of 3.17 ($p < 0.001$ uncorrected) was selected.

Voxel data from the T-statistic maps were charted to MNI coordinates and grouped into regions defined by the Automated Anatomical Labeling Atlas (AAL) (Tzourio-Mazoyer et al., 2002) using a custom MATLAB program and functions from SPM12 and xiView 9.6[2]. The AAL atlas was chosen due to its widespread use and recognizability in SPM12, python, and the general fMRI community. The catalog of AAL regions with centers of mass and voxels per region was shown in **Table S1** and **Figure S1**[3]. The numbers of significant voxels per AAL region for each individual were the independent input variables that were fed into the feature selection process and logistic regression learner model. Our approach utilized a machine learning algorithm applied to the 3-D matrix of voxel data split using the binary outcome variable of CFS vs. control status.

## Feature Extraction

Model features (AAL regions with total activated voxels) were selected by a multistep feature reduction process.

Pearson's correlation coefficients were used as a preliminary variable selection methodology to determine highly correlation regions of brain activity. The number of significant voxels in every AAL region in the entire dataset (Testing + Training) was compared to every other AAL region to determine multicollinearity or which regions, if any, could be linearly predicted from the others with a substantial degree of accuracy. This created a matrix of correlations depicting Pearson's Correlation Coefficient for every region. When regions have a Pearson's Correlation Coefficient (R) of ≥0.9, it can be assumed that multicollinearity exists and that these regions should be removed or combined. Multicollinearity may not affect the overall predictive power of a model, but can impact

[1]http://www.pstnet.com/eprime.cfm

[2]http://www.alivelearn.net/xjview/

[3]https://figshare.com/articles/_Abbreviations_and_MNI_coordinates_of_AAL_/184981

the residual calculations of individual predictors and render the overall coefficients invalid (Belsley, 1991; O'Brien, 2007). Perfect multicollinearity causes the design matrix to have one less full rank and does not allow the ordinary least squares estimator to be inverted (Farrar and Glauber, 1967). Eliminating multicollinearity prevents against inaccurate machine learning algorithms, excessive standard errors for coefficients, and overfitting of models (Kumar, 1975; O'Hagan and McCabe, 1975). Multicollinearity was tested for three times on the training set, testing set, and training and testing set combined due to the small number of samples to ensure no multicollinearity existed for any combination of variables and that ordinary least squares (OLS) estimators could be obtained. The matrix depicting the training and testing set is depicted in the results.

Next the list of variables for model inputs was reduced using recursive feature elimination (RFE). Only data from the training set was fed into RFE and later, the logistic regression model. RFE is a feature selection method that fits a model by removing the weakest model input (feature) until a specified number of attributes remains or total accuracy level is reached. By eliminating a small number of inputs per loop in an iterative process, RFE attempts to reduce variable dependencies and collinearity that could otherwise impact a model. This data reduction step used the default recursive feature elimination (RFE) algorithm in the scikit-learn python package[4]. The principle of Occam's razor governs that the simplest set of inputs into a machine learning algorithm often leads to the most accurate result, as such this process attempted to whittle down the variables to as few as possible while still controlling for accuracy (Gauch, 2003).

Recursive feature elimination is a greedy feature elimination algorithm similar to sequential backward selection as found in a stepwise logistic regression. It was ultimately determined to use recursive feature elimination to remove excess inputs rather than use stepwise logistic regression to decrease bias in $R^2$ values, increase standard errors of the parameter estimates, increase confidence intervals, increase $p$-values, and unbias parameter estimates. Stepwise logistic regression can also exacerbate collinearity problems, which was especially important to account for given the small sample size.

## Predictive Model Build

Multiple predictive models were initially tested and evaluated before final presentation of results. These included a Support Vector Machine (SVM), Random Forest, Decision Tree, and Neural Net. Logistic Regression was the algorithm ultimately selected as it fast to build, repeatedly produced the most accurate and generalizable results, and is easy to implement in practice. Logistic regression is an algorithm used to determine the probability of a binary response to be dependent on one or more independent input variables (Walker and Duncan, 1967). A logistic regression works by attempting to fit a model that minimizes coefficients assigned to model inputs and maximizes total differentiated subgroups that fall into the classification region. Coefficients estimate the logarithm of the odds (log-odds) for a dependent variable based on the independent variables (Biondo et al., 2000). Corresponding coefficients for input variables are "regressed" from the data (Freedman, 2009). The model fits the data to the logit equation:

$$p(x) = 1/\left(1 + e^{-(\beta 0+\beta 1x1+\beta 2x2\ldots+\beta ixi)}\right)$$

where $\beta_0$ is the intercept (constant term), and $\beta_1$ and $\beta_2$ are the coefficients for variables $x_1$ and $x_2$, and $\beta_i$ represents coefficients for all subsequent variables (http://www.alivelearn.net/xjview/). Features fed into the model are assigned a coefficient that is reduced according to stochastic gradient descent until the best possible model (highest accuracy) remains. Stochastic gradient descent is a first order optimization algorithm that seeks to find the minimum of a function by taking steps proportional to the negative gradient of the function at every point (Barzilai and Borwein, 1998). The model was trained on a subgroup of the total dataset that was split into a stratified sample of disease and control subjects and tested on the remaining subgroup. This created a designated "training" and "testing" set. Each testing set was a distinct validation set created for each training set that did not overlap with the training set used to build the predictive model. Recursive feature elimination was run before each predictive model build on each respective training set. The ratio of training to test subjects was varied from 50:50 to 90:10 with the grouping of 70:30 selected to give the optimal model validation. This optimal ratio was determined by evaluating the final predictive power on the model. For example if a model rebuilt twice on two overlapping separate samples with a 90:10 ratio resulted in 89% accuracy and subsequent 15% accuracy on each respective 10% testing set, it was determined that this ratio resulted in overfitting and lack of generalizability once validated on the testing set. In contrast, the final selected ratio of 70:30 gave similar accuracies upon testing set validation across multiple re-sampled training and testing sets and multiple predictive model rebuilds.

## Validation and Evaluation of Performance

Model accuracy was tested by examining the total false positive rates, specificity, sensitivity on the designated testing set. Model generalizability was tested by cross validation on the testing set. Cross validation is a resampling method used to evaluate machine learning models such as logistic regression on limited data samples. Cross-validation seeks to understand the model's ability to predict new data that was not used in creation of the model. Cross-validation helps identify common problems such as overfitting and selection bias to evaluate how the predictive model might perform in practice. The model was cross validated 10 times using 10 subgroups (k-cross validation with $k = 10$) randomly drawn from the 30% test set (out-of-sample testing) to ensure generalizability. This cross-validation was done only within the testing set and included no data from the training set. Although for each partition the same training data and model is used, the 10 subgroups sampled from the test set are non-overlapping. Cross-validation was done for every predictive model rebuild on every ratio of training:testing set data and every re-sampled training set. The average results from the

[4]http://scikit-learn.org/stable/modules/feature_selection.html#rfe

cross-validations was used to estimate the model's predictive performance on future datasets. The averaged set of cross-validated outcomes provides a more accurate estimate of a model's predictive capability (Grossman et al., 2010).

The predictive model was iteratively re-built until the best set of inputs and model coefficients remained to allow for a high rate of accuracy and generalizability, or ability to be applied to new populations and maintain the same result. This means that the predictive model was built on multiple different re-sampled ratio's of training: testing set samples. For each training to testing split, the training set was also re-sampled from the original sample and subsequently validated and cross-validated on its respective testing set. The final result and model outcome was the ability of the combination of independent variables (model features) to predict the dependent binary variable (CFS vs. control status).

To test the significance of model accuracy, the models were then subjected to a "Shuffle Test." The labels on the subjects (CFS or SC) were shuffled in python using the built in sample function and passed through the model process 1,000 additional times to test if the original accuracy could be incurred by random chance. Each of the 1,000 runs trained the original model coefficients on a randomly selected stratified sample of 70% of each respective new shuffled sample (training set) and then tested the resultant model on a 30% randomly selected stratified sample (testing set) from this shuffled set to mimic the original conditions. The process was repeated on the entire shuffled sample 1,000 times. This process was repeated an additional 10,000 times if no accuracy greater than or equal to the original model accuracy could be obtained. If the Shuffle Test produced the model accuracy greater than or equal to the original model accuracy <5% of the time, it was determined the model was significant at the $p < 0.05$ level.

Logistic model coefficients must be treated with care. The logistic coefficient quantifies the rate of change in the "log odds" of the dependent variable as the input variable changes. The y-intercept term ($\beta_0$) is the log-odds of an outcome variable when all predictors are 0. In a multivariate model, the coefficients represented by $\beta_1$ to $\beta_i$ show the increase in log-odds relative to each other. For example, a coefficient of $\beta_i = 1$ multiplies the odds of $x_i$ by $10^1 = 10$, while a coefficient of 2 multiples the odds by $10^2 = 100$. These coefficients are highly dependent on other variable inputs to the model. A negative coefficient could indicate a negative relationship with the outcome variable and surrounding variables just as a positive coefficient could indicate a positive relationship, however one cannot ascertain the direction of correlation between any pair of variables in the outcome due to the nature of multivariate models and interactions between multiple variables.

## Visualization

The Wake Forest PICK ATLAS was used to select the AAL regions that contributed to each significant model and then data were imported into marsbar. These were displayed as color-coded axial slices (MRIcron).

Pearson's correlation coefficient was used to visually examine differences between CFS and sedentary control on Days 1 and 2 (Before and after exercise).

# RESULTS

## Demographics

All subjects had a sedentary lifestyle with <40 min of active aerobic work or exercise per week. The subjects spanned a similar age and BMI range, however due to the wider range of ages in the control group, age and gender were controlled in the final model build. CFS had significantly worse symptoms (Baraniuk et al., 2013) and quality of life (Ware and Gandek, 1998; **Table 1**).

## Selection of Threshold

Significant voxels were identified by calculating the number of voxels per brain scan at different levels of significance. Ultimately it was determined to use a threshold of $T \geq 3.17$ ($p \leq 0.001$) (**Figure 1**) due to its ability to allow a workable number of voxels while preserving significance.

## Feature Selection

Pearson's correlation coefficients were calculated between all AAL regions in the combined CFS and control dataset. All correlation coefficients were below 0.8 indicating that there was no collinearity between AAL regions or no significant dependency within model parameters on Day 1 and Day 2 for the groups (**Figure 2**). All regions were retained in the model because

**TABLE 1** | Demographics (mean ± SD).

| Group | SC | CFS |
|---|---|---|
| N | 31 | 38 |
| Age | 43.9 ± 16.3 | 47.74 + 16.46 |
| BMI | 28.4 ± 4.5 | 26.20 + 4.52 |
| Male | 19 (61.3%) | 10 (26.3%)† |
| White | 23 (74.2%) | 34 (89.4%)† |
| **CFS symptom severity scores** | | |
| Fatigue | 1.2 ± 1.0 | 3.4 + 0.8** |
| Memory and concentration | 1.0 ± 1.2 | 2.9 + 0.9** |
| Sore throat | 0.2 ± 0.6 | 1.0 + 1.0* |
| Sore lymph nodes | 0.1 ± 0.4 | 1.0 + 1.1* |
| Muscle pain | 0.6 ± 0.9 | 2.5 + 1.3** |
| Joint pain | 0.8 ± 1.0 | 1.8 + 1.4* |
| Headaches | 1.0 ± 1.3 | 2.0 + 1.3* |
| Sleep | 1.7 ± 1.4 | 3.2 + 0.9** |
| Exertional exhaustion | 0.5 ± 1.0 | 3.5 + 0.8** |
| **MOS SF-36** | | |
| Physical functioning | 88.8 ± 21.1 | 46.2 ± 26.3** |
| Role physical | 86.8 ± 31.5 | 9.2 ± 25.0** |
| Bodily pain | 85.9 ± 19.2 | 46.7 ± 26.7** |
| General health | 73.8 ± 21.9 | 34.6 ± 23.4** |
| Vitality | 64.9 ± 20.8 | 18.9 ± 15.7** |
| Social functioning | 85.3 ± 22.1 | 32.6 ± 27.0** |
| Role emotional | 90.2 ± 27.9 | 70.2 ± 44.4 |
| Mental health | 76.1 ± 16.9 | 67.6 ± 16.8 |
| Chalder fatigue score | 12.1 ± 4.5** | 22.8 ± 6.4** |

*Scale: 0 = none, 1 = trivial, 2 = mild, 3 = moderate, 4 = severe. Mean ± SD.
*$p < 0.001$ and **$p < 0.000001$ by 2-tailed unpaired Student's t-tests with Bonferroni corrections; †$p < 0.001$ by Fisher's Exact Test.

*R* was less than the cutoff point of 0.9 needed to justify removal. Although all regions were included, ultimately many of the 117 AAL regions were later removed in the recursive feature selection step and logistic regression.

## Model Results

### Significantly Activated Regions

Region of interest analysis identified areas that were significantly activated in each group (**Figure 3**). BOLD patterns for the 2-back > 0-back residual condition (2 > 0-back condition) were similar between CFS and controls and between Days 1 and 2 (**Figure 4**). Bilateral dorsolateral prefrontal cortex extending to the anterior insulae, dorsal anterior cingulate cortex, lateral parietal, and dorsal medial precuneus were activated. These match frontal parietal executive control, anterior salience, and dorsal attention networks (Laird et al., 2011; Rottschy et al., 2012). Exercise did not cause significant changes in BOLD of these regions.

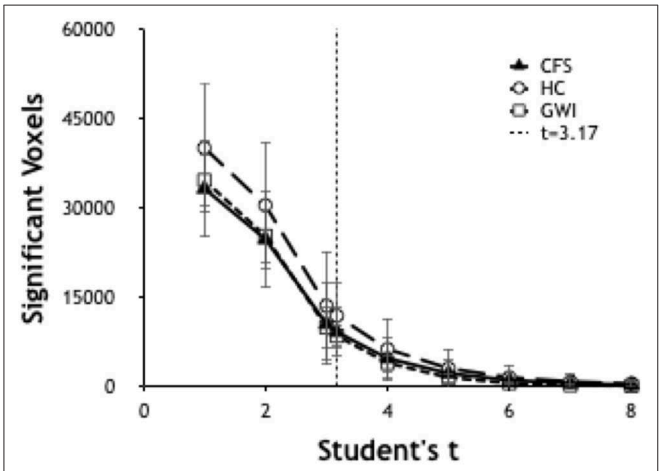


**FIGURE 1 |** Number of significant voxels plotted vs. *T*-values for each group. The number of significant voxels decreases at higher *T*-values and more rigorous *p*-values. *T* of 3.17 indicated $p < 0.001$.

### Differentially Activated Regions Found by Predictive Model Build

Logistic regression and recursive feature elimination identified three general patterns for regions that were differentially activated between CFS and controls. Ten AAL regions were selected by the logistic regression models on both days (**Table 2**), suggesting these ten regions may represent persistent indicators of CFS pathologies. In addition, 19 were significant only on Day 1, and 18 only on Day 2.

These 10 AAL regions were the right caudate, left and right putamen, left supramarginal gyrus (SMG), right postcentral gyrus (POST), right parahippocampus (PHIP), left inferior frontal gyrus orbital (F3O), right middle temporal gyrus (T2), left temporal pole; superior temporal gyrus (T1P), and the right cerebellum 8.

The 19 regions significant on Day 1 only were the left superior frontal gyrus; dorsolateral (F1), right superior frontal gyrus; dorsolateral (F1), right superior frontal gyrus; medial (F1M), right middle frontal gyrus; orbital (F2O), right gyrus rectus (GR), left middle frontal gyrus; orbital (F2O), right temporal pole; middle temporal gyrus (T2P), right supramarginal gyrus (SMG), left cerebellum 4 5, right vermis 6, left cerebellum 6, right supplementary motor area (SMA), left paracentral lobule (PCL), right rolandic operculum (RO), right cuneus (Q), right lingual gyrus (LING), left superior occipital lobe (O1), right middle occipital lobe (O2), right fusiform gyrus (FUSI).

The 18 regions significant on Day 2 only were the left and right pallidum (PAL), left and right calcarine fissure and surrounding cortex (V1), left middle occipital lobe (O2), left inferior occipital

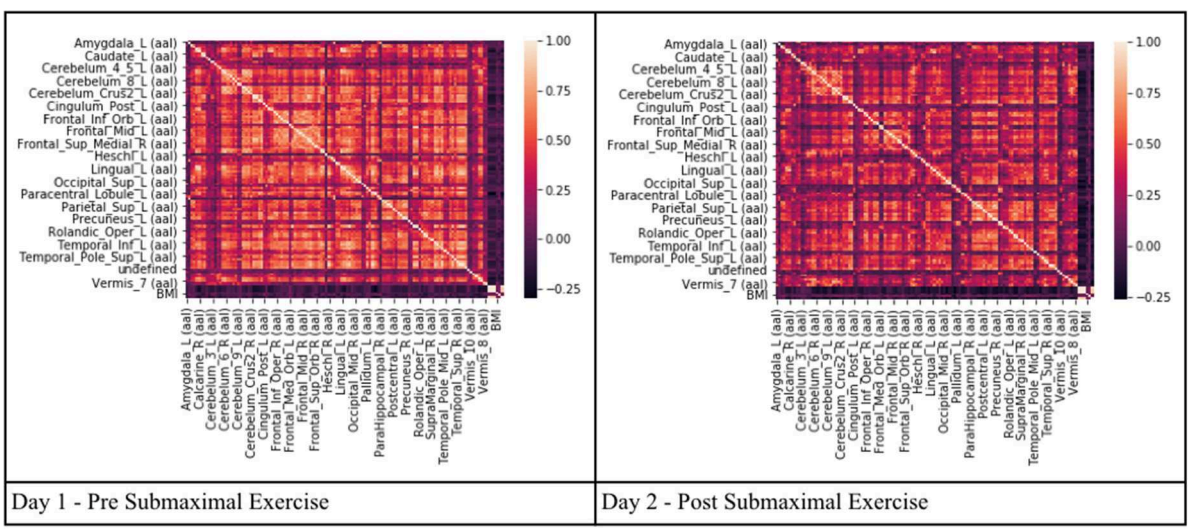


**FIGURE 2 |** Heat maps depicting Pearson's correlation coefficients (*R*) for all AAL regions in CFS and control datasets. The diagonal white line indicates $R = 1$. The x and y axis correspond to different regions of the brain according to the AAL atlas respectively, such that the diagonal line should be a perfect correlation (One region measured against itself) and the remaining are the cross product of the rest.

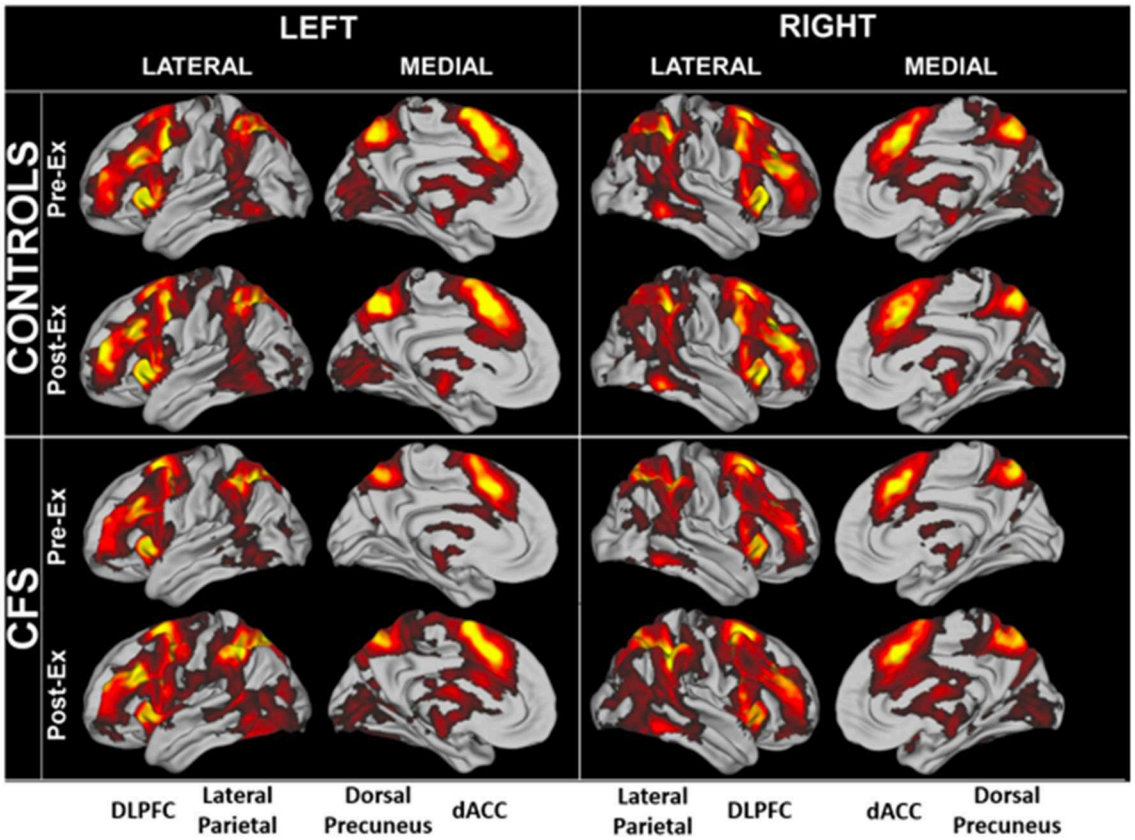


**FIGURE 3 |** Significantly elevated BOLD activity during the 2 > 0 back condition in CFS and control groups before and after exercise.

lobe (O3), right superior temporal gyrus (T1), right inferior frontal gyrus opercular (F3OP), right inferior gyrus triangular (F3T), right superior frontal gyrus orbital (F10), right superior frontal gyrus medial orbital (F1MO), left and right precuneus (PQ), left middle temporal gyrus (T2), left rolandic operculum (RO), left postcentral gyrus (POST), right cerebellum crus 1, and left cerebellum 9.

Another indication of significant differences was shown by looking at the patterns of Pearson's correlation coefficients between individual AAL regions between groups and days (**Figure 5**). SC on Day 1 had the highest number of correlations with $R \geq 0.7$ suggesting that control subjects were focused on the task on Day 1. SC subjects had fewer correlations with $R > 0.7$ on Day 2 suggesting that they exhibited learning, automaticity, and required a lower level of focus to complete the n-back task. CFS had fewer correlations on the pre-exercise MRI scan, different patterns of correlations from SC on both days, and poor similarity between Days 1 and 2. The different patterns of correlations between AAL regions supported the logistic regression analysis and demonstrated differences in connectivity between brain regions for CFS and SC before and after exercise. These outcomes predict that more advanced measures of functional connectivity (Rubinov and Sporns, 2010) will differ between CFS and SC before and after exercise and when depicting changes related to post-exertional malaise.

The composite multivariate pattern of activation differentiated CFS from Control with 80.9% accuracy on Day 1 and 76.1% accuracy on Day 2. Cross validation performed better than random on both days with a 65% accuracy on Day 1 and 57.5% accuracy on Day 2 (**Table 3**). Both the Day 1 and Day 2 models were able to correctly predict CFS from a SC greater than random chance (>0.5) due to this high predictive performances, however the Day 1 predictive model showed greater predictive power than the Day 2 model upon cross-validation. More samples in a future study would assist in validating this predictive performance.

The Shuffle Test reproduced an accuracy of 65% on 0 of the 1,000 shuffled test runs for Day 1. The Day 1 Shuffle Test had an average of 44% accuracy and mode of 37.5% accuracy for the 1,000 test runs. To ensure the statistical rigor of this method, the Shuffle Test was repeated for an additional 10,000 permutations on Day 1. A maximum accuracy of 69% was obtained and results for 65% accuracy or greater were found 11 times of 10,000 runs. Thus, it was determined the Day 1 model was significant at a $p < 0.01$ level. The Shuffle Test for Day 2 reproduced an accuracy of 57% or higher on 40 of 1,000 test runs. The Day 2 Shuffle Test had an average of 46% accuracy and mode of 43.5% accuracy for the

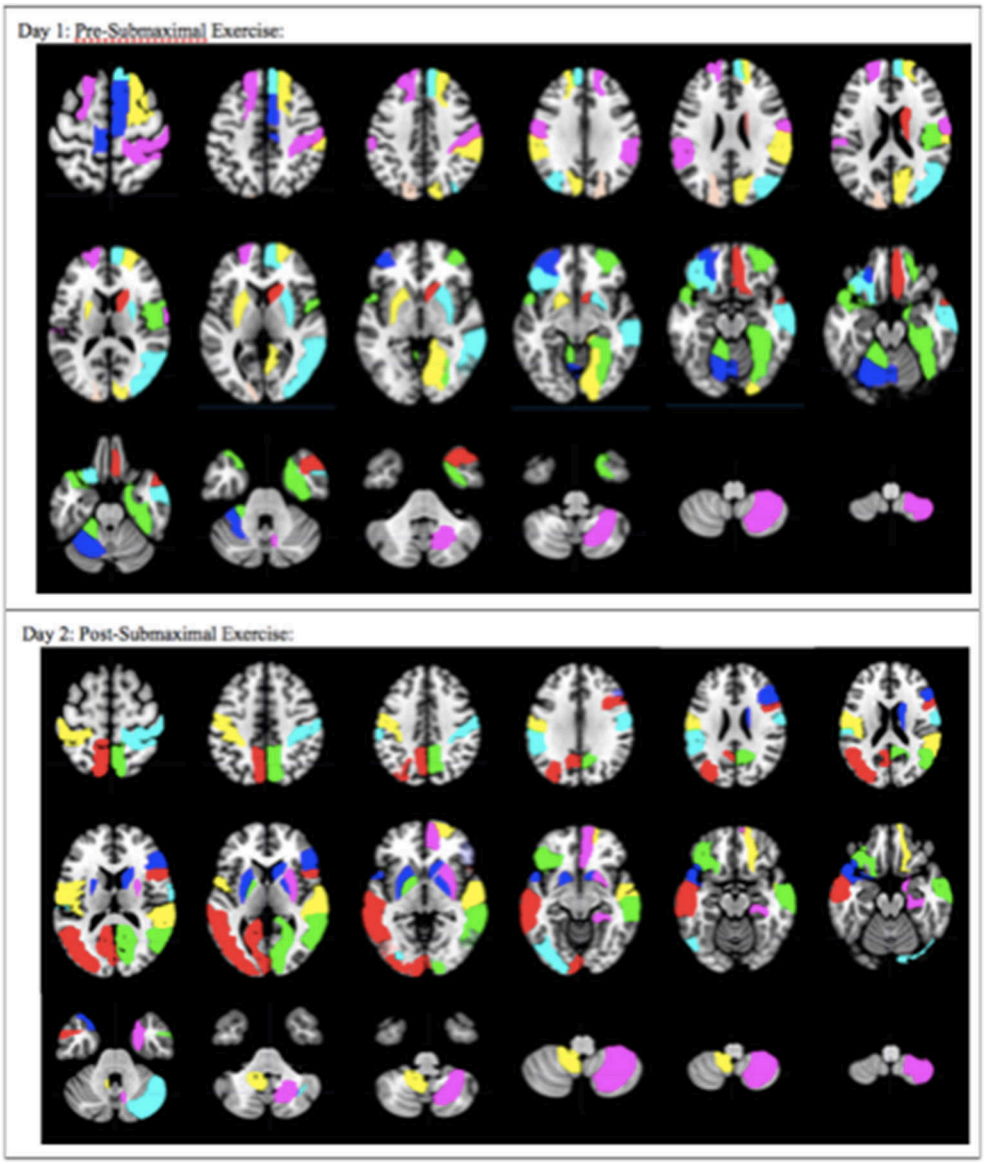


**FIGURE 4 |** Overall pattern depicting the difference in brain activation between CFS and sedentary control groups on Day 1 and Day 2. Axial slices show the pattern of 29 AAL regions that had significantly different numbers of activated voxels ($t > 3.17$, $p < 0.001$) in the 2 > 0-black condition based on logistic regression analysis. The complete pattern reflects the overall changes in all regions. Individual AAL regions are color coded for clarity. The colors do not indicate differences in BOLD signal intensity, *t*-values, logistic regression coefficients or Pearson's correlation coefficients for any single region between the two groups.

1,000 test runs. As both tests reproduced the model accuracy on <5% of 1,000 shuffled runs, it was determined that each model was significant at $p < 0.05$.

## DISCUSSION

This machine-learning approach was able to uncover a pattern of activated neurological regions that differentiated CFS from control subjects. The results of these two models indicate that machine learning algorithms combined with the voxel counts for activated regions grouped into the AAL Atlas was able to differentiate CFS from sedentary controls with good accuracy. The outcome indicates that analysis of fMRI data by machine learning algorithm(s) may lead to their use as part of a diagnostic tool that relies on cognitive aspects of CFS and their response to the physiological stressor of exercise. This may provide objective support for the concept of post-exertional malaise that is a central tenet of current subjectively defined CFS diagnostic criteria (Fukuda et al., 1994; Carruthers et al., 2003).

Ten AAL regions were significantly different according to the predictive model between SC and CFS before and after exercise and may represent persistent indicators of CFS pathologies. Left and right putamen and right caudate of the basal ganglia may be part of the Affective Network that has been identified by

**TABLE 2 |** AAL regions and logistic regression coefficients.

| AAL ID | AAL abbreviation | Day 1 | Day 2 |
|---|---|---|---|
| 72 | R Caudate_R (CAU) | 0.054 | 0.010 |
| 73 | L Putamen_L (PUT) | −0.375 | 0.065 |
| 74 | R Putamen_R (PUT) | −0.449 | 0.010 |
| 63 | L Supramarginal gyrus (SMG) | 0.294 | −0.008 |
| 58 | R Postcentral gyrus (POST) | −0.379 | −0.161 |
| 40 | R Parahippocampus (PHIP) | 0.092 | 0.015 |
| 15 | L Inferior frontal gyrus, orbital (F3O) | 0.142 | 0.214 |
| 86 | R Middle temporal gyrus (T2) | −0.106 | −0.153 |
| 83 | L Temporal pole; superior temporal gyrus (T1P) | −0.128 | −0.427 |
| 104 | R Cerebellum 8 | 0.115 | −0.028 |
| 3 | L Superior frontal gyrus, dorsolateral (F1) | −0.117 | |
| 4 | R Superior frontal gyrus, dorsolateral (F1) | 0.107 | |
| 24 | R Superior frontal gyrus, medial (F1M) | 0.303 | |
| 10 | R Middle frontal gyrus, orbital (F2O) | −0.262 | |
| 28 | R Gyrus rectus (GR) | −0.081 | |
| 9 | L Middle frontal gyrus, orbital (F2O) | 0.193 | |
| 88 | R Temporal pole; middle temporal gyrus (T2P) | 0.000 | |
| 64 | R Supramarginal gyrus (SMG) | 0.206 | |
| 97 | L Cerebellum 4 5 | 0.340 | |
| 112 | R Vermis 6 | −0.374 | |
| 99 | L Cerebellum 6 | −0.278 | |
| 20 | R Supplementary motor area (SMA) | −0.145 | |
| 69 | L Paracentral lobule (PCL) | −0.176 | |
| 18 | R Rolandic operculum (RO) | 0.534 | |
| 46 | R Cuneus (Q) | 0.566 | |
| 48 | R Lingual gyrus (LING) | −0.292 | |
| 49 | L Superior occipital lobe (O1) | 0.290 | |
| 52 | R Middle occipital lobe (O2) | −0.262 | |
| 56 | R Fusiform gyrus (FUSI) | 0.269 | |
| 75 | L Pallidum_L (PAL) | | −0.172 |
| 76 | R Pallidum_R (PAL) | | −0.062 |
| 43 | L Calcarine fissure and surrounding cortex (V1) | | −0.134 |
| 44 | R Calcarine fissure and surrounding cortex (V1) | | 0.122 |
| 51 | L Middle occipital lobe (O2) | | 0.203 |
| 53 | L Inferior occipital lobe (O3) | | −0.209 |
| 82 | R Superior temporal gyrus (T1) | | 0.252 |
| 12 | R Inferior frontal gyrus, opercular (F3OP) | | 0.139 |
| 14 | R Inferior frontal gyrus, triangular (F3T) | | −0.148 |
| 6 | R Superior frontal gyrus, orbital (F1O) | | −0.039 |
| 26 | R Superior frontal gyrus, medial orbital (F1MO) | | −0.178 |
| 67 | L Precuneus (PQ) | | −0.172 |
| 68 | R Precuneus (PQ) | | 0.107 |
| 85 | L Middle temporal gyrus (T2) | | 0.091 |
| 17 | L Rolandic operculum (RO) | | 0.542 |
| 57 | L Postcentral gyrus (POST) | | 0.071 |
| 92 | R Cerebellum crus 1 | | 0.086 |
| 105 | L Cerebellum 9 | | 0.095 |

*Ten regions were differentially activated between CFS and SC on both Days 1 and 2, with 17 regions only on Day 1 and 16 other regions only after exercise.*

meta-analysis of studies in anxiety (Xu et al., 2019). The primary sensory region (right S1) has been associated with heightened sensory awareness in panic disorder (Kim and Yoon, 2018). AAL regions of the ventromedial prefrontal cortex, temporal lobe, and parahippocampus overlapped with nodes from the default mode network (DMN) (Fox et al., 2015). They may be more related with subsets of the DMN related to rest and retrieval than forward thinking (Bellana et al., 2017). The left cerebellar hemisphere region eight has motor functions but is adjacent to regions having cognitive effects (Schmahmann, 2019).

Nineteen regions distinguished CFS from control only on Day 1 before exercise. They fit into several general patterns. Ventromedial and dorsomedial prefrontal cortex, parahippocampus and temporal pole regions are part of the default mode network (Mazziotta et al., 1995; Fox et al., 2015; Bellana et al., 2017). Cerebellar regions mediated working memory and emotional processing, and may have interacted with supplementary motor areas in pain and interoceptive dysfunction (Schmahmann, 2019). Occipital regions implicated visual functions. Bilateral supramarginal gyri suggested a role in the systemic hyperalgesia found in CFS (Lanz et al., 2011).

These 29 regions were differentially activated in CFS and controls before exercise. They included six ventromedial and dorsomedial frontal cortex regions of the anterior division of the DMN, and the right cuneus, right middle temporal gyrus (T2P), and right supramarginal gyrus from the posterior DMN (Laird et al., 2009; Fox et al., 2015). Heightened sensory awareness was implicated by activation of five regions of the visual network, right Rolandic operculum, supplementary motor areas (SMA), and cerebellum. The Rolandic operculum is the "little lid" of the parietal lobe that folds over the posterior insula. The bilateral Rolandic operculum integrates exteroceptive and interoceptive signals that are necessary for bodily self-consciousness and interoceptive awareness (Wager et al., 2013; Blefari et al., 2017) and is activated for maintenance of vigilant attention during simple tasks such as the stimulus-response 0-back task and discrimination tasks that require continuous decisions about alternative responses (e.g., go vs. no-go tasks) (Langner and Eickhoff, 2013).

The right supplementary motor area (SMA) and left paracentral lobule are functionally connected to cerebellar regions during pain processing (Coombes and Misra, 2016). Left cerebellar hemispheres 4, 5, and 6 have been implicated in working memory and generalized aversive processing, while right vermis six functions in emotional processing (Schmahmann, 2019). Visual regions may be differentially activated for attention (Vossel et al., 2014) or visual memory (Baldassano et al., 2016) during the n-back task.

After exercise, 18 other regions were activated. They included bilateral pallidum, precuneus, and superior frontal gyri, and visual cortex. These 28 regions were differentially activated on Day 2. Left and right pallidum joined other basal ganglia regions of the affective network (Xu et al., 2019). Bilateral precuneus, anterior insula, and sensorimotor cortex (Vossel et al., 2014), ventromedial frontal cortex and temporal regions suggest activation of the rostromedial frontal—lateral temporal subnetwork of the DMN (Bellana et al., 2017). Left and right precuneus may indicate DMN activation, or recruitment for cognitive compensation during the challenging 2-back task. Even though the dorsal precuneus is a node

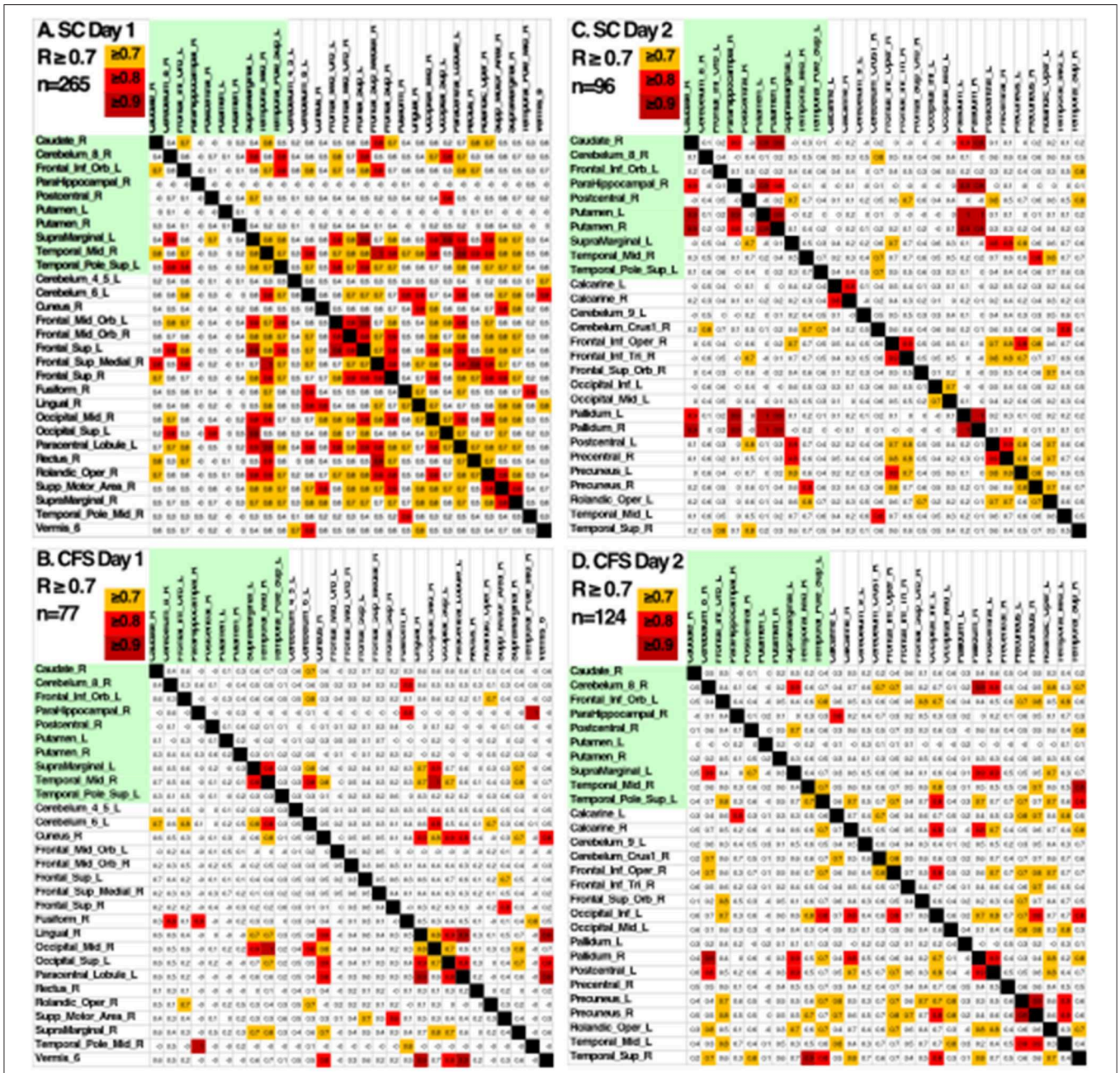


**FIGURE 5 |** Pearson's correlation coefficients between the BOLD activities in AAL regions that were differentially activated in SC and CFS on Days 1 and 2. AAL regions were listed in alphabetical order on the x- and y-axes for the 10 regions that were activated on Days 1 and 2 (yellow), on only on Day 1 or Day 2. Correlations were color coded in orange ($R > 0.7$), red ($R > 0.8$) and dark red ($R > 0.9$). As predicted by the logistic regression, CFS and SC had different patterns of correlations for the 10 shared regions (green) on Days 1 and 2, and between CFS and SC for the regions that were significantly different on Day 1 and Day 2.

in the DMN, it can be recruited into task systems during n-back testing (Rzucidlo et al., 2013). This is in contrast to the ventral precuneus that is only associated with DMN functions. Sensory activation was suggested by activation of bilateral postcentral gyri (S1) as found in panic disorder (Kim and Yoon, 2018). Occipital lobe visual network regions were particularly noteworthy and may indicate heightened vigilance, visual and memory analysis, or general sensory hypersensitivity. The ventral attention network was suggested by activation of the right ventrolateral prefrontal cortex (Fox et al., 2015). Attention and vigilance were implied from the activation of visual regions that can interact with dorsal attention network nodes in the intraparietal sulcus, and the right temporal parietal junction of the ventral attention network (Vossel et al., 2014). The left Rolandic operculum had the highest coefficient of any region, and was notable for its association with bodily self-consciousness, interoceptive and pain networks (Blefari et al., 2017).

**TABLE 3 |** Model results for day 1 (pre-submaximal exercise) and day 2 (post submaximal exercise).

| | Pre-exercise (day 1) | Pre-exercise (day 2) |
|---|---|---|
| Accuracy | 80.9% | 76.1% |
| 10x cross validation frequency | 65% | 57.5% |
| Sensitivity | 87.5% | 76.9% |
| Specificity | 76.9% | 75% |
| PPV | 70% | 83.3% |
| NPV | 90.9% | 66.7% |
| Significance as determined from Shuffle Test | $p < 0.01$ | $p < 0.05$ |
| Shuffle test average | 44% | 46% |
| Shuffle test mode | 37.5% | 43.5% |

*The accuracy and 10x cross validation frequency represent the corresponding model accuracy and accuracy after being ran through 10 smaller sub samplings of the initial testing set.*

The specific AAL regions that were differentially activated and selected by the logistic regression model were different between CFS and control on Days 1 and 2, but many of the regions were closely related because they belonged to the same functionally defined brain networks. Many belonged to the default mode network (DMN). Differential activation was found in the ventromedial and dorsomedial prefrontal cortex, hippocampus, lateral and temporal poles of the DMN, but with no significant differences for the medial posterior DMN nodes in the retrosplenial and posterior cingulate cortex regions (Laird et al., 2011; Fox et al., 2015). Affective network regions included basal ganglia, dorsal precuneus, sensorimotor regions, dACC and anterior insulae (Kim and Yoon, 2018). However, the amygdala was not differentially activated in CFS vs. control. The logistic regression included cerebellar and supplementary motor regions involved in working memory suggesting they were recruited as cognitive compensation or because of their interactions with sensorimotor regions during pain processing (Schmahmann, 2019). Cognitive compensation was suggested by the inclusion of the dorsal precuneus in the logistic regression on Day 2. Multimodal sensory intergration was suggested by visual and sensorimotor nodes, and on Day 2 by the addition of the ventral attention network. The Rolandic operculum, affective, cognitive, sensory, and attention network changes on Day 2 after exercise provocations may point to regions involved in post-exertional malaise in CFS. Involvement of these networks in the logistic regression was consistent with attention, memory and other cognitive dysfunction, chronic pain, systemic hyperalgesia and allodynia, negative emotion, and labile arousal that are part of the clinical presentation of CFS.

A limitation was the small sample size that created relatively small training and validation sets. The results of this pilot study can now be used to power larger studies to test the hypotheses proposed above. The nature of logistic regression means that individual regions of activation or deactivation of pathological significance for CFS cannot be determined from the model results alone. The coefficients assigned to input features are the "log odds" for the statistical models and not actual representations of increased or decreased BOLD activities. Because the variables depend on one another, it is the collective grouping of all AAL regions from the regression that ultimately show the difference between CFS and control. It is the entire pattern that transforms the fMRI data into a potential diagnostic biomarker. This methodology may be generalizable to allow sharing of fMRI data and creation of a diagnostic tool.

## CONCLUSION

The logistic regression model performed on fMRI data significantly differentiated CFS from control with model accuracy of 80.9% on Day 1 before exercise and 76.1% on Day 2 during the period of post-exertional malaise. Before exercise, CFS and control groups were different because of differential activation in default mode network nodes, and sensory perception networks involving visual, somatic, supplementary motor areas and cerebellar regions. These differences suggested dysfunction of attention and potential distraction by sensory processing in pain and interoception. Differential activation after exercise may indicate objective alterations related to post-exertional malaise involving frontal and lateral temporal nodes of the default mode network, sensory hypervigilance and attention using the left Rolandic operculum, visual network and the ventral attention network, and basal ganglia in the Affective Network.

## DATA AVAILABILITY STATEMENT

The final compiled data that was generalized and analyzed for this study can be found within article tables. Individual patient records are not published, however could be de-identified and made available upon request.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Georgetown University Institutional Review Board (IRB 2009-229, 2013-0943, 2015-0579) and U.S. Army Medical Research and Material Command (USAMRC) Human Research Protection Office (HRPO A-155547.0, A-18749). The patients/participants provided their written informed consent to participate in this study.

## AUTHOR CONTRIBUTIONS

The predictive model experiment design, predictive model build, data analysis, visualization, and final article write up were performed by DP. SW performed data pre-processing, visualization, and review of article write up. JB performed initial data collection, visualization, article review, and project oversight.

## FUNDING

The study was supported by funding from The Sergeant Sullivan Circle, Dr. Barbara Cottone, Dean Clarke Bridge

Prize, Department of Defense Congressionally Directed Medical Research Program (CDMRP) W81XWH-15-1-0679 and W81-XWH-09-1-0526, and the National Institute of Neurological Disorders and Stroke R21NS088138 and RO1NS085131.

## ACKNOWLEDGMENTS

We would like to acknowledge Georgetown University, The Sergeant Sullivan Circle, Dr. Barbara Cottone, Dean Clarke Bridge Prize, Department of Defense Congressionally Directed Medical Research Program, and the National Institute of Neurological Disorders and Stroke for support of this study that aided in the efforts of the authors.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.00002/full#supplementary-material

## REFERENCES

Baldassano, C., Esteva, A., Fei-Fei, L., and Beck, D. M. (2016). Two Distinct Scene-Processing Networks Connecting Vision and Memory. *eNeuro*. 3:ENEURO.0178-16.2016. doi: 10.1523/ENEURO.0178-16.2016

Baraniuk, J. N., Adewuyi, O., Merck, S. J., Ali, M., Ravindran, M. K., Timbol, C. R., et al. (2013). A Chronic Fatigue Syndrome (CFS) severity score based on case designation criteria. *Am. J. Transl. Res.* 5, 53–68.

Barzilai, J., and Borwein, J. M. (1998). Two-point step size gradient methods. *IMA J. Numerical Anal.* 8, 141–148. doi: 10.1093/imanum/8.1.141

Bellana, B., Liu, Z. X., Diamond, N. B., Grady, C. L., and Moscovitch, M. (2017). Similarities and differences in the default mode net-work across rest, retrieval, and future imagining. *Hum. Brain Mapp.* 38, 1155–1171. doi: 10.1002/hbm.23445

Belsley, D. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York, NY: Wiley.

Biondo, S., Ramos, E., Deiros, M., Ragué, J. M., De Oca, J., Moreno, P., et al. (2000). Prognostic factors for mortality in left colonic peritonitis: a new scoring system. *J. Am. Coll. Surg.* 191, 635–642. doi: 10.1016/S1072-7515(00)00758-4

Blefari, M. L., Martuzzi, R., Salomon, R., Bello-Ruiz, J., Herbelin, B., Serino, A., et al. (2017). Bilateral Rolandic operculum processing underlying heartbeat awareness reflects changes in bodily self-consciousness. *Eur. J. Neurosci.* 45, 1300–1312. doi: 10.1111/ejn.13567

Carruthers, B. M., Jain, A. K., De Meirleir, K. L., Peterson, D. L., Klimas, N. G., Lerner, A. M., et al. (2003). Myalgic encephalomyelitis/chronic fatigue syndrome. *J. Chronic Fatigue Syndr.* 11, 7–115. doi: 10.1300/J092v11n01_02

Caseras, X., Mataix-Cols, D., Giampietro, V., Rimes, K. A., Brammer, M., Zelaya, F., et al. (2006). Probing the working memory system in chronic fatigue syndrome: a functional magnetic resonance imaging study using the n-back task. *Psychosom. Med.* 68, 947–955. doi: 10.1097/01.psy.0000242770.50979.5f

Centers for Disease Control and Prevention (2006). *Chronic Fatigue Syndrome*. Available online at: www.cdc.gov/cfs (accessed September 8, 2018).

Committee on the Diagnostic Criteria for Myalgic Encephalomyelitis/Chronic Illness (2015). *Fatigue Syndrome, Board on the Health of Select Populations, Institute of Medicine. Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Redefining an Illness*. Washington, DC: National Academies Press (US). Available online at: https://www.nap.edu/catalog/19012/beyond-myalgic-encephalomyelitischronic-fatigue-syndrome-redefining-an-illness (accessed February 10, 2017).

Cook, D. B., Light, A. R., Light, K. C., Broderick, G., Shields, M. R., Dougherty, R. J., et al. (2017). Neural consequences of post-exertion malaise in myalgic encephalomyelitis/chronic fatigue syndrome. *Brain Behav. Immun.* 62, 87–99. doi: 10.1016/j.bbi.2017.02.009

Coombes, S. A., and Misra, G. (2016). Pain and motor processing in the human cerebellum. *Pain* 157, 117–127. doi: 10.1097/j.pain.0000000000000337

Cox, D. D., and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage* 19, 261–270. doi: 10.1016/S1053-8119(03)00049-1

De Lange, F. P., Kalkman, J. S., Bleijenberg, G., Hagoort, P., van der Werf, S. P., van der Meer, J. W., et al. (2004). Neural correlates of the chronic fatigue syndrome: an fMRI study *Brain*. 127(pt 9), 1948–1957. doi: 10.1093/brain/awh225

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., and Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage* 43, 44–58. doi: 10.1016/j.neuroimage.2008.06.037

Farrar, D. E., and Glauber, R. R. (1967). Multicollinearity in regression analysis: the problem revisited. *Rev. Econ. Stat.* 49, 92–107. doi: 10.2307/1937887

Fox, K. C., Spreng, R. N., Ellamil, M., Andrews-Hanna, J. R., and Christoff, K. (2015). The wandering brain: meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *Neuroimage* 111, 611–621. doi: 10.1016/j.neuroimage.2015.02.039

Freedman, D. A. (2009). *Statistical Models: Theory and Practice*. Berkeley, CA: Cambridge University Press.

Fukuda, K., Straus, S. E., Hickie, I., Sharpe, M. C., Dobbins, J. G., and Komaroff, A. (1994). The chronic fatigue syndrome: a comprehensive approach to its definition and study. International Chronic Fatigue Syndrome Study Group. *Ann Intern Med.* 121, 953–959. doi: 10.7326/0003-4819-121-12-199412150-00009

Gauch, H. G. Jr. (2003). *Scientific Method in Practice*. New York, NY: Cambridge University Press.

Grossman, R., Seni, G., Elder, J., Agarwal, N., and Liu, H. (2010). *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Chicago, IL: Morgan Claypool. doi: 10.2200/S00240ED1V01Y200912DMK002

Haynes, J. D., Sakai, K., Rees, G., Gilbert, S., Frith, C., and Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Curr. Biol.* 17, 323–328. doi: 10.1016/j.cub.2006.11.072

Jones, J. F., Lin, J. M., Maloney, E. M., Boneva, R. S., Nater, U. M., Unger, E. R., et al. (2009). An evaluation of exclusionary medical/psychiatric conditions in the definition of chronic fatigue syndrome. *BMC Med.* 7:57. doi: 10.1186/1741-7015-7-57

Kim, Y. K., and Yoon, H. K. (2018). Common and distinct brain networks underlying panic and social anxiety disorders. *Prog. Neuropsychopharmacol. Biol. Psychiatr.* 80(Pt B), 115–122. doi: 10.1016/j.pnpbp.2017.06.017

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. *Proc. Natl. Acad. Sci. U.S.A.* 103, 3863–3868. doi: 10.1073/pnas.0600244103

Kumar, T. K. (1975). Multicollinearity in regression analysis. *Rev. Econ. Stat.* 57, 365–366. doi: 10.2307/1923925

Laird, A. R., Eickhoff, S. B., Li, K., Robin, D. A., Glahn, D. C., and Fox, P. T. (2009). Investigating the functional heterogeneity of the default mode network using coordinate-based meta-analytic modeling. *J. Neurosci.* 29, 14496–14505. doi: 10.1523/JNEUROSCI.4004-09.2009

Laird, A. R., Fox, P. M., Eickhoff, S. B., Turner, J. A., Ray, K. L., McKay, D. R., et al. (2011). Behavioral interpretations of intrinsic connectivity networks. *J. Cogn. Neurosci.* 23, 4022–4037. doi: 10.1162/jocn_a_00077

Lange, G., Steffener, J., Cook, D. B., Bly, B. M., Christodoulou, C., Liu, W. C., et al. (2005). Objective evidence of cognitive complaints in chronic fatigue syndrome: a BOLD fMRI study of verbal working memory. *Neuroimage* 26, 513–524. doi: 10.1016/j.neuroimage.2005.02.011

Langner, R., and Eickhoff, S. B. (2013). Sustaining attention to simple tasks: a meta-analytic review of the neural mechanisms of vigilant attention. *Psychol. Bull.* 139, 870–900. doi: 10.1037/a0030694

Lanz, S., Seifert, F., and Maihöfner, C. (2011). Brain activity associated with pain, hyperalgesia and allodynia: an ALE meta-analysis. *J. Neural. Transm*. 118, 1139–1154. doi: 10.1007/s00702-011-0606-9

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., and Lancaster, J. (1995). A probabilistic atlas of the human brain: theory and rationale for its development: The International Consortium for Brain Mapping (ICBM). *NeuroImage* 2, 89–101. doi: 10.1006/nimg.1995.1012

Mourão-Miranda, J., Bokde, A. L., Born, C., Hampel, H., and Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: support vector machine on functional MRI data. *Neuroimage* 28, 980–995. doi: 10.1016/j.neuroimage.2005.06.070

Nater, U. M., Lin, J. M., Maloney, E. M., Jones, J. F., Tian, H., Boneva, R. S., et al. (2009). Psychiatric comorbidity in persons with chronic fatigue syndrome identified from the Georgia population. *Psychosom. Med*. 71, 557–565. doi: 10.1097/PSY.0b013e31819ea179

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Qual. Quant*. 41:673. doi: 10.1007/s11135-006-9018-6

O'Hagan, J., and McCabe, B. (1975). Tests for the severity of multicolinearity in regression analysis: a comment. *Rev. Econ. Stat*. 57, 368–370. doi: 10.2307/1923927

Okada, T., Tanaka, M., Kuratsune, H., Watanabe, Y., and Sadato, N. (2004). Mechanisms underlying fatigue: a voxel-based morphometric study of chronic fatigue syndrome. *BMC Neurol*. 4:14. doi: 10.1186/1471-237 7-4-14

Owen, A. M., McMillan, K. M., Laird, A. R., and Bullmore, E. (2005). N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies. *Hum. Brain Mapp*. 25, 46–59. doi: 10.1002/hbm. 20131

Pearce, J. M. (2006). The enigma of chronic fatigue. *Eur. Neurol*. 56, 31–36. doi: 10.1159/000095138

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*. 45, S199–S209. doi: 10.1016/j.neuroimage.2008.11.007

Pichot, P. (1994). Neurasthenia, yesterday and today. *Encephale* 20(Spec No 3), 545–549.

Rayhan, R. U., Stevens, B. W., Raksit, M. P., Ripple, J. A., Timbol, C. R., Adewuyi, O., et al. (2013). Exercise challenge in Gulf War Illness reveals two subgroups with altered brain structure and function. *PLoS ONE* 8:e63903. doi: 10.1371/journal.pone.0063903

Rottschy, C., Langner, R., Dogan, I., Reetz, K., Laird, A. R., Schulz, J. B., et al. (2012). Modelling neural correlates of working memory: a coordinate-based meta-analysis. *Neuroimage* 60, 830–846. doi: 10.1016/j.neuroimage.2011. 11.050

Rubinov, M., and Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* 52, 1059–1069. doi: 10.1016/j.neuroimage.2009.10.003

Ryali, S., Supekar, K., Abrams, D. A., and Menon, V. (2010). Sparse logistic regression for whole brain classification of fMRI data. *NeuroImage* 51, 752–764. doi: 10.1016/j.neuroimage.2010.02.040

Rzucidlo, J. K., Roseman, P. L., Laurienti, P. J., and Dagenbach, D. (2013). Stability of whole brain and regional network topology within and between resting and cognitive states. *PLoS ONE* 8:e70275. doi: 10.1371/journal.pone.0070275

Schmahmann, J. D. (2019). The cerebellum and cognition. *Neurosci. Lett*. 688, 62–75. doi: 10.1016/j.neulet.2018.07.005

Sen, B., Borle, N. C., Greiner, R., and Brown, M. R. G. (2018). A general prediction model for the detection of ADHD and Autism using structural and functional MRI. *PLoS ONE* 13:e0194856. doi: 10.1371/journal.pone.0194856

Steele, L. (2000). Prevalence and patterns of Gulf War illness in Kansas Veterans: association of symptoms with characteristics of person, place, and time of Military Service. *Am. J. Epidemiol*. 152, 992–1002. doi: 10.1093/aje/152.10.992

Tanaka, M., Sadato, N., Okada, T., Mizuno, K., Sasabe, T., Tanabe, H. C., et al. (2006). Reduced responsiveness is an essential feature of chronic fatigue syndrome: a fMRI study. *BMC Neurol*. 6:9. doi: 10.1186/1471-2377-6-9

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Vossel, S., Geng, J. J., and Fink, G. R. (2014). Dorsal and ventral attention systems: distinct neural circuits but collaborative roles. *Neuroscientist* 20, 150–159. doi: 10.1177/1073858413494269

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C. W., and Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *N. Engl. J. Med*. 368, 1388–1397. doi: 10.1056/NEJMoa1204471

Walker, S. H., and Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika* 54, 167–178. doi: 10.2307/2333860

Ware, J. E., and Gandek, B. (1998). Overview of the SF-36 Health Survey and the International Quality of Life Assessment (IQOLA) Project. *J. Clin. Epidemiol*. 51, 903–912. doi: 10.1016/S0895-4356(98)00081-X

Whitfield-Gabrieli, S., and Nieto-Castanon, A. (2012). Conn: A functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect*. 2, 125–141. doi: 10.1089/brain.2012.0073

Xu, J., Van Dam, N. T., Feng, C., Luo, Y., Ai, H., Gu, R., et al. (2019). Anxious brain networks: a coordinate-based activation likelihood estimation meta-analysis of resting-state functional connectivi-ty studies in anxiety. *Neurosci. Biobehav. Rev*. 96, 21–30. doi: 10.1016/j.neubiorev.2018.11.005

**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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 04 February 2020
doi: 10.3389/fncom.2020.00005

# Modeling the Effect of Temperature on Membrane Response of Light Stimulation in Optogenetically-Targeted Neurons

*Helton M. Peixoto*[1,2,3], *Rossana M. S. Cruz*[4], *Thiago C. Moulin*[5] *and Richardson N. Leão*[2,3]*

[1] School of Science and Technology (ECT), Federal University of Rio Grande do Norte (UFRN), Natal, Brazil, [2] Neurodynamics Lab, Brain Institute, Federal University of Rio Grande do Norte, Natal, Brazil, [3] Developmental Genetics Unit, Neurodynamics Lab, Department of Neuroscience, Uppsala, Sweden, [4] Electrical Engineering Department, Federal Institute of Paraiba (IFPB), Joao Pessoa, Brazil, [5] Institute of Medical Biochemistry, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil

Optogenetics is revolutionizing Neuroscience, but an often neglected effect of light stimulation of the brain is the generation of heat. In extreme cases, light-generated heat kills neurons, but mild temperature changes alter neuronal function. To date, most *in vivo* experiments rely on light stimulation of neural tissue using fiber-coupled lasers of various wavelengths. Brain tissue is irradiated with high light power that can be deleterious to neuronal function. Furthermore, absorbed light generates heat that can lead to permanent tissue damage and affect neuronal excitability. Thus, light alone can generate effects in neuronal function that are unrelated to the genuine "optogenetic effect." In this work, we perform a theoretical analysis to investigate the effects of heat transfer in rodent brain tissue for standard optogenetic protocols. More precisely, we first use the Kubelka-Munk model for light propagation in brain tissue to observe the absorption phenomenon. Then, we model the optothermal effect considering the common laser wavelengths (473 and 593 *nm*) used in optogenetic experiments approaching the time/space numerical solution of Pennes' bio-heat equation with the Finite Element Method. Finally, we then modeled channelrhodopsin-2 in a single and spontaneous-firing neuron to explore the effect of heat in light stimulated neurons. We found that, at commonly used light intensities, laser radiation considerably increases the temperature in the surrounding tissue. This effect alters action potential size and shape and causes an increase in spontaneous firing frequency in a neuron model. However, the shortening of activation time constants generated by heat in the single firing neuron model produces action potential failures in response to light stimulation. We also found changes in the power spectrum density and a reduction in the time required for synchronization in an interneuron network model of gamma oscillations. Our findings indicate that light stimulation with intensities used in optogenetic experiments may affect neuronal function not only by direct excitation of light sensitive ion channels and/or pumps but also by generating heat. This approach serves as a guide to design optogenetic experiments that minimize the role of tissue heating in the experimental outcome.

**Keywords: optogenetics, bio-heat, temperature, finite element method, Hodgkin-Huxley model**

OPEN ACCESS

***Edited by:***
*Germán Mato, Bariloche Atomic Centre (CNEA), Argentina*

***Reviewed by:***
*Marcus Thomas Wilson, University of Waikato, New Zealand*
*Albrecht Stroh, Johannes Gutenberg University Mainz, Germany*

***Correspondence:**
*Richardson N. Leão*
*richardson.leao@neuro.ufrn.br*

***Received:*** *18 June 2019*
***Accepted:*** *14 January 2020*
***Published:*** *04 February 2020*

***Citation:***
*Peixoto HM, Cruz RMS, Moulin TC and Leão RN (2020) Modeling the Effect of Temperature on Membrane Response of Light Stimulation in Optogenetically-Targeted Neurons. Front. Comput. Neurosci. 14:5. doi: 10.3389/fncom.2020.00005*

## INTRODUCTION

Optogenetics refers to a group of techniques that rely on genetics and optics for the deterministic control or study of (generally excitable) cells from a similar genetic background (Fenno et al., 2011). The radical idea of using light-driven ion channels and pumps from unicellular organisms to modulate neurons was pioneered by Deisseroth, Nagel, and Boyden and has now spread to neuroscience laboratories throughout the world (Knöpfel et al., 2010; Fenno et al., 2011). Limiting factors of the technique include the availability of genetic markers (Lerchner et al., 2014), the invasiveness of the gene delivery and especially difficulties of delivering light throughout large brain volumes (Lerchner et al., 2014). Perhaps for these reasons, optogenetics studies are vastly more common in small animals, especially mice and rats (Aravanis et al., 2007; Madisen et al., 2012).

To date, most *in vivo* experiments rely on light stimulation of neural tissue using fiber-coupled lasers of various wavelengths. Blue and yellow lasers are broadly employed for optogenetic experiments, but due to poor penetration of these light frequencies in the brain, high laser power and/or fibers of high numerical aperture are often used to achieve functional stimulation of deep brain regions (Adamantidis et al., 2014; Adelsberger et al., 2014). Hence, brain tissue is irradiated with high light power that can be deleterious to neuronal function, but surprisingly little attention has been paid on the effects of light stimulation itself in optogenetic experiments. Absorbed light generates heat that can lead to permanent tissue damage. Additionally, neuronal excitability is acutely affected by temperature through the changes in Nernst equilibrium potential and by altering the gating properties of ion channels (Andersen and Moser, 1995; Kim and Connors, 2012). Thus, light alone can generate effects in neuronal function that are unrelated to the genuine 'optogenetic effect'. In modeling studies, an empirical factor ($Q_{10}$) is used to multiply rate constants to add temperature dependence to the classical Hodgkin and Huxley formalism (Fitzhugh, 1966).

Fiber optics delivered light in biological tissues is partially reflected at the fiber-tissue interface and partially transmitted through the tissue. A previous study (Stujenske et al., 2015) demonstrates that light emitted into the brain through fiber optic delivery is sufficient to increase local temperature and cortical firing rates of single neurons during optogenetics experiments. They also show that *in vivo* temperature recordings validate model predictions of heat induction. They provide an optogenetics MATLAB package for predicting light and heat spread in human brain tissue. On the other hand, the study of Arias-Gil and colleagues (Arias-Gil et al., 2016) uses thermal imaging to directly measure temperature rises at the surface of live mouse brains during laser illumination, with wavelengths and intensities typically used for optogenetics. They use a simple logarithmic model to validate their empirical model by predicting the temperature rise caused by pulsed stimulation paradigms.

The absorbed light is converted to heat, radiated in the form of fluorescence and/or consumed in photobiochemical reactions. The time-dependent heat production in brain tissue can be described by the bio-heat equation (Pennes, 1948), in which changes in tissue temperature can be calculated in time and space. These equations can also account for the buffering of temperature by blood perfusion. Furthermore, laser radiation increases stored energy that results in the diffusion of heat away from the irradiated area in proportion to the temperature gradients generated within the tissue (Welch and Van Gemert, 2011). Therefore, the conclusion drawn from optogenetic experiments may be hindered if the direct heat effect of light stimulation is not accounted for.

In this work, we model the optothermal effect in mice brain tissue produced by visible light laser sources (with a Gaussian profile) in both continuous and pulsed modes (Aravanis et al., 2007; Bernstein et al., 2008) to understand how heat can affect the transfer function of single neurons and how it can alter their response to photocurrents. We first approach the time/space numerical solution of Pennes' bio-heat equation comprising the effects of blood perfusion and metabolism with the finite element method (FEM) (Zimmerman, 2004). We then simulate the effect of varying heat in two single neuron models (Wang and Buzsáki, 1996; Rothman and Manis, 2003) that include a voltage and light-dependent current based on the channelrhodopsin-2 dynamics (Williams et al., 2013) to demonstrate that heat itself can considerably alter neuronal dynamics.

## METHODS

### Absorption

Absorption is a process involving the extraction of energy from light by a molecular species. It is important in diagnostic and therapeutic applications in biomedical photonics. The concept of the cross section is used for absorption, where the power absorbed is part of the incident intensity. Therefore, for a given absorber, the absorption cross-section, $\sigma_a$, can be defined as (Welch and Van Gemert, 2011; Vo-Dinh, 2014):

$$\sigma_a(\hat{a}) = \frac{P_a}{I_w}, \tag{1}$$

where, $\hat{a}$ is the propagation direction of the plane wave relative to the absorber, $P_a$ is the absorbed power, and $I_w$ is the intensity of the wave. Therefore, a medium with absorbing particles can be characterized by the absorption coefficient, $\mu_a$:

$$\mu_a = \rho_a \sigma_a, \tag{2}$$

where, $\rho_a$ represents the numeric density ($m^{-3}$) of the absorbers. Similar equations are found in the literature to explain the scattering phenomenon (Welch and Van Gemert, 2011; Vo-Dinh, 2014).

### Refraction

The relation between the angle of incidence, $\theta_1$, and the angle of refraction, $\theta_2$, for the transmitted light is given by Snell's law (Balanis, 2012; Peatross and Ware, 2015):

$$sin(\theta_2) = \frac{n_1}{n_2} sin(\theta_1). \tag{3}$$

Similarly, the relation between the incident wavelength (medium 1) and the refracted wavelength (medium 2) can be obtained by (Vo-Dinh, 2014):

$$\lambda_2 = \frac{n_1}{n_2}\lambda_1 \,. \tag{4}$$

## Photon Flux

Since light frequency does not depend on the refractive index, the photon energy is always the same as in a vacuum, according to (Welch and Van Gemert, 2011; Vo-Dinh, 2014):

$$E = hf \,, \tag{5}$$

where, $h = 6.626 \cdot 10^{-34}\ J \cdot s$ is Planck's constant and $f$ is the photon frequency ($Hz$).

Photon flux in a laser light beam is defined as the total number of photons crossing a particular section of the light beam, per unit area and per unit time (Svelto and Hanna, 2010). The number of photons emitted per second is given by:

$$N_p/s = P\frac{\lambda}{hc} \,, \tag{6}$$

in which, $P$ is the laser power. Then, the photon flux, $\phi_p$, can be obtained as a function of the cross section area ($A, m^2$) of the light beam as well as the intensity ($I, W/m^2$) of the light beam, according to (Svelto and Hanna, 2010):

$$\phi_p = \frac{P}{A}\frac{\lambda}{hc} = I\frac{\lambda}{hc} \,. \tag{7}$$

## Gaussian Laser Beam

Assuming that a laser beam in the $z$ direction attenuates exponentially with the distance $d$ in the tissue (Welch and Van Gemert, 2011), the irradiance can be defined as the radiant energy flux incident on the point of the surface, divided by the area of the surface. Many laser sources emit beams that approximate a Gaussian profile, in which case the propagation mode of the beam is the fundamental transverse electromagnetic mode ($TEM_{00}$) (Balanis, 2012; Sadiku, 2014).

Gaussian functions can assume multidimensional forms by composing the exponential function with a concave quadratic function (Weisstein, 2015). A particular example of a two-dimensional Gaussian function, in the $x - y$ plane, is:

$$f(x,y) = A\exp\left[-\left(\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2}\right)\right] . \tag{8}$$

Considering a bell curve shape for the Gaussian function, the parameter $A$ is the maximum amplitude of the curve, $x_0$ and $y_0$ are the center position of the curve in $x$ and $y$ axis, and $\sigma_x$ and $\sigma_y$ are the $x$ and $y$ spreads or standard deviations of the Gaussian curve.

## Light Propagation in Brain Tissue

*In vitro* and *in vivo* optogenetic experiments commonly use a relatively simple setup that consists of laser sources coupled to optical fibers to deliver light to a region of interest (ROI) in the tissue, in an accurate and efficient manner. *In vivo* experiments in deep regions of the brain, for example, also require a stereotactic surgery to position the tip of the optical fiber in the ROI into the brain (Zhang et al., 2015). Depending on the distance from the fiber tip and the optical properties of the surrounding tissue, the emitted light can propagate with uneven intensity.

The transmittance, $T$, is the relationship between the light intensity measured in the tissue at a distance $d$, and the light intensity measured without tissue, $\frac{I(d)}{I(d=0)}$, considering both scattering and absorption effects, and is given by (Vo-Dinh, 2014):

$$T = \frac{b}{a\sinh(bd\mu_s) + b\cosh(bd\mu_s)} \,, \tag{9}$$

in which, $\mu_s$ is the scattering coefficient and can be given in $mm^{-1}$ (Aravanis et al., 2007; Bernstein et al., 2008), $d$ is the distance in the brain tissue ($mm$), and $a$ and $b$ are given by (Vo-Dinh, 2014):

$$a = 1 + \frac{\mu_a}{\mu_s} \,, \tag{10}$$

$$b = \sqrt{a^2 - 1} \,. \tag{11}$$

here, $\mu_a$ can also be given in $mm^{-1}$ (Aravanis et al., 2007; Bernstein et al., 2008).

The light intensity can be estimated by the product between the transmittance $T$ and the geometric loss $g_{loss}$ due to light spreading in the tissue. The geometric loss is obtained by the decrease in light intensity due to the conical shape observed from the fiber tip ($d = 0$) to a certain distance $d$ in the tissue. The divergence angle, $\theta_{div}$, for a multimode fiber is given by (Aravanis et al., 2007):

$$\theta_{div} = \sin^{-1}\left(\frac{NA_{fib}}{n_t}\right) \,, \tag{12}$$

where, $n_t$ is the refractive index of the tissue and $NA_{fib}$ is the numerical aperture of the optical fiber. Considering the conservation of energy, we can calculate the geometric loss, $g_{loss}$, to a given distance, $d$, in the tissue as (Aravanis et al., 2007):

$$g_{loss} = \frac{\rho^2}{(d+\rho)^2} \,, \tag{13}$$

with,

$$\rho = r\sqrt{\left(\frac{n_t}{NA_{fib}}\right)^2 - 1} \,, \tag{14}$$

in which, $r$ is the fiber core radius. In this way, the expression for the normalized light intensity, $I_N$ ($mW/mm^2$), considering scattering, absorption and geometric loss is given by:

$$I_N = \frac{I(d)}{I(d=0)} = g_{loss} \cdot T \,. \tag{15}$$

**TABLE 1 |** Parameters used in scattering and absorption simulations.

| Parameters | Values | References |
|---|---|---|
| Fiber core radius ($r$) | 0.2 $mm$ | dat, 2015 |
| Fiber numerical aperture ($NA$) | 0.48 | dat, 2015 |
| Fiber core refractive index ($n_1$) | Blue: 1.4644 Yellow: 1.4587 | dat, 2015 |
| Scattering coefficient ($\mu_s$) | Blue: 10.0 $mm^{-1}$ Yellow: 9.0 $mm^{-1}$ | Bernstein et al., 2008 |
| Absorption coefficient ($\mu_a$) | Blue: 0.070 $mm^{-1}$ Yellow: 0.027 $mm^{-1}$ | Bernstein et al., 2008 |
| Laser input power ($P$) | 20 $mW$ | |
| Laser coupling fraction ($\eta$) | 1 or 100% | |

We can consider $I(d = 0)$ as the light intensity at the fiber tip that can be obtained in $mW/mm^2$ simply by:

$$I(d = 0) = \frac{P}{A\eta}, \tag{16}$$

where, $P$ is the power emitted by the light source ($mW$), $A = \pi r^2$ is the area of the optical fiber ($mm^2$), and $\eta$ is the coupling efficiency between the optical fiber and the light source (dimensionless). We chose $\eta = 1$ for all the scattering and absorption simulations.

Finally, the light intensity ($mW/mm^2$) at a region of interest in the tissue, assuming a distance $d$ ($mm$) from the fiber tip, is given by:

$$I(d) = I(d = 0) \cdot I_N. \tag{17}$$

We used MATLAB commercial software to simulate scattering and absorption characteristics in mice brain tissue. **Table 1** shows the parameters and respective values used for these simulations.

## Heat Transfer in Mice Brain Tissue

Heat transfer is a known physical problem already modeled in many areas of knowledge (Ahmed et al., 2019; Taheripour et al., 2019). For biology, heat is inevitable when light propagates and is absorbed by biological tissues.

The traditional bio-heat equation describes the change in tissue temperature over time that can be expressed at a distance $d$ in the tissue. Furthermore, blood perfusion occurs in living tissues, and the passage of blood modifies the heat transfer in tissues. Pennes (1948) has established a simplified bio-heat transfer model to describe heat transfer in tissue by considering the effects of blood perfusion, $\omega_b$, and metabolism, $H_m$ (Elwassif et al., 2006; Vo-Dinh, 2014):

$$\rho C_p \frac{\partial T}{\partial t} = \nabla(k\nabla T) - \rho_b \omega_b C_b (T - T_b) + H_s + H_m, \tag{18}$$

where, $\rho$ is the tissue density ($kg/m^3$), $C_p$ is the specific heat of the tissue ($J/kg^{\circ}C$), $k$ is the thermal conductivity of the tissue ($W/m^{\circ}C$), $\rho_b$ is the blood density ($kg/m^3$), $\omega_b$ is the blood perfusion ($1/s$), $C_b$ is the specific heat of the blood ($J/kg^{\circ}C$), $T$ is the temperature of the tissue ($^{\circ}C$), $T_b$ is the blood temperature ($^{\circ}C$), $H_s$ is the heat source due to photon absorption ($W/m^3$), and $H_m$ is the term that represents heat generated by metabolism ($W/m^3$). Equation (18) is almost linear for small temperature changes, therefore, it is expected that temperature rises are approximately proportional to the energy input (that is, duty cycle).

The interaction between metabolic heat generation and blood perfusion was investigated, and it was proved that the temperature increases during Deep Brain Stimulation (DBS). Other environmental interactions that can affect the stored energy include radiation and convection from the sample surface, the loss of vapor phase water from the sample, and convection with blood that is perfused through the vascular network from arterial and venous sources. This network has a very specific geometry that is unique to a tissue or organ and can affect significantly the capability to exchange heat with the tissue in which it is embedded (Welch and Van Gemert, 2011).

Additionally, thermal boundary interactions occur over the surface area with the environment and are often characterized as convective and irradiative processes. Laser irradiation process increases the stored energy from its initial state and, as a result, it diffuses the heat away from the irradiated area in proportion to the temperature gradients developed in the tissue. A quantitative characterization of the formation of these gradients and the heat flow that they drive are the focus of heat transfer analysis (Welch and Van Gemert, 2011).

In the case of convective boundary conditions, heat transfer occurs when a solid substrate is in contact with a fluid at a different temperature (Welch and Van Gemert, 2011). The magnitude of the heat exchange can be calculated according to Newton's law of cooling, that describes the convective flow, $H_{conv}$ ($W/m^2$), at the surface in terms of the convective heat transfer coefficient, $h$ ($W/m^{2\circ}$C) and the temperatures of the sample, $T$, and the external environment, $T_{ext}$, in $^{\circ}$C:

$$H_{conv} = h(T - T_{ext}). \tag{19}$$

We consider the geometry and shape of the boundary layer region of the fluid in which convection occurs, to calculate the free convective flow. Convective effects are hard to estimate once different process characteristics must be considered depending on the convective transport problem. Typical values of $h$ for free convection in liquids are in the range of 20–1,000 ($W/m^{2\circ}C$) (Welch and Van Gemert, 2011). It is important to choose small values of $h$, such as 25 $W/m^{2\circ}C$, so that the temperature variations between the environment and the sample are properly evidenced.

Heating generated within the biological material is governed by the following expression (Elwassif et al., 2006):

$$H(x, y, z) = P(1 - R)\frac{\mu_a}{\pi\sigma_x\sigma_y} \exp\left[-\left(\frac{(x - x_0)^2}{2\sigma_x^2}\right) + \frac{(y - y_0)^2}{2\sigma_y^2}\right] \exp(-\mu_a z), \tag{20}$$

in which, the first exponential function represents the two-dimensional Gaussian distribution in $x - y$ plane, in accordance

**TABLE 2 |** Parameters and material properties used in heat transfer simulations.

| Parameters | Values | References |
|---|---|---|
| Refractive index of the tissue ($n_t$) | 1.36 (gray matter) | Vo-Dinh, 2014 |
| Specific heat of the tissue ($C_p$) | 3650 $J/kg°C$ | Elwassif et al., 2006 |
| Density of the tissue ($\rho$) | 1040 $kg/m^3$ | Elwassif et al., 2006 |
| Thermal conductivity of the tissue ($k$) | 0.527 $W/m°$C | Elwassif et al., 2006 |
| Metabolic heat ($H_m$) | 13698 $W/m^3$ | Elwassif et al., 2006 |
| Blood density ($\rho_b$) | 1057 $kg/m^3$ | Elwassif et al., 2006 |
| Blood perfusion ($\omega_b$) | 0.012 $1/s$ | Elwassif et al., 2006 |
| Specific heat of the blood ($C_b$) | 3600 $J/kg°$C | Elwassif et al., 2006 |
| Temperature of the tissue ($T$) | 37°C | Elwassif et al., 2006 |
| Blood temperature ($T_b$) | 36.7°C | Elwassif et al., 2006 |
| Heat transfer coefficient ($h$) | 25 $W/m^2$ °C | Welch and Van Gemert, 2011 |
| Standard deviations in $x$ and $y$ axis ($\sigma_x, \sigma_y$) | 0.5 | |
| Reflection coefficient ($R$) | 0 | |

to Equation (8). The second exponential function represents the exponential decay due to absorption (Yang and Miklavcic, 2005).

Some considerations in using Equation (20) are: the reflection ($R$) and absorption coefficients are assumed to be constant; the sample is assumed to have a planar surface aligned with the $xy$-plane of the global coordinate system and whose top matches $z = 0$ (distance at the fiber tip); the center of the beam can be easily shifted by changing $x_0$ and $y_0$; the beam width can be easily controlled by the standard deviation parameters $\sigma_x$ and $\sigma_y$. We assumed $R = 0$ and $\sigma_x = \sigma_y = 0.5$ for the analysis of heat transfer performed in this work.

Heat transfer simulations were accomplished using the computational modeling software, COMSOL Multiphysics 4.4, that allows numerical solutions for partial differential equations based on the Finite Element Method (FEM) (Zimmerman, 2004). Laser heating was simulated considering two stationary conditions: continuous mode and pulsed mode. We used biological material with mice brain tissue characteristics (gray matter). The material properties were assumed to be constant and are shown in **Table 2**.

## Channelrhodopsin-2 and Neuron Models

We first modeled the effect of temperature alone in a pyramidal cell model and in a network of basket cells known to generate gamma oscillations. We have implemented a single compartment CA1 neuron model described by Migliore (Migliore, 1996). He has implemented a multicompartment model in his original work, but here we only employ the soma with an inactivating sodium conductance (max. 30 $nS$), a delayed rectifier $K^+$ conductance (max. 10 $nS$), conductance from an $M$ current (max. 0.6 $nS$) and from an $H$ current (max. 0.3 $nS$). Kinetics for all currents were download from ModelDB (https://senselab.med.yale.edu/modeldb/, Accession:2937).

In addition, we have used the same $Q_{10}$ values for all voltage-gated currents as the original publication (Wang and Buzsáki, 1996). Temperature values from the heat transfer simulation were fed to the neuron model by a "look up time/temperature table" where each rounded ms value corresponded to a single temperature value. Simulations were run for 90 s (30 s for stabilization with constant temperature and 60 s with variable temperature). The model was solved in MATLAB using the built-in solver "ode23". The interneuron network gamma model was simulated using Neuron with no changing in parameters from the model available from ModelDB (Accession:26997) exception by setting the temperature to 37 or 39°C. These simulations were run for 500 ms with a constant temperature. Note that the original study of Wang and Buzsáki did not account for temperature; however, the uploaded model in ModelDB includes $Q_{10}$ for kinetic variables (Wang and Buzsáki, 1996).

Power spectrum density analysis and cross-correlation of action potentials were calculated from spike trains transformed in a series of 0 s (no spike) and 1 s (spike) with 0.1 $ms$-precision (Hilscher et al., 2013). Power spectral density analysis of binary spike series was performed using Welch's method (pwelch command in MATLAB). Cross-correlograms (CCGs) were calculated as described previously (Hilscher et al., 2013) and then smoothed by a moving average filter with a span of 10 ms (Hilscher et al., 2013). Cross-correlations over a lag range of $\pm$0.1 s. Synchrony index (SI) is defined as the maximum value of the CCG.

We have implemented the channelrhodopsin-2 empirical model (Williams et al., 2013) in two single neuron models to test the interaction of temperature and optocurrents: a single basket cell from Wang and Buzsaki network model (Wang and Buzsáki, 1996) and an anteroventral cochlear nucleus bushy cell model (Rothman and Manis, 2003). The equations and parameters from the neuron models can be found in the original publications (Wang and Buzsáki, 1996; Rothman and Manis, 2003) and equations and parameters for channelrhodopsin optocurrents are found in (Williams et al., 2013). All models were implemented in MATLAB (Mathworks), and the codes can be downloaded from https://github.com/cineguerrilha/Neurodynamics/tree/master/Cell_Models.

# RESULTS

In this work, we first simulated the light propagation and absorption in the brain of mice in a typical optogenetic setup. **Figure 1A** shows a diode pumped solid state - DPSS laser source coupled to a multimode optical fiber that transmits light directly to the region where the brain implant was performed (Zhang et al., 2015).

Subsequently, we simulated the effect of heat in single neurons and networks. We have also examined the additive effect of heat and light in simulations that included a channelrhodopsin-2 model (Williams et al., 2013). The bio-heat transfer was solved numerically using Pennes' equation with the finite element method and temporal changes in temperature at a given point in space were applied to a single compartment neuron model (with Hodgkin and Huxley formalism).

We first simulated beam geometry and light spreading. A DPSS laser emits a Gaussian beam that the propagation

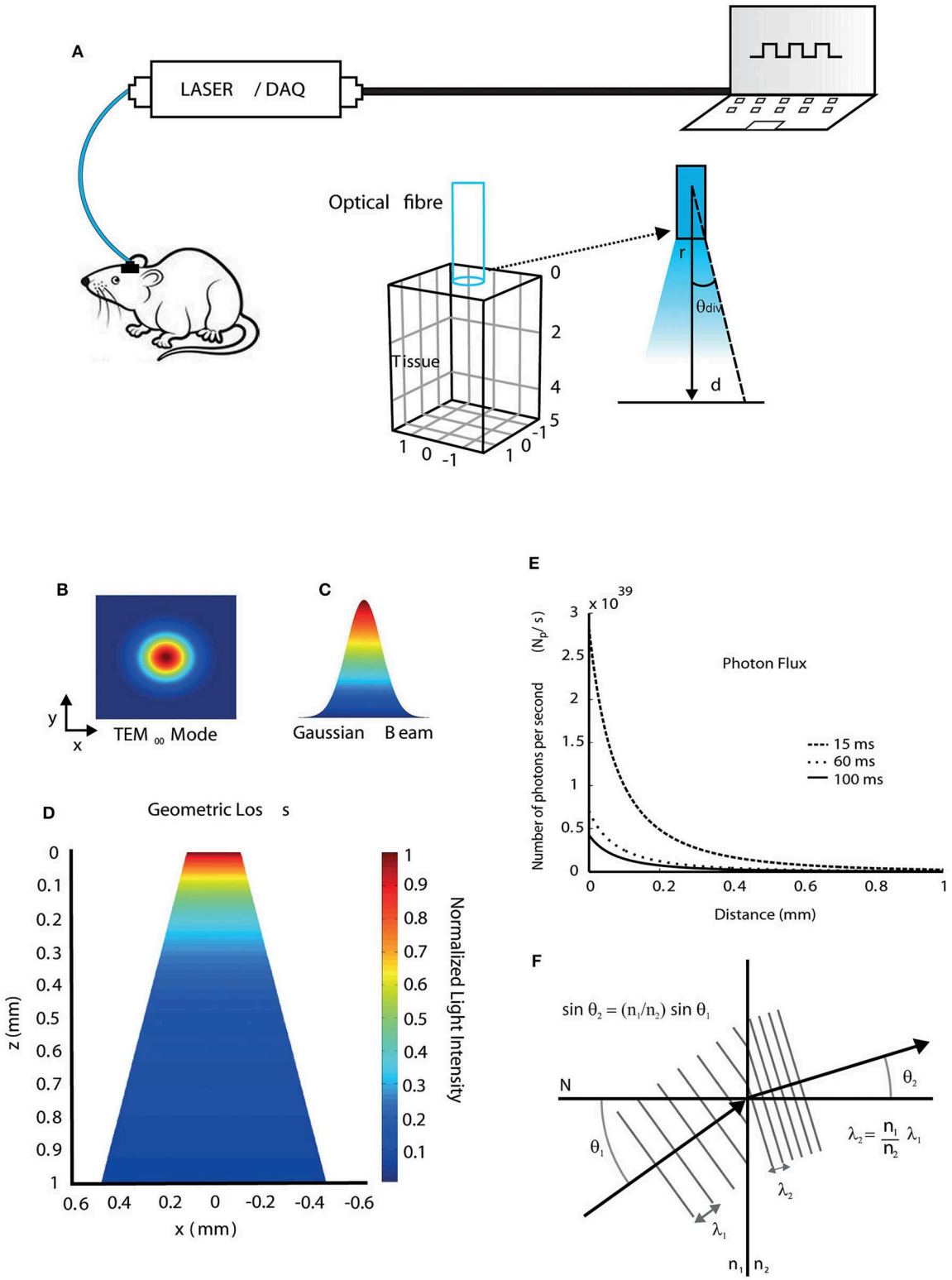


**FIGURE 1 |** Light propagation properties when interacting with brain tissue. **(A)** Diagram showing a typical optic stimulation setup used in freely moving animals. The setup consists of a computer, a data acquisition (DAQ) board, and a laser source coupled to a fiber transmitting light to a target region into the mouse brain at a divergence angle ($\theta_{div}$) calculated using Equation (12). **(B)** Transversal electromagnetic fundamental propagation mode ($TEM_{00}$) of the laser source. **(C)** Gaussian beam shape. **(D)** 2D view of the geometric loss due to light spreading in the tissue (conical shape) at a certain distance from the fiber tip. **(E)** The flux of irradiated photons as a function of distance during 15, 60, and 100 *ms* light pulses considering a region of unit area. **(F)** Wavelength shift during light propagation through different media.

mode is the fundamental transversal electromagnetic ($TEM_{00}$) (**Figures 1B,C** and Equation 8). **Figure 1D** shows the normalized geometric loss due to light spreading in *z-x* plane within the tissue as a function of the distance from the fiber tip in *z* direction. The divergence angle is determined by the optical fiber numerical aperture, according to Equation (12). After light power at a given point is calculated, photon flux (number of irradiated photons per unit time and per unit area) at that point can be obtained by Equation (7). Photon flux can then be correlated to photocurrents in channelrhodopsin models (Foutz et al., 2012). Photon flux simulations are shown in **Figure 1E**, in which, a 20 $mW$, 473 nm laser is pulsed with durations of 15, 60, and 100 ms. The different pulse durations were chosen to illustrate that the pulse width changes alter the amount of photons passing through a surface. Light speed is altered during propagation because of the difference of refractive indices and their dependence with wavelength. Consequently, the wavelength can change during propagation and this effect is not only observed in the interface between fiber and tissue, but also within the tissue, due to its anisotropic refractive indexes between different brain regions. The wavelength change between two different media, which is calculated using Snell's law (Equation 3), is illustrated in **Figure 1F**. Assuming that light propagates from an optical fiber (medium 1) to the tissue (medium 2), where N is a perpendicular line to the surface of separation between the two media, and considering $n_{1b}$ = 1.4644 as the refractive index of the fiber core at 473 nm, $n_{1y}$ = 1.4587 the refractive index of the fiber core at 593 nm, and $n_2$ = 1.36 the refractive index of the tissue (mouse brain, gray matter), the wavelength shifts for blue (473 nm) and yellow (593 nm) lights due to refraction are 36 nm and 43 nm, respectively, according to Equation (4). Yet small, wavelength shifts have to be considered specially in modeling studies as there is an obvious relationship between wavelength and light absorption in both light-sensitive ion channels and fluorescent proteins (Zhang et al., 2015), even if the photon energy remains the same, once small changes in the wavelength affect the response of the light-sensitive ion channels and fluorescent proteins.

We then used the Kubelka-Munk model to calculate light intensity vs. distance considering absorption (Mobley and Vo-Dinh, 2003). Light absorption by the tissue has no direct relation to the production of photocurrents by channelrhodopsin; however, absorption produces heat, a side effect of light stimulation (Shapiro et al., 2012). Light absorption also changes (although slightly) the relation between light intensity and tissue depth (**Figure 2A**). Assuming a threshold of 10 $mW/mm^2$ (green line), which is a sound intensity value when stimulating a large group of stimulated cells (Bernstein et al., 2008), the depth for channelrhodopsin-2 activation is 0.39 $mm$ (473 $nm$) and for halorhodopsin activation is 0.42 $mm$ (593 $nm$). **Figure 2B** shows the transmittance (Equation 9) as a function of distance $d$, considering both scattering and absorption effects. These simulations indicate that only cells and neurites at the vicinity of the fiber are affected by light stimulation and are in agreement with a previous study (Stujenske et al., 2015).

We next computed the production of heat in the tissue caused by light absorption using FEM. For heat transmission simulations, we used a rectangular prism of dimensions equal to 3.5 × 3.5 × 5 ($mm^3$) representing a mouse brain tissue. Optogenetic experiments often use specific stimulation protocols with yellow light to activate halorhodopsin and blue light to activate channelrhodopsin (Cardin et al., 2009; Mikulovic et al., 2016). We, therefore, simulated the interaction between the mouse brain and the yellow light radiation (593 nm wavelength), with the laser source operating in continuous mode, while the blue light radiation (473 nm wavelength) laser source operating in pulsed mode.

Temperature changes at a distance $d$ = 10 $\mu m$ from the fiber tip caused by continuous light radiation (593 $nm$) as a function of time are shown in **Figure 3A**. We simulated heat transfer due to continuous yellow light for different values of power emitted by the laser source: 1, 10, 20, 30, and 40 $mW$. According to **Figure 3A**, during the first 5 s, the rate of temperature variation is higher. After that, the temperature continues to increase more slowly moving toward the steady state condition. For light power up to 10 $mW$, temperature increases about 0.5°C. For 20, 30, and 40 $mW$, the increase in temperature after 1 min of radiation is between 1 and 2°C. **Figure 3B** shows a temperature distribution in 3D view, 2D top view (*x-y*), and 2D slice center view (*z-x*, constant *y*), for continuous yellow light radiation (20 $mW$ and 60 s, indicated by the red asterisk shown in **Figure 3A** and pulsed blue light radiation (473 nm), 12 Hz and 18% of duty cycle-percentage of a period in which the light is turned on (black asterisk indicated in **Figure 3C**). We have also computed temperature changes for 20 $mW$ blue light, at 60 s and 10 $\mu m$ from the fiber tip, for frequencies varying from 1 to 40 $Hz$ with duty cycles varying from 1% to 100% (**Figure 3C**). These results show that lower duty cycles minimize temperature changes by light stimulation.

Currents produced by voltage-gated ion channels are directly influenced by temperature. It is known for decades that channel opening and closing are generally faster in higher temperatures and conductance/voltage relationship and ion reversal potential are also be affected by temperature (Fitzhugh, 1966). To illustrate the effect of temperature in firing, we used a basket cell model (Wang and Buzsáki, 1996). For these simulations, we used two temperatures (37°C and 39°C the latter can be quickly produced by a pulsed laser at 40 $Hz$ and 90% duty cycle and at 10 $\mu$m distance from the center of the fiber tip **Figure 4**). In the model implemented here, action potentials become smaller and briefer (**Figures 4A,B**). Spontaneous firing frequency of the neuron used in this simulation also increases (**Figure 4C**). Optogenetics has been used to study the mechanisms behind neuronal synchrony and brain rhythm generation (Cardin et al., 2009). Hence, we further investigated the effect of heat generated by light stimulation itself (rather than photocurrents in channelrhodopsin-expressing neurons) in a network model comprised solely by basket cells that synchronize in gamma frequency (Wang and Buzsáki, 1996). The model is composed of 100 interconnected fast spiking interneurons (same as in **Figure 4**) (Wang and Buzsáki, 1996). In the Wang and Buzsákis model (Wang and Buzsáki, 1996), neurons in the network take around 200–300 ms to fire in gamma frequency from a relatively asynchronous onset (**Figures 4A,D**). If the temperature

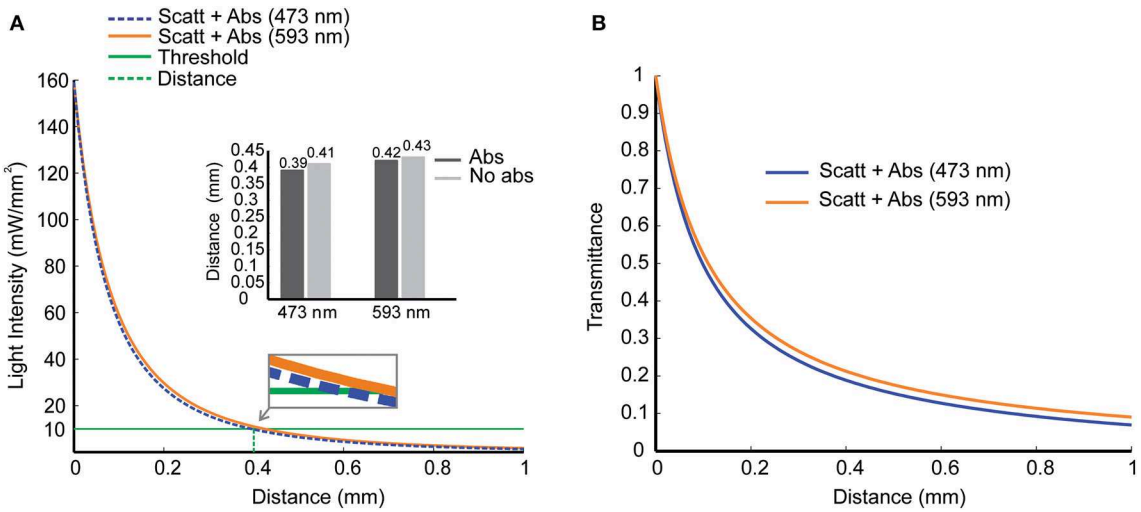


**FIGURE 2 |** Scattering and absorption effects as light propagate into mouse brain tissue. **(A)** Light intensity vs. penetration distance for 473 *nm* (blue) and 593 *nm* (yellow) wavelengths. At a distance $d$ = 0.4 *mm* from the fiber tip (dashed green line), a reference value for light intensity of 10 mW/mm$^2$ (solid green line) was chosen for blue and yellow light with (solid lines) and without (dashed lines) absorption. Inset. Distance in which light decays to 10 mW/mm$^2$ in simulations with and without absorption. **(B)** Transmittance vs. penetration distance for blue and yellow lights, including scattering and absorption effects.

is raised by 2°C the network is synchronized in less than 50 ms (**Figures 4A,D**) from the onset of simulation. Firing frequency of the interneurons in the network also increased by raising the temperature in 2°C (**Figure 4C**). This changing in frequency caused a shift in the peak of 'gamma oscillation' in the power spectrum (**Figure 4C**). Hence, heat itself can theoretically facilitate the generation of oscillations and/or alter their frequency.

We further assess the effect of raising the temperature in neuronal synchronization using previously described synchrony metrics (Leao et al., 2005; Hilscher et al., 2013). Autocorrelation histograms of all 100 neurons in the model are shown in **Figure 5A** for 37°C and at 39°C. Heating the network model caused neurons to fire at greater rhythmicity (**Figure 5A**). In addition, cross-correlogram also showed greater synchrony when simulations were executed at 39°C (compared to 37°C). This increase in synchrony is reflected by a significant rise in the synchronization coefficient (**Figure 5B**). The mean synchronization index (SI) for all possible neuron pair combinations (9,900 pairs) was equal to 0.16 for 37°C and 0.22 for 39°C. These results show that heating can, not exclusively, change the frequency of brain oscillations but also alter the coordination and synchrony of neuronal firing.

We then combine temperature and irradiation in modeled neurons that also contained a channelrhodopsin-2-driven photocurrents (Wang and Buzsáki, 1996; Williams et al., 2013). We have used two distinct cell models to illustrate the interaction of channelrhodopsin photocurrents with other ionic currents in the neuron. The basket cell shows high-frequency firing that increases proportionally to the injected current (Martina et al., 1998) and a bushy cell of the dorsal cochlear nucleus that show single action potentials in response to continuously injected currents (Leao et al., 2006). At 1 *mW* power, the basket cell model fired action potentials at the beginning of each pulse whether at 37°C or 39°C (**Figure 6A**). However, the bushy cell model only fired APs at physiological temperature (**Figure 6A**). The tissue reaches 39°C quickly for duty 50% or 90% duty cycles, but the temperature only rises mildly for 10% duty cycle (**Figure 6B**). Nevertheless, even at 10% duty cycle, bushy cell light-elicited AP amplitude is still affected by the small increase in temperature (**Figures 6C,D**). Taken together, this data suggests that temperature can alter the efficiency of photocurrents in eliciting APs. Most importantly, the effect of temperature and light stimulation interaction in the membrane is greatly dependent on native voltage-gated channels.

## DISCUSSION

In the context of optogenetics, the first study that addressed the interaction of light emanating from an optical fiber with brain tissue omitted absorption (Aravanis et al., 2007). Aravanis and colleagues argued that the effect of light (400–900 *nm*) absorption could be neglected when simulating light transmission in the brain (Aravanis et al., 2007). However, while absorption does not affect significantly the spatial computation of light intensity (as most of the loss occurs through scattering), it is through absorption that heat is generated. Also, we opt to use the simpler Kubelka-Munk model for light transmission instead of a more accurate Monte Carlo method as the former generates values that approximate empirical results for short distances ($\sim$ 1 *mm*) (Aravanis et al., 2007; Džimbeg-Malčić et al., 2011).

Our bio-heat transfer results corroborate with recent studies found in the literature (Stujenske et al., 2015; Arias-Gil et al., 2016). These authors were the first to explores heat generation by light in optogenetic experiments and compare simulations with empirical measurements. Our work, instead, explore

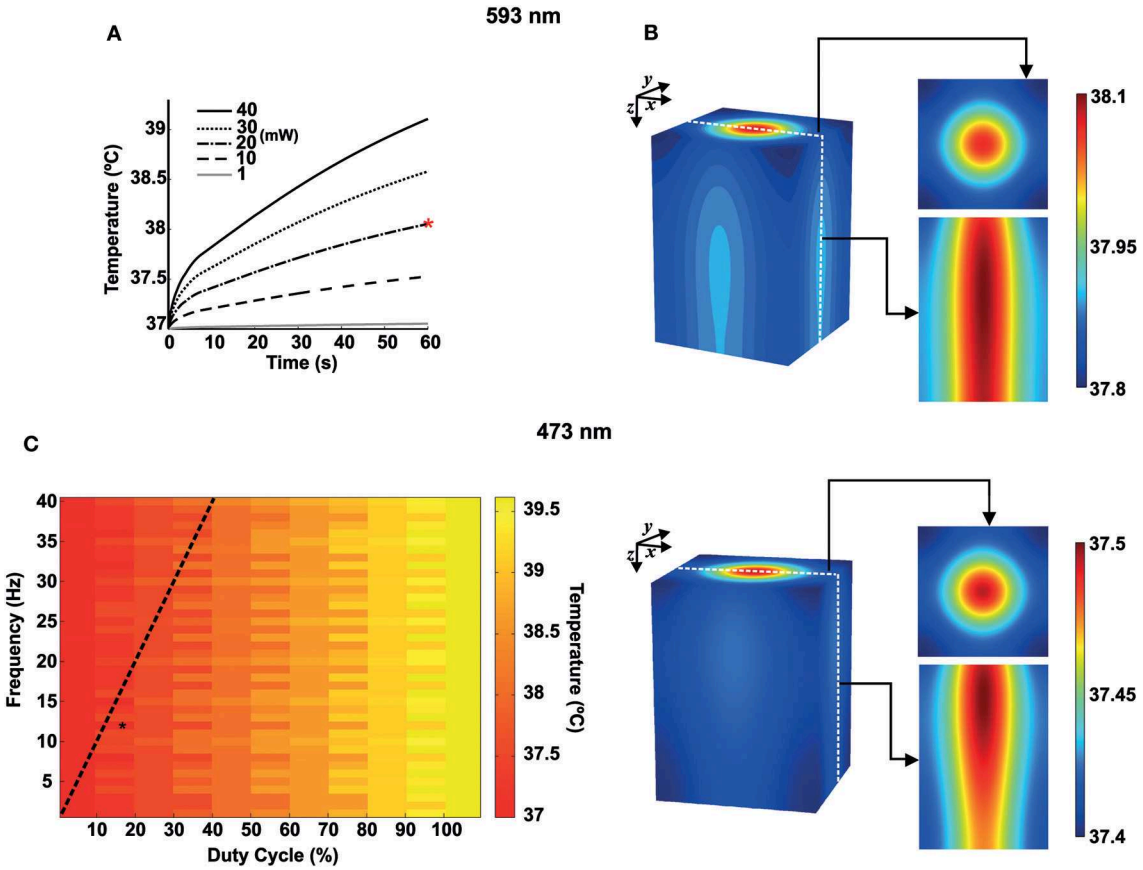


**FIGURE 3 |** Heat transfer simulations for blue and yellow light in mouse brain tissue. **(A)** Temperature variations for 593 *nm* wavelength as a function of time for 1, 10, 20, 30, and 40 *mW* of continuous radiation. The red asterisk indicates continuous yellow light radiation for 20 *mW* and 60 s. **(B)** Temperature distribution in space for 593 *nm* and 473 *nm*. Right. Top. 2D Gaussian beam (*x*-*y*) for the top view and with $z \rightarrow 0$. Bottom 2D slice view (*z*-*x*) of the temperature distribution. **(C)** Heat map for the temperature distribution (473 *nm*) as a function of frequency (1–40 *Hz*, bin size of 1 *Hz*) and duty cycle (1–100%, bin size of 10%) at 60 *s* of light radiation (10 $\mu m$ from the fiber tip). The black asterisk indicates pulsed blue light radiation, 12 *Hz* and 18% of duty cycle. The dashed black line shows a pulse width of 10 *ms*.

the effect of bio-heat transfer in neurons and networks, in particular, with a few differences compared to the study by Stujenske and colleagues (Stujenske et al., 2015). For instance, these authors used light absorption and scattering coefficients obtained from human brain tissue interpolated from different wavelengths while here we employ coefficients obtained from rodent brains in specific wavelengths used in optogenetic experiments (Bernstein et al., 2008; Stujenske et al., 2015). Besides, we have calculated temporo-spatial photon flux in brain tissue. Ultimately, photon flux determines the opening of channelrhodopsin pores, and these values could be directly used for simulation of channelrhodopsin activation (Zhang et al., 2015).

We used homogeneous absorption coefficients for a given wavelength, but it is clear from optical measurements that light is unevenly absorbed in the brain (Jacques, 2013). Thus, the temperature can also increase unevenly based on anisotropic absorption coefficients. Besides, blood vessels are not homogeneously distributed in all brain regions; therefore, spatial differences in temperature buffering will further complicate the network effect of heat generation by optical stimulation. In other words, the effect of the increase in temperature in optogenetic experiments will depend on the region, neuron type, and connections and can significantly affect neuronal processing. Minimizing stimulation time may help to prevent unwanted heat effects in neuronal function. In experiments where long stimulation times are desirable, step-function opsins may be the tool of choice for avoiding heat-related changes in firing and behavior.

The temperature effect in the gating of voltage-dependent channels is classically modeled by using an empirical factor ($Q_{10}$) to multiply rate constants (incorporating temperature dependence to the classical Hodgkin and Huxley formalism) (Thompson et al., 1985). In addition, ion reversal potentials in semipermeable membranes are directly proportional to temperature. We simulated the effect of a 2°C change in a classical model of interneuron network gamma (ING) oscillation (Wang and Buzsáki, 1996). The idea that gamma oscillation arises from the interaction of fast spiking interneurons originated from slice and modeling studies (Whittington et al., 1995; Wang and

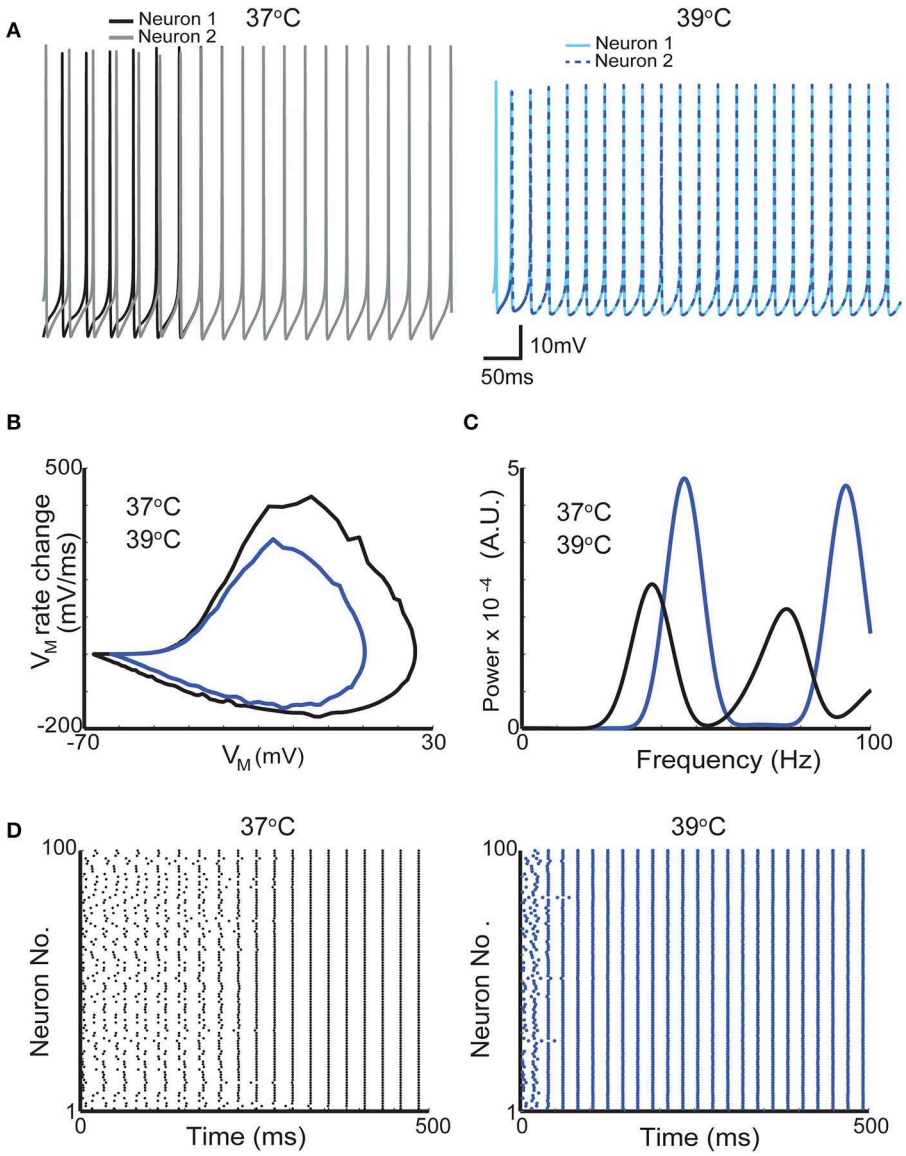


**FIGURE 4 |** A 2°C raise in temperature increases the firing frequency of neurons in a network model of gamma oscillations. **(A)** Membrane potential of two neurons from a network of 100-interneuron network when simulation was executed with temperatures of 37°C (gray and black traces left) and 39°C (blue and dashed dark blue right). **(B)** Phase plots from one action potential of one interneuron at 37°C and at 39°C (black and dark blue traces, respectively). **(C)** Mean firing power spectrum density (see section Methods) of the 100 interneurons in the network at 37°C and at 39°C (black and dark blue traces, respectively). **(D)** Scatter plots showing the action potential firing of the gamma network at 37°C (left) and at 39°C (right).

Buzsáki, 1996) and it was demonstrated by a highly influential optogenetics study (Cardin et al., 2009). Cardin and colleagues elicited gamma oscillation in the neocortex by rhythmical optical stimulation of cells expressing the enzyme Cre recombinase (and channelrhodopsin) in a Parvalbumin-Cre animal (Cardin et al., 2009). To generate gamma oscillations, the authors optically stimulated neurons at the same frequency as the recorded local field potential (Cardin et al., 2009). It is known that rhythmical stimulation is likely to interfere with the local field potential recording due to the optoelectric effect (Mikulovic et al., 2016). However, the effect of temperature caused by optical stimulation in network responses is largely unexplored. Parvalbumin is especially found in soma targeting fast spiking interneurons (but it is also found in several other types of interneurons) (Klausberger et al., 2005; Mikulovic et al., 2016). Using Wang and Buzsaki's model of ING (1996), we found that an increase of two degrees significantly organizes the inhibitory neuron network. At 39°C, firing in gamma can be observed in less than 50 ms from the simulation onset (when firing of individual neurons is random) while at 37°C, that network takes almost 5 times longer to organize its spikes at gamma frequency. Also, network firing frequency increases in several *Hz*. Changes in gamma oscillation

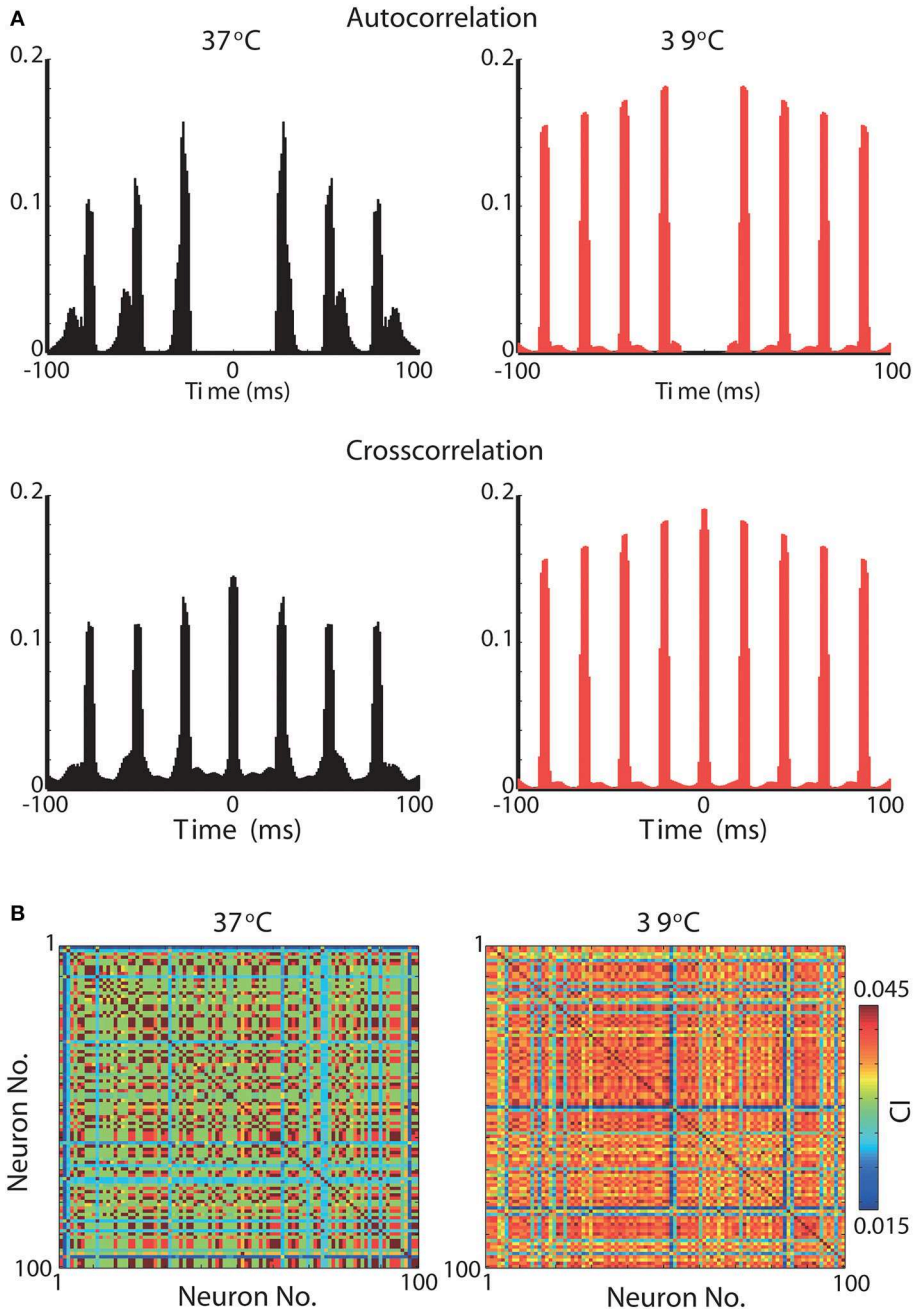


**FIGURE 5 |** Synchrony is greatly increased in a gamma oscillation network model by a 2°C raise in temperature. **(A)** Top, Normalized autocorrelograms of all 100 neurons in the network at 37°C (left) and at 39°C (right). Bottom, Normalized crosscorrelograms of all 100 neurons crosscorrelated with all 100 neurons in the network at 37°C (left) and at 39°C (right). **(B)** Peak normalized correlation index between all 100 neurons when simulations were performed at temperatures of 37°C (left) and 39°C (right).

frequency by temperature has been observed experimentally (Leao et al., 2009), and as the increase in temperature depends on the proximity of targeted neurons to the optical fiber, light stimulation could generate small networks that oscillate incoherently from non-heated networks and this effect is not directly associated to opsin expression.

Here, we show that different types of neurons can have very different responses to similar light pulses. There has been little concern in optogenetic experiments regarding native currents of neuronal populations of interest (Adamantidis et al., 2015). However, we show that native voltage-gated currents can have a huge impact on how neurons fire to light stimulation. For

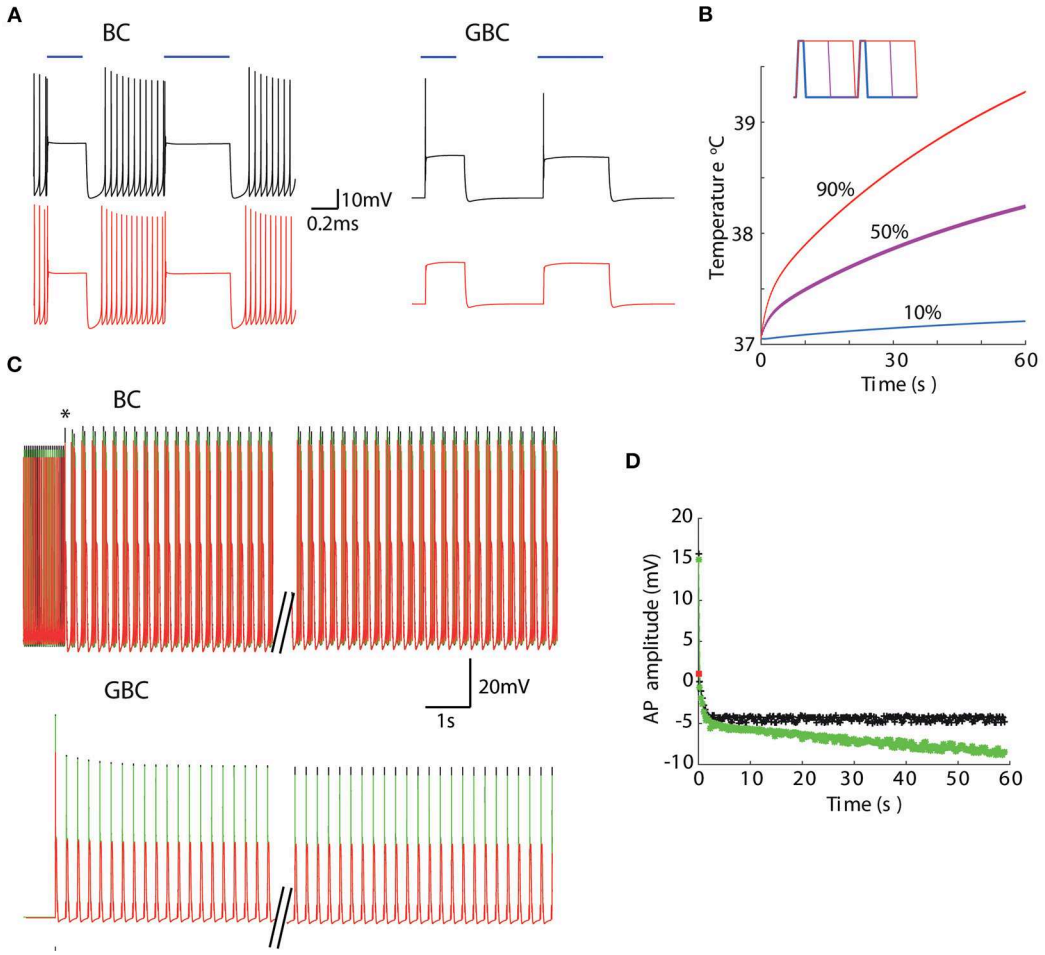


**FIGURE 6 |** Temperature changes caused by light absorption affects membrane response to photocurrents. **(A)** Membrane potential of a basket cell (BC) and a dorsal cochlear nucleus bushy cell (GBC) models to 10 *mW*-473 *nm* light pulses at 37°C (top) and 39°C (bottom). **(B)** Temperature at 10 $\mu m$for 4 Hz stimulation (20 *mW*) for 10% (blue), 50% (magenta) and 90% (red) duty cycles (inset shows 0.5 s pulses with the three different duty cycles). **(C)** BC and GBC responses for 10% duty cycle (4 *Hz*) light pulses with fixed temperatures (37°C black and 39°C red) and when temperature raises (green) in response to light pulses (black trace in **B**). **(D)** Action potential amplitude evolution in time of GBC model in response to light pulses in **(C)**. The red square is the amplitude of the single AP the GBC model fired when temperature was set to 39°C.

example, neurons that express strong low threshold $K+$ currents to avoid repetitive firing when currents are injected will only fire one to a couple of spikes independent of the duration of the light pulse (Leao et al., 2008). On the other hand, fast spike neurons expressing high-threshold $K+$ currents like basket cells (Martina et al., 1998) will respond, most likely, with multiple spikes after each light pulse. Neurons with strong inward currents activated by hyperpolarization (e.g., Ih) could also produce strong depolarizations (and action potentials) by activation of Ih rather than the reversal of $Cl-$ gradients (Leao et al., 2011; Adamantidis et al., 2015). It is important to note that the simple ChR2 model used here describes well the behavior of macroscopic photocurrents for short periods (that cover a large number of optogenetic experiments) (Williams et al., 2013). Hence, this ChR2 model could be added to specific cell models that are readily available in databases like the ModelDB (McDougal et al., 2017) for optimization of light protocol design.

Finally, temperature affects the transfer function of a given neuron according to the diversity of ion channels in it (Cao and Oertel, 2005). For that reason, while some neuron types increase spontaneous firing, other populations may become quiet when the temperature is changed (Kim and Connors, 2012). Most importantly, changes in temperature and native channels may hinder optogenetic stimulation. Our optogenetic simulations using the bushy cell model showed that light pulses are unable to elicit spikes when the cell is heated to 39°C. Bushy cells are known to express low threshold potassium channels (Kv1) (Rothman and Manis, 2003), and these channels prevent the firing of multiple APs in response to tonic currents (Couchman et al., 2011). Thus, accelerating the opening of Kv1 channels could prevent spike generation by photocurrents. However,

the interaction of channelrhodopsin photocurrents with native voltage-gated currents of a given cell is a subject largely explored, especially when changes in temperature caused by the light stimulation affects the gating dynamics of native channels. Future studies should assess the interaction of photocurrents with native voltage-gated currents and examine the effect of temperature.

## CONCLUSION

In this work, we have used the finite element method to address brain temperature changes caused by light stimulation in optogenetics and its effect in neuron firing. We found that temperature can increase by about 2.6°C in 1 min for blue light stimulation (20 mW of power, **Figure 3C**). A two-degree change in temperature, when applied to a model of a spontaneous firing neuron, caused a dramatic increase in firing frequency and change in action potential shape. Conversely, a 2°C-increase in temperature in a fast spiking interneuron network model of gamma oscillation produced a large increase in neuronal synchrony and oscillation frequency. Moreover, the effect of channelrhodopsin-driven photocurrents on membrane potential is dramatically affected by temperature changes provoked by light stimulation itself, especially in the single-firing cell model.

In summary, we have shown that temperature increase caused by brain optical stimulation, with light intensities commonly used in optogenetic experiments (Cardin et al., 2009; Adamantidis et al., 2011) can considerably affect neuron and network properties independently of opsin expression. Moreover, the temperature can alter cellular responses to optical stimulation. As the usage of channelrhodopsin becomes widespread, studies tend to assume that optical stimulation elicits spiking activity without assessing cellular responses (Ahlbeck et al., 2018; Almada et al., 2018). Thus, the whole cell current- and voltage-clamp assessment of the cell response to optical stimulation may still be necessary to determine optimal light stimulation protocols.

## DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

## AUTHOR CONTRIBUTIONS

HP performed the COMSOL/MATLAB modeling and simulations about light propagation in brain tissue. RC performed the optical theoretical analysis. TM modeled the interaction of channelrhodopsin photocurrents with the ionic currents in the neuron. RL modeled the synchrony in a gamma oscillation network during temperature changes. HP, RC, TM, and RL wrote the paper. The authors read and approved the final manuscript.

## FUNDING

This work was supported by the Swedish Brain Foundation and the Brazilian agency Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES/STINT program).

## ACKNOWLEDGMENTS

The authors would like to thank the Neurodynamics Laboratories at UFRN and Uppsala University, as well as the CAPES/STINT program. This manuscript has been released as a Pre-Print at Peixoto et al. (2018).

## REFERENCES

(2015). *0.48 NA Polymer Clad Multimode Fiber*. Thorlabs, Inc.

Adamantidis, A., Arber, S., Bains, J. S., Bamberg, E., Bonci, A., Buzsáki, G., et al. (2015). Optogenetics: 10 years after ChR2 in neurons–views from the community. *Nat. Neurosci.* 18, 1202–1212. doi: 10.1038/nn.4106

Adamantidis, A. R., Tsai, H.-C., Boutrel, B., Zhang, F., Stuber, G. D., Budygin, E. A., et al. (2011). Optogenetic interrogation of dopaminergic modulation of the multiple phases of reward-seeking behavior. *J. Neurosci.* 31, 10829–10835. doi: 10.1523/JNEUROSCI.2246-11.2011

Adamantidis, A. R., Zhang, F., de Lecea, L., and Deisseroth, K. (2014). Establishing a fiber-optic-based optical neural interface. *Cold Spring Harb. Protoc.* 2014, 839–844. doi: 10.1101/pdb.prot083337

Adelsberger, H., Grienberger, C., Stroh, A., and Konnerth, A. (2014). *In vivo* calcium recordings and channelrhodopsin-2 activation through an optical fiber. *Cold Spring Harb. Protoc.* 2014, pdb–prot084145. doi: 10.1101/pdb.prot084145

Ahlbeck, J., Song, L., Chini, M., Bitzenhofer, S. H., and Hanganu-Opatz, I. L. (2018). Glutamatergic drive along the septo-temporal axis of hippocampus boosts prelimbic oscillations in the neonatal mouse. *eLife* 7:e33158. doi: 10.7554/eLife.33158

Ahmed, N., Ali Shah, N., Ahmad, B., Shah, S. I., Ulhaq, S., and Gorji, M. R. (2019). Transient MHD convective flow of fractional nanofluid between vertical plates. *J. Appl. Comput. Mech.* 5, 592–602.

Almada, R. C., Genewsky, A. J., Heinz, D. E., Kaplick, P. M., Coimbra, N. C., and Wotjak, C. T. (2018). Stimulation of the nigrotectal pathway at the level of the superior colliculus reduces threat recognition and causes a shift from avoidance to approach behavior. *Front. Neural Circuits* 12:36. doi: 10.3389/fncir.2018.00036

Andersen, P., and Moser, E. I. (1995). Brain temperature and hippocampal function. *Hippocampus* 5, 491–498. doi: 10.1002/hipo.450050602

Aravanis, A. M., Wang, L.-P., Zhang, F., Meltzer, L. A., Mogri, M. Z., Schneider, M. B., et al. (2007). An optical neural interface: *in vivo* control of rodent motor cortex with integrated fiberoptic and optogenetic technology. *J. Neural Eng.* 4:S143. doi: 10.1088/1741-2560/4/3/S02

Arias-Gil, G., Ohl, F. W., Takagaki, K., and Lippert, M. T. (2016). Measurement, modeling, and prediction of temperature rise due to optogenetic brain stimulation. *Neurophotonics* 3:045007. doi: 10.1117/1.NPh.3.4.045007

Balanis, C. A. (2012). *Advanced Engineering Electromagnetics*. NJ: John Wiley & Sons.

Bernstein, J. G., Han, X., Henninger, M. A., Ko, E. Y., Qian, X., Franzesi, G. T., et al. (2008). "Prosthetic systems for therapeutic optical activation and silencing of genetically targeted neurons," in *Biomedical Optics (BiOS) 2008* (International Society for Optics and Photonics), 68540H.

Cao, X.-J., and Oertel, D. (2005). Temperature affects voltage-sensitive conductances differentially in octopus cells of the mammalian cochlear nucleus. *J. Neurophysiol.* 94, 821–832. doi: 10.1152/jn.01049.2004

Cardin, J. A., Carlén, M., Meletis, K., Knoblich, U., Zhang, F., Deisseroth, K., et al. (2009). Driving fast-spiking cells induces gamma rhythm and controls sensory responses. *Nature* 459:663. doi: 10.1038/nature08002

Couchman, K., Garrett, A., Deardorff, A. S., Rattay, F., Resatz, S., Fyffe, R., et al. (2011). Lateral superior olive function in congenital deafness. *Hear. Res.* 277, 163–175. doi: 10.1016/j.heares.2011.01.012

Džimbeg-Malčić, V., Barbarić-Mikočević, Ž., and Itrić, K. (2011). Kubelka-munk theory in describing optical properties of paper (i). *Tehnički vjesnik* 18, 117–124.

Elwassif, M. M., Kong, Q., Vazquez, M., and Bikson, M. (2006). Bio-heat transfer model of deep brain stimulation-induced temperature changes. *J. Neural Eng.* 3:306. doi: 10.1088/1741-2560/3/4/008

Fenno, L., Yizhar, O., and Deisseroth, K. (2011). The development and application of optogenetics. *Annu. Rev. Neurosci.* 34, 389–412. doi: 10.1146/annurev-neuro-061010-113817

Fitzhugh, R. (1966). Theoretical effect of temperature on threshold in the hodgkin-huxley nerve model. *J. Gen. Physiol.* 49, 989–1005. doi: 10.1085/jgp.49.5.989

Foutz, T. J., Arlow, R. L., and McIntyre, C. C. (2012). Theoretical principles underlying optical stimulation of a channelrhodopsin-2 positive pyramidal neuron. *J. Neurophysiol.* 107, 3235–3245. doi: 10.1152/jn.00501.2011

Hilscher, M. M., Leão, K. E., and Leão, R. N. (2013). Synchronization through nonreciprocal connections in a hybrid hippocampus microcircuit. *Front. Neural Circuits* 7:120. doi: 10.3389/fncir.2013.00120

Jacques, S. L. (2013). Optical properties of biological tissues: a review. *Phys. Med. Biol.* 58:R37. doi: 10.1088/0031-9155/58/11/R37

Kim, J., and Connors, B. (2012). High temperatures alter physiological properties of pyramidal cells and inhibitory interneurons in hippocampus. *Front. Cell. Neurosci.* 6:27. doi: 10.3389/fncel.2012.00027

Klausberger, T., Marton, L. F., O'Neill, J., Huck, J. H., Dalezios, Y., Fuentealba, P., et al. (2005). Complementary roles of cholecystokinin-and parvalbumin-expressing gabaergic neurons in hippocampal network oscillations. *J. Neurosci.* 25, 9782–9793. doi: 10.1523/JNEUROSCI.3269-05.2005

Knöpfel, T., Lin, M. Z., Levskaya, A., Tian, L., Lin, J. Y., and Boyden, E. S. (2010). Toward the second generation of optogenetic tools. *J. Neurosci.* 30, 14998–15004. doi: 10.1523/JNEUROSCI.4190-10.2010

Leao, K. E., Leao, R. N., Sun, H., Fyffe, R. E., and Walmsley, B. (2006). Hyperpolarization-activated currents are differentially expressed in mice brainstem auditory nuclei. *J. Physiol.* 576, 849–864. doi: 10.1113/jphysiol.2006.114702

Leao, K. E., Leao, R. N., and Walmsley, B. (2011). Modulation of dendritic synaptic processing in the lateral superior olive by hyperpolarization-activated currents. *Eur. J. Neurosci.* 33, 1462–1470. doi: 10.1111/j.1460-9568.2011.07627.x

Leao, R. N., Leao, F. N., and Walmsley, B. (2005). Non-random nature of spontaneous mipscs in mouse auditory brainstem neurons revealed by recurrence quantification analysis. *Proc. R. Soc. Lond. B Biol. Sci.* 272, 2551–2559. doi: 10.1098/rspb.2005.3258

Leao, R. N., Leao, R. M., Da Costa, L. F., Rock Levinson, S., and Walmsley, B. (2008). A novel role for mntb neuron dendrites in regulating action potential amplitude and cell excitability during repetitive firing. *Eur. J. Neurosci.* 27, 3095–3108. doi: 10.1111/j.1460-9568.2008.06297.x

Leao, R. N., Tan, H. M., and Fisahn, A. (2009). Kv7/KCNQ channels control action potential phasing of pyramidal neurons during hippocampal gamma oscillations *in vitro*. *J. Neurosci.* 29, 13353–13364. doi: 10.1523/JNEUROSCI.1463-09.2009

Lerchner, W., Corgiat, B., Der Minassian, V., Saunders, R., and Richmond, B. (2014). Injection parameters and virus dependent choice of promoters to improve neuron targeting in the nonhuman primate brain. *Gene Ther.* 21, 233–241. doi: 10.1038/gt.2013.75

Madisen, L., Mao, T., Koch, H., Zhuo, J.-M., Berenyi, A., Fujisawa, S., et al. (2012). A toolbox of cre-dependent optogenetic transgenic mice for light-induced activation and silencing. *Nat. Neurosci.* 15, 793–802. doi: 10.1038/nn.3078

Martina, M., Schultz, J. H., Ehmke, H., Monyer, H., and Jonas, P. (1998). Functional and molecular differences between voltage-gated K+ channels of fast-spiking interneurons and pyramidal neurons of rat hippocampus. *J. Neurosci.* 18, 8111–8125. doi: 10.1523/JNEUROSCI.18-20-08111.1998

McDougal, R. A., Morse, T. M., Carnevale, T., Marenco, L., Wang, R., Migliore, M., et al. (2017). Twenty years of modeldb and beyond: building essential modeling tools for the future of neuroscience. *J. Comput. Neurosci.* 42, 1–10. doi: 10.1007/s10827-016-0623-7

Migliore, M. (1996). Modeling the attenuation and failure of action potentials in the dendrites of hippocampal neurons. *Biophys. J.* 71, 2394–2403. doi: 10.1016/S0006-3495(96)79433-X

Mikulovic, S., Pupe, S., Peixoto, H. M., Do Nascimento, G. C., Kullander, K., Tort, A. B., et al. (2016). On the photovoltaic effect in local field potential recordings. *Neurophotonics* 3:015002. doi: 10.1117/1.NPh.3.1.015002

Mobley, J., and Vo-Dinh, T. (2003). Optical properties of tissue. *Biomed. Photon. Handbook* 2, 1–2. doi: 10.1201/9780203008997.sec1

Peatross, J., and Ware, M. (2015). *Physics of Light and Optics*. Provo, UT 84602: Brigham Young University, Department of Physics.

Peixoto, H. M., Moreno, R., Moulin, T., and Leão, R. N. (2018). *Modeling the Effect of Temperature on Membrane Response of Light Stimulation in Optogenetically-Targeted Neurons*. Technical report, PeerJ Preprints. doi: 10.7287/peerj.preprints.27248v1

Pennes, H. H. (1948). Analysis of tissue and arterial blood temperatures in the resting human forearm. *J. Appl. Physiol.* 1, 93–122. doi: 10.1152/jappl.1948.1.2.93

Rothman, J. S., and Manis, P. B. (2003). The roles potassium currents play in regulating the electrical activity of ventral cochlear nucleus neurons. *J. Neurophysiol.* 89, 3097–3113. doi: 10.1152/jn.00127.2002

Sadiku, M. N. (2014). *Elements of Electromagnetics*. Oxford: Oxford University Press.

Shapiro, M. G., Homma, K., Villarreal, S., Richter, C.-P., and Bezanilla, F. (2012). Infrared light excites cells by changing their electrical capacitance. *Nat. Commun.* 3:736. doi: 10.1038/ncomms1742

Stujenske, J. M., Spellman, T., and Gordon, J. A. (2015). Modeling the spatiotemporal dynamics of light and heat propagation for *in vivo* optogenetics. *Cell Rep.* 12, 525–534. doi: 10.1016/j.celrep.2015.06.036

Svelto, O., and Hanna, D. C. (2010). *Principles of Lasers*. New York, NY: Springer.

Taheripour, S., Saffarian, M. R., and Daneh-Dezfuli, A. (2019). Heat transfer simulation in an industrial journal bearing using vof method. *J. Braz. Soc. Mech. Sci. Eng.* 41:248. doi: 10.1007/s40430-019-1751-6

Thompson, S. M., Masukawa, L. M., and Prince, D. A. (1985). Temperature dependence of intrinsic membrane properties and synaptic potentials in hippocampal CA1 neurons *in vitro*. *J. Neurosci.* 5, 817–824. doi: 10.1523/JNEUROSCI.05-03-00817.1985

Vo-Dinh, T. (2014). *Biomedical Photonics Handbook: Therapeutics and Advanced Biophotonics*. New York, NY: CRC Press.

Wang, X.-J., and Buzsáki, G. (1996). Gamma oscillation by synaptic inhibition in a hippocampal interneuronal network model. *J. Neurosci.* 16, 6402–6413. doi: 10.1523/JNEUROSCI.16-20-06402.1996

Weisstein, E. W. (2015). *Fourier Transform–Gaussian. From MathWorld–A Wolfram Web Resource*.

Welch, A. J., and Van Gemert, M. J. (2011). *Optical-Thermal Response of Laser-Irradiated Tissue*, Vol. 2. New York, NY: Springer.

Whittington, M. A., Traub, R. D., and Jefferys, J. G. (1995). Synchronized oscillations in interneuron networks driven by metabotropic glutamate receptor activation. *Nature* 373:612. doi: 10.1038/373612a0

Williams, J. C., Xu, J., Lu, Z., Klimas, A., Chen, X., Ambrosi, C. M., et al. (2013). Computational optogenetics: empirically-derived voltage-and light-sensitive channelrhodopsin-2 model. *PLoS Comput. Biol.* 9:e1003220. doi: 10.1371/journal.pcbi.1003220

Yang, L., and Miklavcic, S. J. (2005). Revised kubelka–munk theory. III. A general theory of light propagation in scattering and absorptive media. *JOSA A* 22, 1866–1873. doi: 10.1364/JOSAA.22.001866

Zhang, F., Tsai, H.-C., Airan, R. D., Stuber, G. D., Adamantidis, A. R., De Lecea, L., et al. (2015). Optogenetics in freely moving mammals: dopamine and reward. *Cold Spring Harb. Protoc.* 2015, 715–724. doi: 10.1101/pdb.top086330

Zimmerman, W. B. (2004). *Process Modelling and Simulation With Finite Element Methods*, Vol. 1. Singapore: World Scientific.

**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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 09 March 2020
doi: 10.3389/fncom.2020.00019

# Unsupervised Domain Adaptation With Optimal Transport in Multi-Site Segmentation of Multiple Sclerosis Lesions From MRI Data

*Antoine Ackaouy[1], Nicolas Courty[2], Emmanuel Vallée[3], Olivier Commowick[1], Christian Barillot[1] and Francesca Galassi[1]**

[1] Empenn, INRIA, IRISA, CNRS, INSERM, Rennes, France, [2] Panama/Obélix, INRIA, IRISA, Université de Bretagne Sud, Vannes, France, [3] Orange Labs, Lannion, France

Automatic segmentation of Multiple Sclerosis (MS) lesions from Magnetic Resonance Imaging (MRI) images is essential for clinical assessment and treatment planning of MS. Recent years have seen an increasing use of Convolutional Neural Networks (CNNs) for this task. Although these methods provide accurate segmentation, their applicability in clinical settings remains limited due to a reproducibility issue across different image domains. MS images can have highly variable characteristics across patients, MRI scanners and imaging protocols; retraining a supervised model with data from each new domain is not a feasible solution because it requires manual annotation from expert radiologists. In this work, we explore an unsupervised solution to the problem of domain shift. We present a framework, Seg-JDOT, which adapts a deep model so that samples from a source domain and samples from a target domain sharing similar representations will be similarly segmented. We evaluated the framework on a multi-site dataset, MICCAI 2016, and showed that the adaptation toward a target site can bring remarkable improvements in a model performance over standard training.

**Keywords: MS lesion segmentation, deep learning, convolutional neural networks, unsupervised domain adaptation, optimal transport**

OPEN ACCESS

***Edited by:***
*Petia D. Koprinkova-Hristova, Institute of Information and Communication Technologies (BAS), Bulgaria*

***Reviewed by:***
*Gongfa Li, Wuhan University of Science and Technology, China*
*Wenjia Bai, Imperial College London, United Kingdom*

****Correspondence:***
*Francesca Galassi*
*francesca.galassi@inria.fr*

***Received:*** *23 September 2019*
***Accepted:*** *12 February 2020*
***Published:*** *09 March 2020*

***Citation:***
*Ackaouy A, Courty N, Vallée E, Commowick O, Barillot C and Galassi F (2020) Unsupervised Domain Adaptation With Optimal Transport in Multi-Site Segmentation of Multiple Sclerosis Lesions From MRI Data. Front. Comput. Neurosci. 14:19. doi: 10.3389/fncom.2020.00019*

## 1. INTRODUCTION

Multiple Sclerosis (MS) is a chronic inflammatory-demyelinating disease of the central nervous system. Magnetic Resonance Imaging (MRI) is fundamental to characterize and quantify MS lesions; the number and volume of lesions are used for MS diagnosis, to track its progression and to evaluate treatments (Smith and McDonald, 1999). Current MRI protocols in MS consists in Fluid-Attenuated Inversion Recovery (FLAIR) and T1-weighted (T1-w) images, offering complementary contrasts that allows to identify different types of lesions. Accurate identification of MS lesions in MRI images is extremely difficult due to variability in lesion location, size, and shape, in addition to anatomical variability across patients. Since manual segmentation requires expert knowledge, it is time consuming and prone to intra- and inter-expert variability, several methods have been proposed to automatically segment MS lesions (García-Lorenzo et al., 2013; Commowick et al., 2018; Galassi et al., 2018).

In recent years, Convolutional Neural Networks (CNNs) have showed better performances in MS lesion segmentation than the traditional unsupervised methods (Commowick et al., 2018; Galassi et al., 2019). Yet, their clinical use remains limited due to a reproducibility issue across different sites or image domains. MRI MS imaging data can have high or subtle variations across individuals, MR scanners, and data acquisition protocols (Galassi et al., 2019; Kushibar et al., 2019; Onofrey et al., 2019). In research, the data used to train and test CNN models are never fully representative of all clinical scenarios, resulting in supervised models that suffer from poor generalization when applied to a new target image domain (Commowick et al., 2018).

A few studies have proposed methods to facilitate model re-training and re-use, such as Transfer Learning strategies (Kushibar et al., 2019), where the weights of an already trained network are tuned to adapt to a new target domain, decreasing the training time and demanding fewer training annotated samples than full training. Recent studies in computer vision propose Unsupervised Domain Adaptation strategies that do not require ground truth segmentation for the target dataset (Kouw and Loog, 2019). Our work deals with this more challenging and common scenario.

Unsupervised Domain Adaptation includes adversarial loss functions and adversarial image generation based methods (Sankaranarayanan et al., 2017; Tzeng et al., 2017). Generative adversarial approaches may generate image samples that are highly different from the actual MRI MS images and therefore make the network learn useless representations. One of the most recent works in Unsupervised Domain Adaptation proposes a solution for a classification task based on Optimal Transport, which learns a shared embedding for the source and target domains while preserving the discriminative information used by the classifier (Damodaran et al., 2018). Our framework is based on the latter approach. Learning a shared representation is suitable and relevant to our task where the aim is segmenting the same objects, MS lesions, within the same structure, the human brain.

In the sections that follow, we describe the use of Optimal Transport for Unsupervised Domain Adaptation and our original proposal, the Seg-JDOT framework. Seg-JDOT performs domain adaptation in a segmentation task thus alleviating the issue of low generalization ability in MS lesions segmentation. We demonstrate the effect of the adaptation on the classifier performance over standard training when training a model using data from a single site only and from multiple clinical sites. We employed the MICCAI 2016 dataset, which includes MRI MS images acquired with different scanners and protocols, and comprises patients with variable size and number of lesions.

# 2. METHODS

## 2.1. Problem Statement

The problem of generalizing across domains can be formally defined. Let $\Omega \in \mathbb{R}$ be an input space of dimension $d$, $\mathcal{C}$ the set of labels, and $\mathcal{P}(\Omega)$ the set of all probability measures over $\Omega$. Let $X$ be the instance space and $Y$ the label space. The differences between domains can be characterized by a change in the marginal feature distributions $\mathcal{P}(X)$ and in the conditional distributions $\mathcal{P}(Y|X)$.

In standard learning for a classification task, one assumes the existence of a source dataset $(\mathbf{X}_s, \mathbf{Y}_s)$, where $\mathbf{X}_s = \{\mathbf{x}_i^s\}_{i=1}^{N_s}$ is the instance data and $\mathbf{Y}_s = \{\mathbf{y}_\mathbf{i}^s\}_{i=1}^{N_s} \in C$ is the corresponding class labels, and a target dataset $\mathbf{X}_t = \{\mathbf{x}_\mathbf{j}^\mathbf{t}\}_{j=1}^{N_t}$ with unknown labels $\mathbf{Y}_t$. To infer the labels on the target dataset, one learns an empirical estimate of the joint probability distribution $\mathcal{P}(X, Y) \in \mathcal{P}(\Omega \times \mathcal{C})$ from $(\mathbf{X}_s, \mathbf{Y}_s)$ by learning a classifier $f$, under the assumption that the source and target data are drawn from the same distribution $\mu \in \mathcal{P}(\Omega)$. However, if the target set is drawn from a slightly different distribution, the learned classifier might under-perform on the target set. If the drift between the two distributions is not too large, a domain adaptation approach can be used to improve learned model generalization.

In our work, we deal with a domain adaptation problem that assumes the existence of two distinct joint probability distributions, $\mathcal{P}_s(X, Y)$ and $\mathcal{P}_t(X, Y)$, corresponding respectively to the source domain and to the target domain, with respective marginal distributions $\mu_s$ and $\mu_t$ over $\Omega$. We aim at leveraging the available information $\{\mathbf{X}_s, \mathbf{Y}_s, \mathbf{X}_t\}$ to learn a classifier $f$, that is a labeling function $\hat{f}$ which approximates $f_s$ and is closer to $f_t$ than any other function $\hat{f}_s$. In order to solve this unsupervised domain adaptation problem, the Optimal Transport theory can be employed (Courty et al., 2017; Damodaran et al., 2018).

### 2.1.1. Optimal Transport for Unsupervised Domain Adaptation

Optimal Transport is a theory that allows to compare and align probability distributions by seeking for a transport plan between them (Villani, 2008). Optimal Transport has been adopted in Unsupervised Domain Adaptation in order to compare the source and target distributions and bring them closer. Earlier use of Optimal Transport in Unsupervised Domain Adaptation involves finding a common latent space between the source and target domains where to learn a unique classifier, or finding a transport plan between the marginal feature distributions $\mu$ under the assumption of label regularity, i.e., the conditional probability remains unchanged (Gopalan et al., 2011; Courty et al., 2015).

Recently, Courty et al. proposed an approach that handles a shift in both the marginal and conditional probabilities, the Joint Distribution Optimal Transport framework (JDOT) (Courty et al., 2017). Formally, following the formulation of Optimal Transport given by Kantorovich (1942), their approach seeks for a transport plan between the two joint distributions $\mathcal{P}_s$ and $\mathcal{P}_t$, or equivalently a probabilistic coupling, $\gamma \in \Pi(\mathcal{P}_s, \mathcal{P}_t)$ such that:

$$\gamma_0 = \underset{\gamma \in \Pi(\mathcal{P}_s, \mathcal{P}_t)}{\arg\min} \int_{\Omega \times \Omega} \mathcal{D}(\mathbf{x}^s, \mathbf{y}^s; \mathbf{x}^t, \mathbf{y}^t) d\gamma(\mathbf{x}^s, \mathbf{y}^s; \mathbf{x}^t, \mathbf{y}^t), \quad (1)$$

where $\mathcal{D}$ is a joint cost function measuring both the dissimilarity between samples $\mathbf{x^s}$ and $\mathbf{x^t}$, and the discrepancy between $\mathbf{y^s}$ and $\mathbf{y^t}$. Because it is an unsupervised problem, the labels $\mathbf{y^t}$ are unknown and replaced by a proxy $f(\mathbf{x^t})$. Hence, they devised an efficient algorithm that aligns jointly the feature space and

label-conditional distributions, by optimizing simultaneously for a coupling $\gamma$ between $\mathcal{P}_s$ and $\mathcal{P}_t$ and a predictive function $f$ embedded in the cost function. The classifier $f$ on a target domain is learned according to the following optimization problem:

$$\min_{f,\gamma\in\Pi} \sum_{ij} \mathcal{D}(\mathbf{x}_i^s, \mathbf{y}_i^s; \mathbf{x}_j^t, f(\mathbf{x}_j^t))\gamma_{ij}, \tag{2}$$

where

$$\mathcal{D}(\mathbf{x}_i^s, \mathbf{y}_i^s; \mathbf{x}_j^t, f(\mathbf{x}_j^t)) = \alpha c(\mathbf{x}_i^s, \mathbf{x}_j^t) + \beta L(\mathbf{y}_i^s, f(\mathbf{x}_j^t)) \tag{3}$$

is a weighted combination of the distances in the feature space and the loss $L$ in the label space, for the i-th source and the j-th target sample.

Two limitations can be identified in the JDOT framework: (i) the cost $c$ is computed in the image space which can be poorly informative of the dissimilarity between samples, and (ii) the problem becomes intractable for large datasets since the coupling $\gamma$ scales quadratically with the number of samples.

Subsequently, Damodaran et al. proposed a deep learning strategy to solve these two drawbacks (Damodaran et al., 2018). Their Deep-JDOT framework (i) minimizes the cost $c$ in a deep layer of a Convolutional Neural Network, which is more informative than the original image space, and (ii) solves the problem with a stochastic approximation via mini-batches from the source and target domains. The Deep-JDOT model is thus composed of an embedding function $g : \mathbf{x} \rightarrow \mathbf{z}$ which maps the input space into a latent space, i.e., the output of a deep layer in the CNN, and a classifier $f : \mathbf{z} \rightarrow \mathbf{y}$ which maps the latent space into the output space. The optimization problem in Equation (2) therefore becomes:

$$\min_{\gamma\in\Pi, f, g} \sum_{ij} \mathcal{D}(g(\mathbf{x}_i^s), \mathbf{y}_i^s; g(\mathbf{x}_j^t), f(g(\mathbf{x}_j^t)))\gamma_{ij}, \tag{4}$$

where

$$\begin{aligned}\mathcal{D}(g(\mathbf{x}_i^s), \mathbf{y}_i^s; g(\mathbf{x}_j^t), f(g(\mathbf{x}_j^t))) \;=\; & \alpha||g(\mathbf{x}_i^s) - g(\mathbf{x}_j^t)||^2 \\ & +\beta L_t(\mathbf{y}_i^s, f(g(\mathbf{x}_j^t))).\end{aligned} \tag{5}$$

The first term in Equation (5) compares the embeddings for the source and the target domain, the second term considers the classification loss in the target domain and its regularity with respect to the labels in the source domain.

Equation (5) optimizes jointly the embedding function and the classifier to provide a model that performs well on a target domain. However, because Equation (5) takes into account the classifier learned in the target domain only, $f(g(\mathbf{x}^t))$, a performance degradation in the source domain might happen. To avoid such a degradation, they reintroduce the loss function $L_s$ evaluating the classifier learned on the source domain, $f(g(\mathbf{x}^s))$, yielding the following optimization problem:

$$\begin{aligned}\min_{\gamma, f, g} \frac{1}{n^s} \sum_i L_s(\mathbf{y}_i^s, f(g(\mathbf{x}_i^s))) \quad & + \sum_{i,j} \gamma_{ij}(\alpha||g(\mathbf{x}_i^s) - g(\mathbf{x}_j^t)||^2 \\ & +\beta L_t(\mathbf{y}_i^s, f(g(\mathbf{x}_j^t))).\end{aligned} \tag{6}$$

With this formulation, the framework learns a common latent space that conveys information for both the source and target domain. The final objective of Deep-JDOT is then to find an embedding function $g$ (which is equivalent to finding a latent space $\mathbf{z}$), a classifier $f$ and a transportation matrix such that inputs from the source and target domains that are similar in the latent space $\mathbf{z}$ are similarly classified. Importantly, solving the optimization problem with a stochastic approximation yields a computationally feasible solution which can be easily integrated into a deep learning framework. This approach is the starting point of our work and it will be further recalled and detailed in the next sections.

## 2.2. The Seg-JDOT Framework

We designed the Seg-JDOT framework to perform simultaneously a segmentation and an adaptation task. An overview of the framework is illustrated in **Figure 1**.

We employed a state-of-the-art deep learning architecture for brain lesion segmentation, a 3D-Unet (Isensee et al., 2018). The architecture was presented at the MICCAI BRATS 2018 segmentation challenge as an optimization of the original 3D-Unet proposed by Ronneberger et al. (2015).

The downward *context pathway* is a succession of *context modules*, with each module comprising two convolutional layers. The upward *localization pathway* combines the deepest representation with spatial information, brought by skip connections. This is achieved by first up-sampling the low dimensional representation and then combining it with the features from the corresponding output of the *context pathway*. To obtain the final segmentation maps, three different feature maps are combined through element-wise summation. Hence, from a compact representation with a low spatial dimension, a segmentation map with the same dimension as the input is obtained.

The model is composed of an embedding function $g: x \rightarrow z$, which maps the input $\mathbf{x}$ into the bottleneck representation $\mathbf{z}$, and a segmenter $f: z \rightarrow y$, which maps the latent space $\mathbf{z}$ into the segmentation space $\mathbf{y}$. Seg-JDOT optimizes jointly the latent space and the segmenter to provide a model that performs well on a target domain. In the sections that follow we provide a thorough description of the framework and the solution to the optimization problem.

### 2.2.1. Defining the Probability Distributions and the Representation Space

As described in the previous section, Optimal Transport allows to align the probability distribution in the source domain, $\mu_s$, and the probability distribution in the target domain, $\mu_t$. Defining the two probability distributions and the space where to compute their coupling $\gamma$ is not trivial and needs attention.

In a statistical context, we hardly have access to the true distribution $\mu$; instead, we work with an empirical distribution $\hat{\mu}_n$. The number of samples $n$ needed for $\hat{\mu}_n$ to be a reasonable proxy of $\mu$ grows with the number of dimensions $d$ of the space in which the distribution lies, a limit known as *the curse of dimensionality* (Bellman, 1961). The Wasserstein distance can be used to quantify the convergence of $\hat{\mu}_n$ to $\mu$. Dudley (1969)

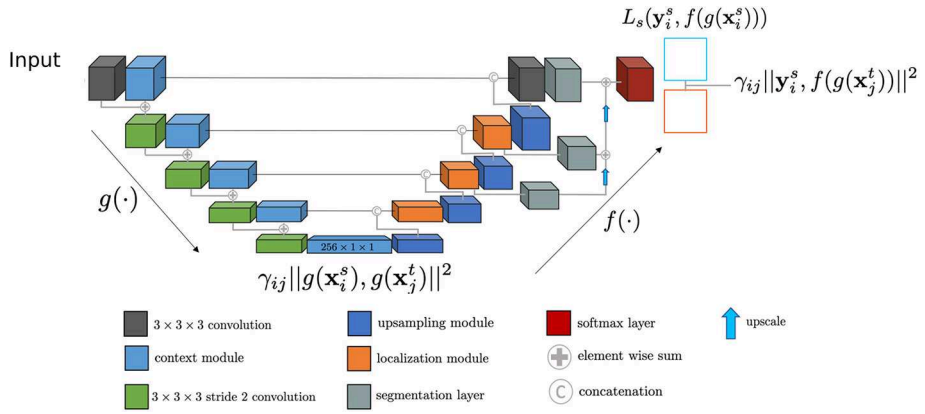


**FIGURE 1 |** The Seg-JDOT framework. $g(\cdot)$ is the embedding function, $f(\cdot)$ is the segmenter and, in between them, the bottleneck representation is the latent space $\boldsymbol{z}$ where we perform the adaptation. We report the terms of the Equation (8) at the levels where they are applied. We represent the output source images with a blue square and the output target images with a red square.

showed that $\mu$ absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ satisfies

$$\mathbb{E}[W_1(\mu, \hat{\mu}_n)] \lesssim n^{-1/d} \tag{7}$$

when $d > 2$. Equation (7) indicates that the expectation of the Wasserstein distance between $\hat{\mu}_n$ and $\mu$ grows exponentially with the number of dimensions $d$, a critical aspect in defining the probability distributions to be aligned.

In our work, we compute the Optimal Transport coupling in a deep layer of the CNN where the representation is compact and rich, which is the bottleneck layer of the 3D-Unet. The use of a compact latent space $\mathbf{z}$ allows to greatly reduce the original input dimensions. Moreover, solving the problem using mini-batches acts as a regularizer, which is important when working in high dimension (Genevay et al., 2019).

In order to define the probability distributions, we employ image patch samples rather than image samples as in Damodaran et al. (2018). The use of image patches enables an higher number of samples and, therefore, a more precise estimation of the true distribution $\mu$. Indeed, five image samples per domain would be insufficient to adequately represent a distribution in $\mathbf{z}$. It is important to notice that aligning patches rather than images is more reasonable for our task: two patches having similar lesions do not necessarily share the same location within the brain anatomy.

### 2.2.2. Defining the Global Loss Function

Damodaran et al. designed the Deep-JDOT framework to solve a classification and adaptation task simultaneously (Damodaran et al., 2018), so that samples from the source and target domain having similar representations in the latent space will be similarly classified by the network. The assumption is that if two images share the same label then they should have similar, if not equal, activation maps at some depth in the network. In their work, the loss functions $L_s$ and $L_t$ in Equation (6), respectively the loss in the label space in the source and in the target domain, were chosen to be the same i.e., the cross-entropy.

In our segmentation task, however, the correspondence between two similar activation maps and two similar segmentation maps is harder to establish. The variety of segmentation maps is generally much higher than the number of classes in a classification task. We cannot expect exact correspondence both in the latent space and in the segmentation space. While we chose the Dice Score as loss $L_s$, the choice of the loss $L_t$ was not trivial.

In order to define $L_t$, we conducted experiments involving the use of the Dice Score and the Squared Euclidean Distance. Results indicated an improved network performance in completing the task when using the Squared Euclidean Distance. Results involving the use of the Dice score can be found in **Supplementary Material**. This behavior might be explained by the fact that if two patches comprise a lesion of similar size and shape but different location within the patch, the Dice Score computed in the output space might be low because sensitive to a lesion location. On the contrary, the distance $||g(\mathbf{x}_i^s) - g(\mathbf{x}_j^t)||^2$ computed at the bottleneck layer of the network, where there is no spatial information, might indicate that the two representations are similar. Yet, for the framework to perform correctly the segmentation and adaptation task simultaneously, there must be an agreement between the distance in the latent space, $c$, and the loss in the output space, $L_t$. The Squared Euclidean distance is less sensitive to a lesion location than the Dice Score and therefore more appropriate for our task. On the basis of such considerations, we formulated the global loss function as:

$$\min_{\gamma, f, g} \frac{1}{n^s} \sum_i L_s(\mathbf{y}_i^s, f(g(\mathbf{x}_i^s))) + \sum_{i,j} \gamma_{ij} (\alpha ||g(\mathbf{x}_i^s) - g(\mathbf{x}_j^t)||^2 + \beta ||\mathbf{y}_i^s - f(g(\mathbf{x}_j^t))||^2). \tag{8}$$

### 2.2.3. Learning With Seg-JDOT

In Equation (8) two groups of variables need to be optimized: the optimal transport matrix $\gamma$ and the functions $g$ and $f$ induced

by the network. As suggested by Courty et al., the problem can be addressed by alternatively solving Equation (8) for $\gamma$, with fixed $g$ and $f$, and computing $g$ and $f$, with fixed $\gamma$ (Courty et al., 2017). When fixing $\hat{g}$ and $\hat{f}$, solving Equation (8) is equivalent to solving a classic Optimal Transport problem with cost matrix $C_{ij} = \alpha||\hat{g}(\mathbf{x}_i^s) - \hat{g}(\mathbf{x}_j^t)||^2 + \beta||\mathbf{y}_i^s - \hat{f}(\hat{g}(\mathbf{x}_j^t))||^2$; similarly, when fixing $\hat{\gamma}$, solving for $g$ and $f$ is a standard deep learning problem.

Damodoran et al. proposed to solve this optimization problem with a stochastic approximation using mini-batches from the source and target domains, so to ease the computation of the Optimal Transport (Damodaran et al., 2018). Using a mini-batch of size $m$ leads to the following optimization problem:

$$\min_{f,g} \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} L_s(\mathbf{y}_i^s, f(g(\mathbf{x}_i^s))) + \min_{\gamma\in\Gamma(\mu_s,\mu_t)} \sum_{i,j=1}^{m} \gamma_{ij}\left(\alpha||g(\mathbf{x}_i^s) - g(\mathbf{x}_j^t)||^2 + \beta||\mathbf{y}_i^s - f(g(\mathbf{x}_j^t))||^2\right)\right], \quad (9)$$

with $\mathbb{E}$ the expected value with respect to the mini-batches from the source and target domains. We summarize this approach in Algorithm 1.

**Algorithm 1:** Seg-JDOT stochastic optimization

**Require:** $\mathbf{x}^s$: source domain images, $\mathbf{x}^t$: target domain images, $\mathbf{y}^s$: source domain segmentation maps

**for** each source batch $(\mathbf{x}_b^s, \mathbf{y}_b^s)$ and target batch $(\mathbf{x}_b^s)$ **do**

fix $\hat{g}$ and $\hat{f}$, find $\gamma$ for the given batch

fix $\hat{\gamma}$, and use gradient descent to update $\hat{f}$ and $\hat{g}$

**end for**

In order to implement Algorithm 1, we separated the global loss function in Equation (9) into two loss functions that are computed at two different levels of the network.

We name the first loss function *representation alignment loss function* and compute it at the output of the bottleneck layer:

$$\sum_{i,j=1}^{m} \gamma_{ij}\alpha||g(\mathbf{x}_i^s) - g(\mathbf{x}_j^t)||^2. \quad (10)$$

**TABLE 1 |** The MICCAI 2016 MS lesion segmentation challenge dataset contains MR images of MS patients from four different MRI scanners.

| Site | MRI scanner | Modality | Train subjects | Test subjects |
|---|---|---|---|---|
| 01 | GE Discovery 3T | 3D FLAIR 3D T1 | 5 | 10 |
| 03 | Philips Ingenia 3T | 3D FLAIR 3D T1 | 0 | 8 |
| 07 | Siemens Aera 1.5T | 3D FLAIR 3D T1 | 5 | 10 |
| 08 | Siemens Verio 3T | 3D FLAIR 3D T1 | 5 | 10 |
| Total | | | 15 | 38 |

*Sites 01, 07, and 08 include 5 train images and 10 test images; site 03 contains 8 test images.*

The *representation alignment loss function* ensures that a source sample and a target sample that are heavily connected (high $\gamma$ value) have representations not far in the Euclidean distance sense. By back-propagating through all the shallower layers, we ensure a domain independent representation.

We name the second loss function *segmentation alignment loss function* and compute it at the final output layer:

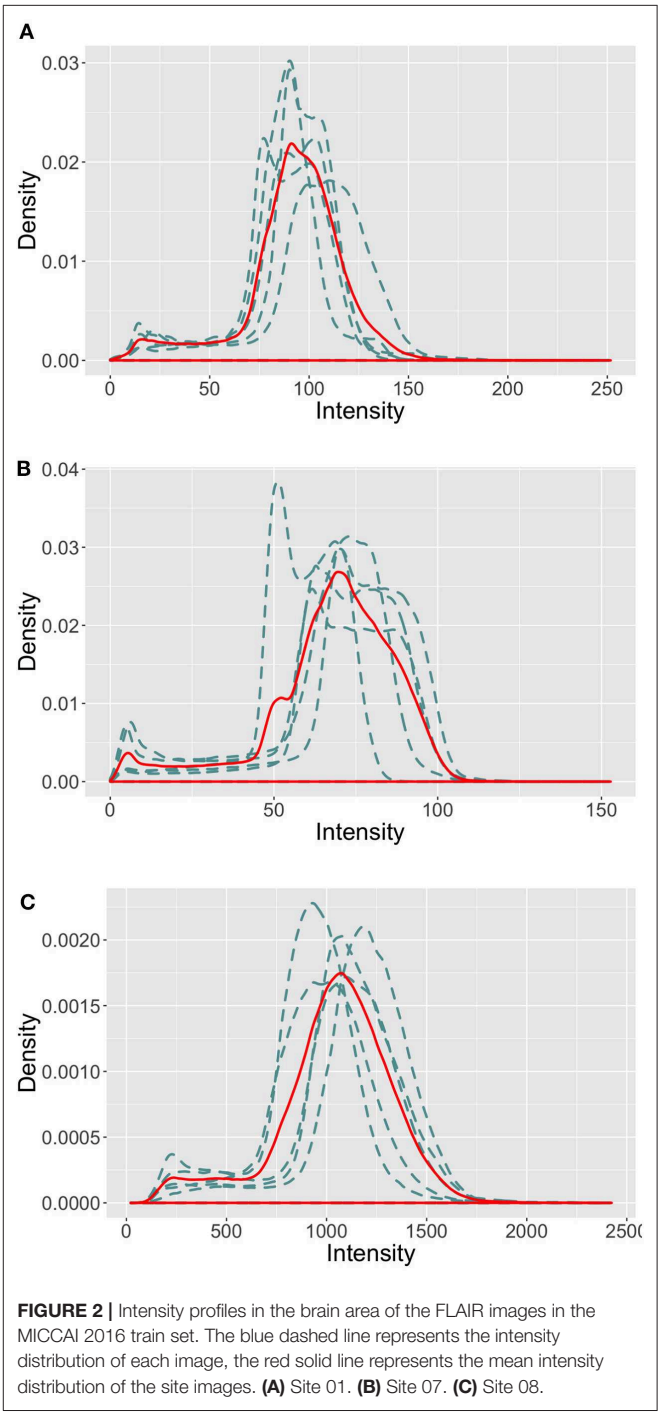


**FIGURE 2 |** Intensity profiles in the brain area of the FLAIR images in the MICCAI 2016 train set. The blue dashed line represents the intensity distribution of each image, the red solid line represents the mean intensity distribution of the site images. **(A)** Site 01. **(B)** Site 07. **(C)** Site 08.

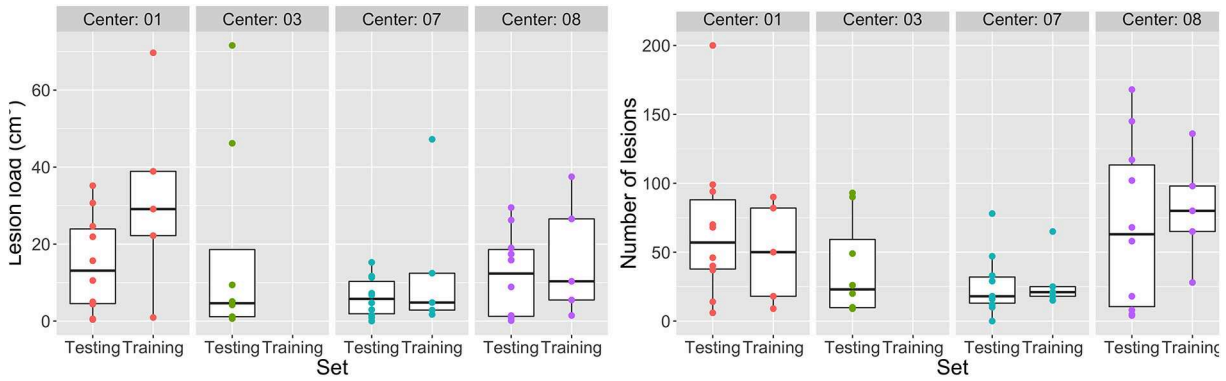


**FIGURE 3 |** Variability in MS lesion volume and number. Lesion load per patient per site **(Left)** and Number of lesions per patient per site **(Right)**.

$$\frac{1}{m}\sum_{i=1}^{m} L_s(\mathbf{y}_i^s, f(g(\mathbf{x}_i^s))) + \sum_{i,j=1}^{m} \gamma_{ij}\beta||\mathbf{y}_i^s - f(g(\mathbf{x}_j^t))||^2. \quad (11)$$

The first term of the *segmentation alignment loss function* allows to avoid a degradation of the performances in the source domain; the second term ensures that a source sample connected to a target sample has an output which is not too far from the true segmentation of the target sample in the Euclidean distance sense.

# 3. EXPERIMENTS AND RESULTS

## 3.1. Dataset

Proper selection of the dataset for the unsupervised domain adaptation experiments is crucial because the domain difference should be present to confirm the framework's robustness. In this work, we employ a well-known dataset, the MICCAI 2016 MS lesion segmentation challenge dataset (Commowick et al., 2018). It contains 53 MRI images of patients suffering from MS, split into 15 train and 38 test images. For each patient, high quality segmentation maps are provided—they were computed from seven independent manual segmentations and using LOPSTAPLE (Akhondi-Asl et al., 2014) so to minimize inter-expert variability.

Images were acquired in four different clinical sites, corresponding to four different MRI scanner models (**Table 1**). Each clinical site includes 5 train and 10 test patients (sites 01, 07, 08), except one site that contains 8 test patients only (site 03). In our experiments, we used the test images for testing purpose only and we never included them in the training or validation or adaptation process.

All MRI imaging protocols included 3D FLAIR and 3D T1-w anatomical images. Image size and resolution were different across the four MRI scanners (more details on the imaging protocol are available on the challenge website[1]). As illustrated in **Figure 2**, the intensity profiles in the brain area vary across the MRI scanners. Sites 01 and 07 follow a similar profile with a maximum intensity $\approx$ 200, while they vary drastically from site 08, where the intensity reaches up to $\approx$ 2,000 (a similar distribution was observed for site 03, test images). This behavior in intensity distribution was observed for both the imaging modalities, train and test patients.

Moreover, patients show a variability in MS lesion volume and number of lesions (**Figure 3**). The median lesion load in the train (test) dataset is for site 01 $\approx$ 30($\approx$ 16) cm$^3$, for site 03 $\approx$ (5) cm$^3$, for site 07 $\approx$ 5(6) cm$^3$, and for site 08 $\approx$ 10(12) cm$^3$. A similar variation across sites was observed in the number of lesions.

Considering these variations across the four clinical sites, the MICCAI 2016 dataset does fit the challenge of the domain shift problem.

## 3.2. Implementation Details

### 3.2.1. Image Pre-processing

Before extracting the patch samples from the image volumes to train the network, we performed a few standard pre-processing steps on the raw MRI images. For each patient, (i) MRI images were denoised (Coupe et al., 2008), (ii) rigidly registered toward the FLAIR modality (Commowick et al., 2012), (iii) skull-stripped (Manjón and Coupé, 2016), and (iv) bias corrected (Tustison et al., 2010). These steps involved the use of Anima, an openly available toolkit for medical image processing developed by the Empenn research team, Inria Rennes[2].

In order to preserve the challenge of the domain shift, we did not standardize intensities across sites. However, as the drastic variation in the intensity profiles would make the training process unnecessarily hard, we adjusted the intensities of each patient image to have zero mean and unit variance.

### 3.2.2. CNN Training

Images were resampled to the same size 128 × 128 × 128; 3D patches of size 16 × 16 × 16 were extracted. We employed a patch overlap of 50%, resulting in 4,096 patches per image. Although overlapping 3D patches contain more surrounding information for a voxel, it is memory demanding; training on patches containing lesions allowed to reduce training time while reducing class imbalance.

[1]https://portal.fli-iam.irisa.fr/msseg-challenge/data

[2]https://github.com/Inria-Visages/Anima-Public

CNN training was performed in batches containing 256 source and 256 target samples, with a total batch size of 512—the maximum size that the employed GPU can handle. Since the quality of approximation of the true optimal transport coupling depends on the number of samples, we chose to use the maximum batch size possible.

### 3.2.3. Technical Details

The Seg-JDOT framework was implemented in Python using the Keras library and the POT library (Flamary and Courty, 2017) which contains helpful functions for the Optimal Transport solver. Experiments were conducted on the GPU NVIDIA Quadro P6000, 24 GB.

## 3.3. Results on the MICCAI 2016 Dataset

We evaluated the segmentation performance when training both on a single site and on multiple clinical sites. The first experiment represents the worst case scenario, with training data acquired on a unique MR scanner; the second experiment reflects a more recurrent situation in the real practice, with training data coming from more than one MR scanner and a model that shall be more robust to variability.

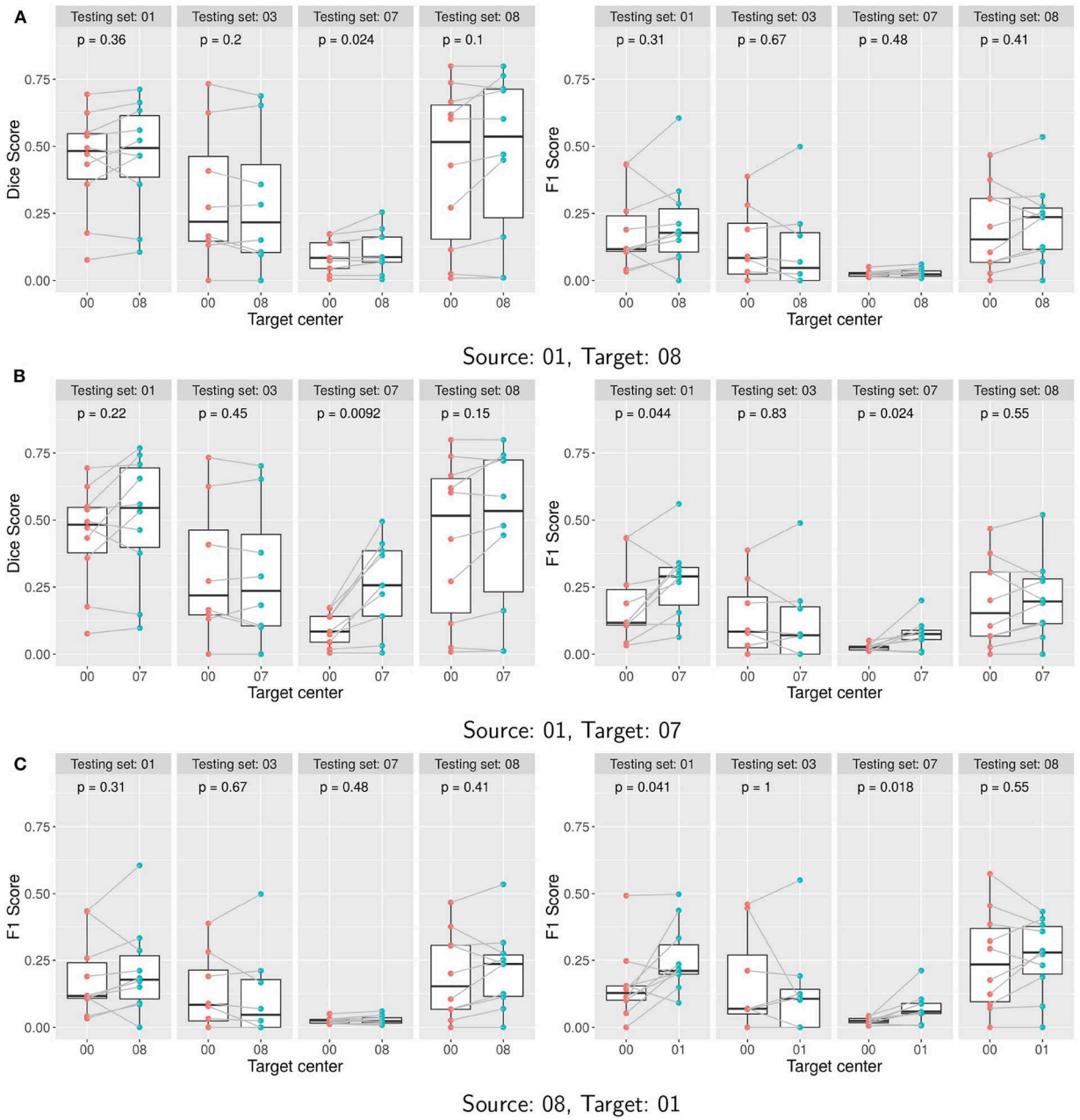


**FIGURE 4 |** Continued

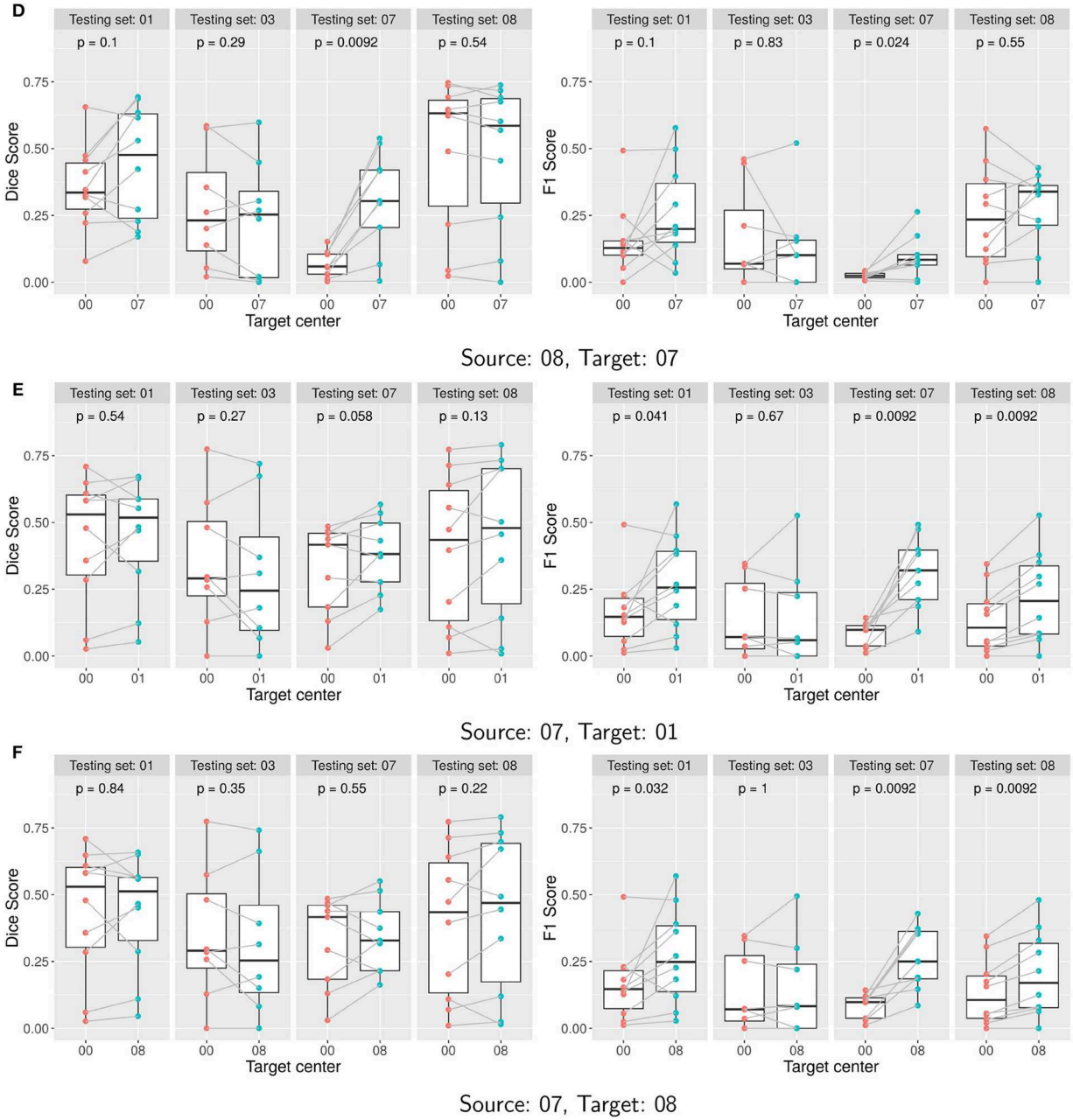


**FIGURE 4 | (A–F)** Performance of Seg-JDOT with single-site source and single-site target domain adaptation. Each row corresponds to a combination of source and target. Dice score (left column) and F1 score (right column) are computed with no adaptation (00) and with Seg-JDOT, where the direction of the domain adaptation is indicated (07, 08, or 01). For each combination of source and target, performances are given for all the four testing sites. Each point is a patient of a given site; performances of a patient *with* and *without* Seg-JDOT are tracked. For each site, the *p*-value of the paired Wilcoxon test is reported.

### 3.3.1. Single-Site Training

First, we evaluated the segmentation performance when training on a single site only. Hence, we applied the Seg-JDOT framework with one site as the source domain and any other site as the target domain. We did not perform adaptation toward the site 03 because it does not contain a train dataset.

The segmentation performance was assessed in terms of Dice score and F1 score. The Dice score is a measure of spatial overlap between the output and the ground truth; the F1 score is a weighted average of the lesion sensitivity and the positive predictive value, hence a metric that is independent of the lesion contour quality.

For each combination source/target, we compared the scores as obtained with the standard training (source only) with the scores as obtained with the adapted model. While the main focus of our study is the variation in performance on the target domain, we also evaluated the scores achieved by

the adapted classifier on the other clinical sites. This allowed us to assess a possible degradation in the source domain performance and the overall effect of the adaptation on the model generalization ability.

Boxplots of the Dice and F1 scores (**Figure 4**) illustrate the effect of the domain adaptation. For each site, we assessed the significance between pair-wise comparisons of the performances of the two learned classifiers. The Shapiro-Wilk's test of normality indicated a non-normal distribution of the samples and thus a paired Wilcoxon test was used (Rey and Neuhäuser, 2011). Reported $p$-values were computed using the paired Wilcoxon test and indicate whether the variations are statistically significant: if the $p$-value is lower than the significance level of 0.05, then we can state that the scores as computed with the two approaches are significantly different.

In **Figure 5**, we report the overall percentage of variation in performance on the target site. A positive variation indicates an improvement in the score. More detailed information can be found in **Supplementary Material**.

Results indicate that target site performances generally improve when applying the Seg-JDOT framework. The domain adaptation toward the site 07 yields the most significant improvement in target performance (**Figures 4B,D**), while the adaptation toward the site 08 yields minor variations only (**Figures 4A,F**).

The highest improvement is registered for the combination source site 08 and target site 07 (**Figure 5**), with a variation in the Dice score and F1 score of about 338 and 295%, respectively. It indicates that the adaptation reduces the effect of the high variability in intensity and lesion load/number that we observed across the two sites. When considering the adaptation in the other direction, i.e., the combination source site 07 and target site 08, we observe that the variability across the two sites did not affect that much the model performance, with a variation in the Dice score and F1 score of about 10 and 51%, respectively. In other words, the model learned on the site 07 appears to be more robust and to generalize better to other sites. This might be due to the fact that the samples within the site 07 are the most challenging and representative among all the sites.

Adapting toward a target domain appears beneficial, or otherwise not detrimental, for the overall generalization ability of a model. For instance, for the combination source site 08 and target site 01 we note a significant improvement in segmentation outcome also on the test site 07 (**Figure 4C**). For the combination source site 01 and target site 08, the adaptation does not yield a significant improvement in performance on the target site (**Figure 4A**); yet, a minor improvement in the Dice score is registered on the test site 07. This suggests that the adaptation toward a target domain allows to learn a classifier that is less specific to the source domain and thus capable to generalize better.

The adaptation can be beneficial for the source site as well. We observe an improvement in the F1 score on the source site for the combination source site 07 and target site 01 (**Figure 4E**) or target site 08 (**Figure 4F**), and for the combination source site 01 and target site 07 (**Figure 4B**). This might be explained by the

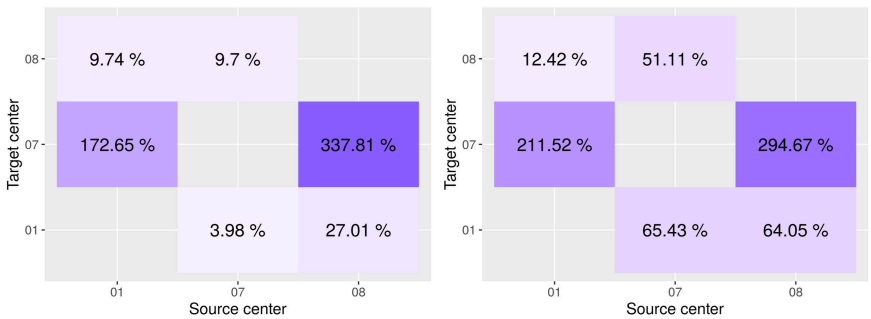


**FIGURE 5 |** Variation in performance on the target site between the model as learned on the source only and adapted on the target domain. Dice score on the left, F1 score on the right. On the $x$-axis is the source center, on the $y$-axis is the target center.

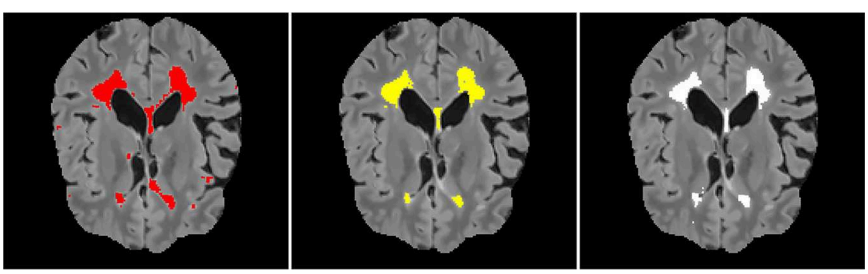

**FIGURE 6 |** A qualitative result for the combination source site 08 and target site 07. The results are shown in the coronal views of the FLAIR image. From the left: a segmentation result on site 07 when training on the site 08, segmentation result after adaptation, ground truth.

fact that the network is trained to minimize the Dice Score rather than the F1 Score and, therefore, the adaptation may move the network away from the optimal Dice Score solution and closer to the optimal F1 Score solution.

A qualitative result on a patient from the site 07 for the combination source site 08 and target site 07 is shown in **Figure 6**. We observe that the adaptation toward the site 07 yields a better segmentation output than training on the source site only. The number of false positives appears greatly reduced.

### 3.3.2. Multi-Site Training

We evaluated the segmentation performance when training on multiple clinical sites. Hence, the source domain comprised multiple sites (two) and the target domain was the remaining one.

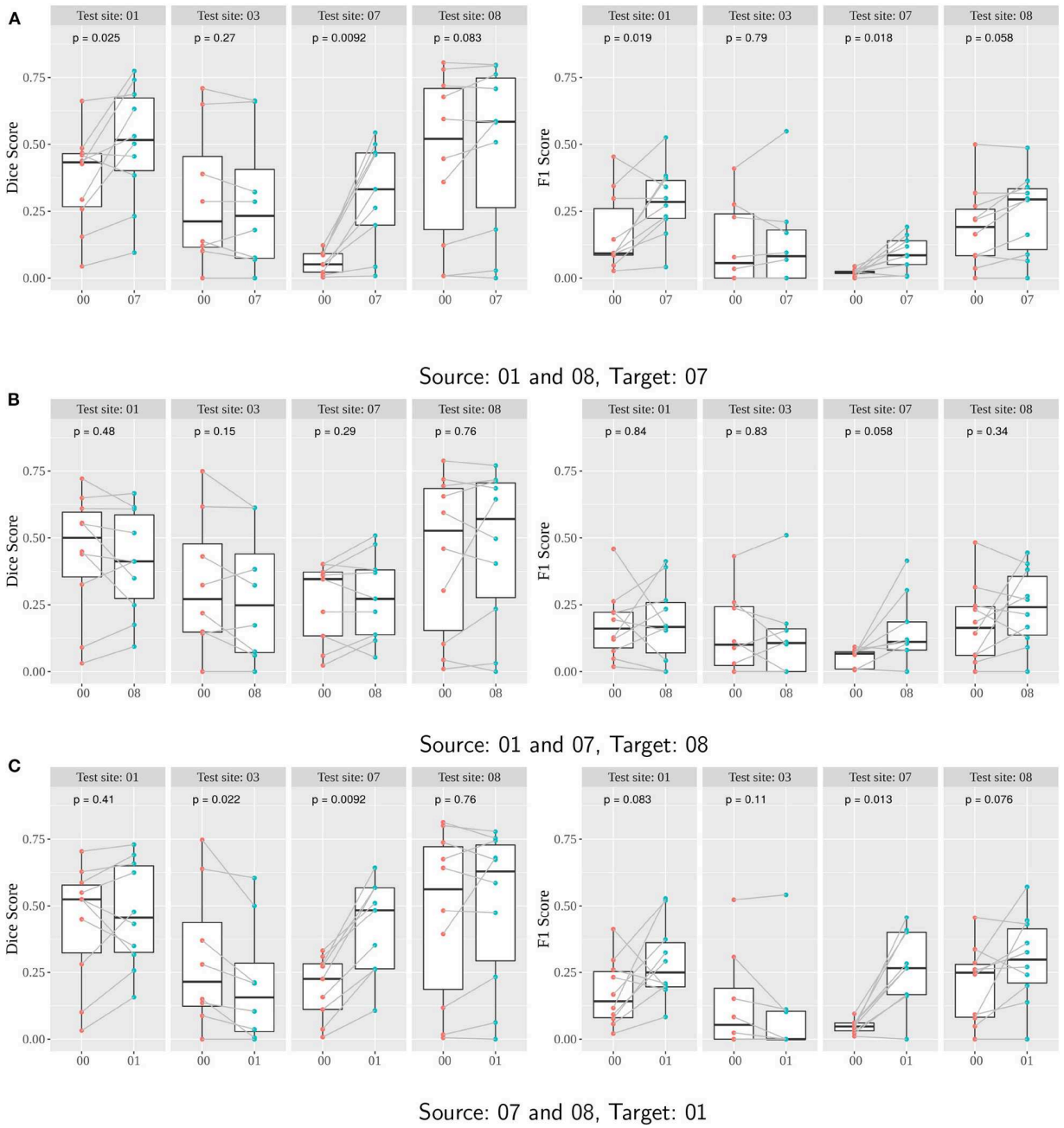


**FIGURE 7 | (A–C)** Performance of Seg-JDOT with multi-site source and single-site target domain adaptation. Each row corresponds to a combination of source and target. Dice score (left column) and F1 score (right column) are computed with no adaptation (00) and with Seg-JDOT, where the direction of the domain adaptation is indicated (07, 08, or 01). For each combination of source and target, performances are given for all the four testing sites. Each point is a patient of a given site; performances of a patient *with* and *without* Seg-JDOT are tracked. For each site, the *p*-value of the paired Wilcoxon test is reported.

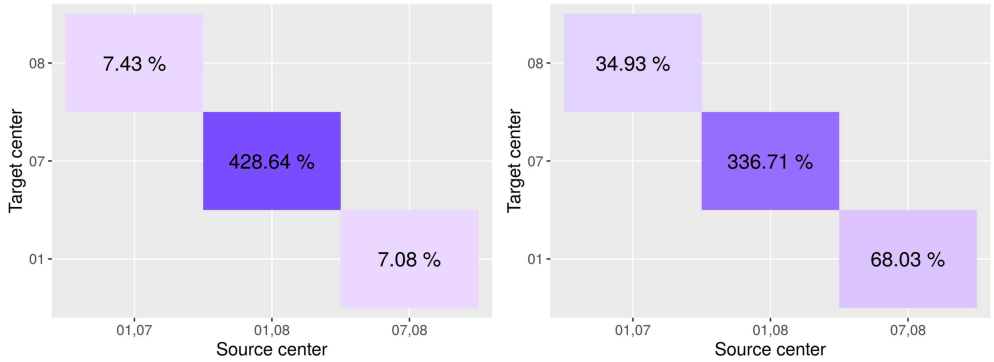


**FIGURE 8 |** Variation in performance on the target site between the model as learned on the source only and adapted on the target domain. On the *x*-axis is the source center, on the *y*-axis is the target center.

The site 03 was used for testing purpose only since it does not include a train dataset.

As for single-site training, the classifier performance was assessed in terms of Dice score and F1 score. For each combination source/target, we tested the classifier as adapted with Seg-JDOT on the target site as well as on the other test sites, so to assess the impact of the adaptation on the source performance and on the overall model generalization ability.

Boxplots of the Dice and F1 scores illustrate the effect of the domain adaptation on a clinical site (**Figure 7**). *P*-values were computed using the paired Wilcoxon-test.

In **Figure 8**, we report the overall percentage of variation in performance on a target site. A positive variation indicates an improvement in the score. More detailed information can be found in **Supplementary Material**.

Results indicate that Seg-JDOT generally improves the performances on the target site. As for single-site training, the most significant improvement is achieved on the target site 07 when the site 08 is a source domain (**Figure 7A**), with an overall variation in the Dice score of about 429% and in the F1 score of about 337% (**Figure 8**), while the least significant improvement is achieved on the target sites 08 (**Figure 7B**) and 01 (**Figure 7C**). This suggests that the less a model generalizes to a site, the more likely the adaptation will improve its performance on the latter, and vice-versa.

The adaptation can be beneficial for a source domain as well. We observe an improvement in the scores on the source site 01 for the combination source sites 01 and 08, and target site 07 (**Figure 7A**). Similarly, the source site 07 benefits from an adaptation toward the target site 01 (**Figure 7C**). For these combinations, the adaptation has thus a regularizing effect that yields an improvement in performance also on the source site.

In order to fully appreciate the effectiveness of the adaptation, we compared Seg-JDOT with training on standardized images. The intensities were standardized using the method of Nyul et al. (2000). Detailed results can be found in the **Supplementary Material**. A significant improvement was still achieved on the target site 07, with an overall variation in the Dice score of about 181% and in the F1 score of about 204%.

## 4. DISCUSSION AND CONCLUSION

In this paper, we presented the Seg-JDOT framework for Unsupervised Domain Adaptation based on Optimal Transport. The framework aims at adapting a model so that samples from a source and a target domain sharing similar representations will yield similar predictions. The framework was designed to perform an MS lesion segmentation task while addressing the recurrent situation of deploying a model on a clinical target site that was not included in the training process. Importantly, the adaptation does not require any manually annotated image in the target domain.

We tested the framework on the MICCAI 2016 MS lesion segmentation challenge dataset which includes four clinical sites presenting variations in intensity profile and lesion load or number. Our results with single-source and multi-source training indicate that the adaptation toward a target site can yield significant improvement in the model performance over standard training. The improvement appears to be the most significant for models having otherwise a low generalization ability. Adaptation toward a target site can bring improvements in the overall generalization ability of the model toward any domains. Also, the source performance is either not affected by the adaptation or an increase in the scores is observed.

A comparison of Seg-JDOT performances with training on standardized images indicates that the domain shift problem is still there after image standardization. This suggests that Seg-JDOT implicitly performs a normalization by adapting the weights to better interpret the features extracted by the network.

Although the approach was shown to be effective to deal with the domain adaptation problem, our dataset included clinical sites comprising five training subjects only. Future work will consider the evaluation of this approach with different data splits, other MS dataset and more subjects. Also, other measures of variability across sites and patients might be taken into account, such MS lesion types or patient age.

Seg-JDOT can easily be adapted to other neural network architectures or tasks. In this work, we have employed a variation of a 3D-Unet architecture recently proposed for a brain lesion

segmentation task. However, the use of image-wise segmentation outputs, rather than voxel-wise, may limit the performance of the framework because the output predictions in the target domain can only approximately fit the target lesion. Future work will consider the evaluation of the framework with other CNN architectures, such as the voxel-wise CNN network proposed by Valverde et al. (2017).

## DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://portal.fli-iam.irisa.fr/msseg-challenge/data.

## AUTHOR'S NOTE

This article has been released as a Preprint at Ackaouy et al. (2019).

## AUTHOR CONTRIBUTIONS

AA conception and design of the work, code implementation and experiments, results interpretation, drafting the article, critical revision of the article. NC theoretical formalism, results interpretation, critical revision of the article. EV results interpretation, drafting the article, critical revision of the article. OC and CB conception and design of the work, data selection. FG conception, design and supervision of the work, data selection, results interpretation, drafting the article, critical revision of the article, final approval of the version to be published.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.00019/full#supplementary-material

## REFERENCES


Ackaouy, A., Courty, N., Vallee, E., Commowick, O., Barillot, C., and Galassi, F. (2019). Preprint: unsupervised domain adaptation with optimal transport in multi-site segmentation of multiple sclerosis lesions from MRI data. Available online at: https://hal.archives-ouvertes.fr/hal-02317028

Akhondi-Asl, A., Hoyte, L., Lockhart, M. E., and Warfield, S. K. (2014). A logarithmic opinion pool based staple algorithm for the fusion of segmentations with associated reliability weights. *IEEE Trans. Med. Imaging* 33, 1997–2009. doi: 10.1109/TMI.2014.2329603

Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour*. Berkeley, CA: Princeton University Press.

Commowick, O., Istace, A., Kain, M., Laurent, B., Leray, F., Simon, M., et al. (2018). Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. *Sci. Rep.* 8:13650. doi: 10.1038/s41598-018-31911-7

Commowick, O., Wiest-Daesslé, N., and Prima, S. (2012). "Block-matching strategies for rigid registration of multimodal medical images," in *9th IEEE International Symposium on Biomedical Imaging (ISBI'2012)* (Barcelona), 700–703.

Coupe, P., Yger, P., Prima, S., Hellier, P., Kervrann, C., and Barillot, C. (2008). An optimized blockwise nonlocal means denoising filter for 3-d magnetic resonance images. *IEEE Trans. Med. Imaging* 27, 425–441. doi: 10.1109/TMI.2007.906087

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. *arXiv e-prints* arXiv:1705.08848. doi: 10.1109/TPAMI.2016.2615921

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2015). Optimal transport for domain adaptation. *CoRR*, abs/1507.00504.

Damodaran, B. B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. (2018). Deepjdot: deep joint distribution optimal transport for unsupervised domain adaptation. *CoRR*, abs/1803.10081.

Dudley, R. M. (1969). The speed of mean glivenko-cantelli convergence. *Ann. Math. Statist.* 40, 40–50. doi: 10.1214/aoms/1177697802

Flamary, R., and Courty, N. (2017). Pot python optimal transport library. Available online at: https://github.com/rflamary/POT

Galassi, F., Commowick, O., Vallée, E., and Barillot, C. (2018). "Voxel-wise comparison with a-contrario analysis for automated segmentation of multiple sclerosis lesions from multimodal MRI," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Vol. 11383. (Granada: Springer Lecture Notes in Computer Science), 11.

Galassi, F., Commowick, O., Vallée, E., and Barillot, C. (2019). "Deep learning for multi-site ms lesions segmentation: two-step intensity standardization and generalized loss function," in *16th IEEE International Symposium on Biomedical Imaging (ISBI)* (Venice), 1.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D. L., and Collins, D. L. (2013). Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. *Med. Image Anal.* 17, 1–18. doi: 10.1016/j.media.2012.09.004

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). "Sample complexity of sinkhorn divergences," in *Proceedings of Machine Learning Research (PMLR)*, Vol. 89, 1574–1583. Available online at: http://proceedings.mlr.press/v89/genevay19a.html

Gopalan, R., Ruonan, L., and Chellappa, R. (2011). "Domain adaptation for object recognition: an unsupervised approach," in *2011 International Conference on Computer Vision* (Barcelone), 999–1006. doi: 10.1109/ICCV.2011.6126344

Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., and Maier-Hein, K. H. (2018). Brain tumor segmentation and radiomics survival prediction: Contribution to the BRATS 2017 challenge. *CoRR*, abs/1802.10508.

Kantorovich, L. V. (1942). On the transfer of masses. *Dokl. Acad. Nauk* 37, 7–8.

Kouw, W. M., and Loog, M. (2019). A review of single-source unsupervised domain adaptation. *CoRR*, abs/1901.05335.

Kushibar, K., Valverde, S., González-Villá, S., Bernal, J., Cabezas, M., Oliver, K., et al. (2019). Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction. *Sci. Rep.* 9:6742. doi: 10.1038/s41598-019-43299-z

Manjón, J. V., and Coupé, P. (2016). Volbrain: an online MRI brain volumetry system. *Front. Neuroinformatics* 10:30. doi: 10.3389/fninf.2016.00030

Nyúl, L. G., Udupa, J. K., and Zhang, X. (2000). New variants of a method of MRI scale standardization. *IEEE Trans. Med. Imaging* 19, 143–150. doi: 10.1109/42.836373

Onofrey, J. A., Casetti-Dinescu, D. I., Lauritzen, A. D., Sarkar, S., Venkataraman, R., Fan, R. E., et al. (2019). "Generalizable multi-site training and testing of deep neural networks using image normalization," in *16th IEEE International Symposium on Biomedical Imaging (ISBI)* (Venice), 348–351. doi: 10.1109/ISBI.2019.8759295

Rey, D., and Neuhäuser, M. (2011). *Wilcoxon-Signed-Rank Test*. Berlin; Heidelberg: Springer, 1658–1659. doi: 10.1007/978-3-642-04898-2_616

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Sankaranarayanan, S., Balaji, Y., Castillo, C. D., and Chellappa, R. (2017). Generate to adapt: aligning domains using generative adversarial networks. *CoRR*, abs/1704.01705.

Smith, K. J., and McDonald, W. I. (1999). The pathophysiology of multiple sclerosis the mechanisms underlying the production of symptoms and the natural history of the disease. *Philos. Trans. R. Soc. B Biol. Sci.* 354, 1649–1673. doi: 10.1098/rstb.1999.0510

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4itk: Improved n3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. *CoRR*, abs/1702.05464.

Valverde, S., Cabezas, M., Roura, E., González-Villá, S., Pareto, D., Vilanova, J. C., et al. (2017). Improving automated multiple sclerosis lesion segmentation with a cascaded 3d convolutional neural network approach. *Neuroimage* 155, 159–168. doi: 10.1016/j.neuroimage.2017.04.034

Villani, C. (2008). *Optimal Transport – Old and New*, Vol. 338 (Berlin; Heidelberg: Grundlehren der mathematischen Wissenschaften. Springer). doi: 10.1007/978-3-540-71050-9

**Conflict of Interest:** EV was employed by Orange Labs for a period of time, but is no longer affiliated. The company had no influence on or contribution to the design, methodology or results of this study.

The remaining 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.

frontiers in Computational Neuroscience

ORIGINAL RESEARCH
published: 15 May 2020
doi: 10.3389/fncom.2020.00024

# Stochastic Resonance Based Visual Perception Using Spiking Neural Networks

*Yuxuan Fu*[1], *Yanmei Kang*[1]* *and Guanrong Chen*[2]

[1] Department of Applied Mathematics, School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China,
[2] Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China

Our aim is to propose an efficient algorithm for enhancing the contrast of dark images based on the principle of stochastic resonance in a global feedback spiking network of integrate-and-fire neurons. By linear approximation and direct simulation, we disclose the dependence of the peak signal-to-noise ratio on the spiking threshold and the feedback coupling strength. Based on this theoretical analysis, we then develop a dynamical system algorithm for enhancing dark images. In the new algorithm, an explicit formula is given on how to choose a suitable spiking threshold for the images to be enhanced, and a more effective quantifying index, the variance of image, is used to replace the commonly used measure. Numerical tests verify the efficiency of the new algorithm. The investigation provides a good example for the application of stochastic resonance, and it might be useful for explaining the biophysical mechanism behind visual perception.

**Keywords: stochastic resonance, spiking networks, visual perception, variance of image, contrast enhancement**

OPEN ACCESS

***Edited by:***
*Yu-Guo Yu,*
*Fudan University, China*

***Reviewed by:***
*Ergin Yilmaz,*
*Bulent Ecevit University, Turkey*
*Huaguang Gu,*
*Tongji University, China*

***Correspondence:**
*Yanmei Kang*
*ymkang@xjtu.edu.cn*

***Received:*** *21 November 2019*
***Accepted:*** *17 March 2020*
***Published:*** *15 May 2020*

***Citation:***
*Fu Y, Kang Y and Chen G (2020) Stochastic Resonance Based Visual Perception Using Spiking Neural Networks. Front. Comput. Neurosci. 14:24. doi: 10.3389/fncom.2020.00024*

## INTRODUCTION

The phenomenon of stochastic resonance, discovered by Benzi et al. (1981), is a type of cooperative effect of noise and weak signal under a certain non-linear circumstance, in which the weak signal can be amplified and detected by a suitable amount of noise (Nakamura and Tateno, 2019). Distinct biological and engineering experiments using crayfish (Douglass et al., 1993; Pei et al., 1996), crickets (Levin and Miller, 1996), rats (Collins et al., 1996), humans (Cordo et al., 1996; Simonotto et al., 1997; Borel and Ribot-Ciscar, 2016; Itzcovich et al., 2017; van der Groen et al., 2018), or optical material (Dylov and Fleischer, 2010) suggested that noise might be helpful for stimuli detection and visual perception.

As the visual perception of images of low contrast can find significance in many fields such as medical diagnosis, flight security, and cosmic exploration, theoretical research on stochastic resonance-based contrast enhancement has become an interesting but challenging topic (Yang, 1998; Ditzinger et al., 2000; Sasaki et al., 2008; Patel and Kosko, 2011; Chouhan et al., 2013; Liu et al., 2019; Zhang et al., 2019). Simonotto et al. (1997) used the noisy static threshold model to recover the picture of *Big Ben*, Patel et al. proposed a watermark decoding algorithm using discrete cosine transform and maximum-likelihood detection (Patel and Kosko, 2011), Chouhan et al. explored contrast enhancement based on dynamic stochastic resonance in the discrete wavelet transform domain (Chouhan et al., 2013), and Liu et al. (2019) applied an optimal adaptive bistable array to reduce noise from the contaminated images. It is more and more evident today that stochastic resonance can be utilized as a visual processing mechanism in nervous systems and neural engineering applications, although many theoretical and technical problems remain to be solved.

There exist at least three issues to be clarified. The first issue is about model selection. In the existing literatures, the neuron model commonly used for image enhancing is the static threshold model. Since the threshold neuron is too oversimplified to contain the evolution of the membrane voltage, a more realistic biological neuron model should be considered. The second issue is that one cannot find enough details from the existing algorithms. For example, in those algorithms, there is nearly no explanation of the choice of the critical threshold, across which the pixel value of a black–white image will switch. Note that a suitable threshold is vital for image enhancement, so the second question we have to face is what a critical threshold should be. The last issue is about the adoption of the quantifying index, which helps one to pick out an optimally detected image. A typical assumption is that one knows a clear or clean reference picture, but in most practical applications, how can one get such reference pictures especially when taking photos in darkness?

To answer the above questions, we consider an integrate-and-fire neuron network with global feedback in this paper. Our work can be divided into two parts. The first part is model preparation, where we theoretically observe stochastic resonance based on linear approximation. In the second part, by integrating all the physiological and biophysical aspects of visual perception, we propose an algorithm for boosting the contrast of an image photographed in darkness. We give a criterion for determining the critical threshold and adopt the variance of image to quantify the quality of the enhanced image. Our numerical tests demonstrate that the new algorithm is effective and robust.

## STOCHASTIC RESONANCE IN AN INTEGRATE-AND-FIRE NEURONAL NETWORK

Consider a global feedback biological network of $N$ integrate-and-fire neurons (Lindner and Schimansky-Geier, 2001; Sutherland et al., 2009). The subthreshold membrane potential of each consisting neuron is governed by

$$C\frac{dV_i}{dt} = -g_L(V_i - V_L) + I_i(t) + Cf(t) + Cs(t), 1 \le i \le N \quad (1)$$

where $V_i$ is the membrane potential, $C$ is the capacitance, $g_L$ is the leaky conductance, $V_L$ is the leaky voltage, and the external synaptic input is

$$dI_i(t) = C\sum\nolimits_{k=1}^{p} a_k d\mathrm{Exc}_{n,k}(t) - C\sum\nolimits_{l=1}^{q} b_l d\mathrm{Inh}_{n,l}(t) \quad (2)$$

with the excitatory synaptic current $Exc_{n,k}(t)$ of rate $\lambda_{E,k}$ and the inhibitory synaptic current $Inh_{n,l}(t)$ of rate $\lambda_{I,l}$, both modeled as i.i.d. homogenous Poisson processes, with $a_k (1 \le k \le p)$ and $b_l (1 \le l \le q)$ denoting the efficacies for excitatory and inhibitory synapses, respectively. Assume that each neuron receives a subthreshold cosine signal, $s(t) = \varepsilon\cos(\Omega t)$, from the external environment. By "subthreshold," it means that, in the absence of the synaptic current input (2), the membrane potential cannot cross the given spiking threshold from below (Kang et al., 2005). Here we use $V_r$ to denote the resetting potential; that is, whenever the $i$th membrane potential reaches the threshold $V_{th}$ from below, the $i$th neuron will emit a spike and then the membrane potential will be reset to $V_r$ immediately. Let $t_{i,k}$ be the $k$th spiking instant recorded from the $i$th neuron; then, the output spike train of the $i$th neuron can be described as $y_i(t) = \sum_k \delta(t - t_{i,k})$. In this network, the output spike trains from every consisting neuron are fed back to the $i$th neuron for $1 \le i \le N$ through the synaptic interaction.

$$f(t) = \frac{G}{N}\int_{\tau_D}^{\infty} d\tau \frac{\tau - \tau_D}{\tau_S^2}\exp(-\frac{\tau - \tau_D}{\tau_S})\sum\nolimits_{n=1}^{N} y_n(t - \tau) \quad (3)$$

Here the global feedback interaction is implemented by a convolution of the sum of all the spike trains with a delayed alpha function. We fix the transmission time delay $\tau_D = 1$ and the synaptic time constant $\tau_S = 0.5$. In Equation (3), the feedback strength $G < 0$ indicates inhibitory feedback, $G > 0$ represents excitatory feedback, and Equation (1) turns into a neuron array model for enhancing information transition (Yu et al., 2012) when $G = 0$.

For simplicity, let us drop the subscripts $k$ and $l$ in the rates and the synaptic efficacies, so $\lambda_E = \lambda_I = \lambda$, $p = q$ and $b = ra$, with $r$ being the ratio between inhibitory and excitatory inputs. Invoking diffusion approximation transforms the synaptic current to

$$dI_i(t) = C(ap(1 - r)\lambda dt + a\sqrt{p\lambda(1 + r^2)}dB_i(t))$$

where $(B_1(t), B_2(t), \ldots, B_N(t))$ is $n$ dimensional standard Brownian motions. With Equation (3) available, Equation (1) can be rewritten as

$$\begin{aligned}\frac{d}{dt}V_i &= -\frac{1}{\tau}(V_i - V_L) + ap(1 - r)\lambda \\ &+ a\sqrt{p\lambda(1 + r^2)}\xi_i(t) + f(t) + s(t)\end{aligned} \quad (4)$$

where $\tau^{-1} = g_L/C$ and $\xi_i(t)$ is Gaussian white noise satisfying $\langle\xi_i(t)\rangle = 0$ and $\langle\xi_i(t + s)\xi_j(t)\rangle = \delta(s)$ for $1 \le i, j \le N$.

It has been shown that the firing rate is approximately a linear function of the external input near the equilibrium point (Gu et al., 2019), so we apply the linear approximation theory (Lindner and Schimansky-Geier, 2001; Pernice et al., 2011; Trousdale et al., 2012) to calculate the response of each neuron. Let $\mu = ap(1 - r)\lambda + V_L/\tau$ and $D = \frac{1}{2}a^2p^2\lambda(1 + r^2)$. Regarding each neuron as linear filter of an external perturbation, we rewrite Equation (4) into Equation (5)

$$\begin{aligned}\frac{dV_i(t)}{dt} &= -\frac{1}{\tau}V_i(t) + (\mu + \langle f(t)\rangle_0) + \sqrt{2D}\xi_i(t) \\ &+ \underbrace{(f(t) - \langle f(t)\rangle_0) + s(t)}_{\text{external perturbation}}, 1 \le i \le N.\end{aligned} \quad (5)$$

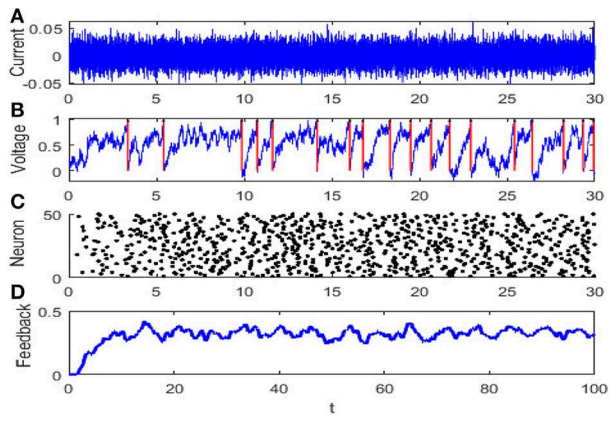


**FIGURE 1 |** The evolution diagram of the integrate-and-fire neuron network: **(A)** diffusion approximation transforming the synaptic current with $r = 1$, **(B)** membrane potential of neuron where the red arrow denotes the discharge time, **(C)** raster plot of the network where every node denotes a spike at a corresponding time and neuron, and **(D)** feedback of the network. The parameters are set as $\mu = 0.8$, $V_T = 1$, $V_R = 0$, $G = 0.5$, $\varepsilon = 0.1$, $\Omega = 1$, $\tau_S = 0.5$, $\tau_D = 1$, $\tau_{ref} = 0$, and $N = 50$.

For simplicity, all of the variables are dimensionless and most of parameters are taken from Lindner et al. (2005), and particularly, time is measured in unit of membrane time constant $\tau$. The dynamical evolution of the network is illustrated in **Figure 1**.

The phenomenon of stochastic resonance is frequently measured by the spectral amplification factor (Liu and Kang, 2018) and the output signal-to-noise ratio (Kang et al., 2005). With the help of the linear approximation theory, both the spectral amplification factor and the output signal-to-noise ratio for the homogeneous network can be explicitly attained. The spectral amplification factor is defined as the ratio of the power denoted by the delta-like spike in the output spectrum at $\pm\Omega$ over the power of the input signal, namely,

$$SAF = \frac{\pi\varepsilon^2\left|A(\Omega,\bar{\mu},D)\right|^2}{\left|1-GA(\Omega,\bar{\mu},D)F(\Omega)\right|}/\pi\varepsilon^2 = \frac{\left|A(\Omega,\bar{\mu},D)\right|^2}{\left|1-GA(\Omega,\bar{\mu},D)F(\Omega)\right|} \quad (6)$$

while the signal-to-noise ratio, defined as the ratio of the power of the signal component over the background noise, is given by

$$SNR = \lim_{\Delta\omega\to 0}\frac{\int_{\Omega\text{-}\Delta\omega}^{\Omega+\Delta\omega} G_{yy}(\omega)d\omega}{S_2(\Omega)} = \frac{N\pi\varepsilon^2|A(\Omega,\bar{\mu},D)|^2}{S_0(\Omega,\bar{\mu},D)} \quad (7)$$

where $A$ is the linear susceptibility, $\bar{\mu}$ is the base current, $F(\omega) = e^{i\omega\tau_D}/(1-i\omega\tau_S)^2$ is the Fourier transform of the kernel in Equation (2) and $S_0(\omega,\mu,D,V_T)$ is the fluctuating spectral density of the unperturbed system. $G_{yy}(\omega)$ is power spectral density of output spike train, which consists of the signal component $S_1(\omega)$ and the fluctuation component $S_2(\omega)$. Actually, within the range of linear response, the power spectrum $G_{yy}(\omega)$ is a sharp power peak at the signal frequency riding over the spectral density of fluctuations, as shown in **Figure 2**. The detailed derivations of power spectral density $G_{yy}(\omega)$, spectral amplification factor *SAF* and output signal-to-noise ratio *SNR* are further described in **Appendix**.

Equation (6) demonstrates that the spectral amplification factor is independent of the network size, whereas Equation (7) shows that the signal-to-noise ratio is proportional to the size. When comparing with the simulation results, **Figure 3** shows that the theoretical results tend to be an overestimated approximation, but the overestimation is reduced as the network size increases. For this reason, the network size is fixed to be large enough in **Figures 4**, **5** so that the theoretical and simulation results are accurately matched.

Since the dependence of the spectral amplification factor or the signal-to-noise ratio on noise intensity is non-monotonic, one can conclude that stochastic resonance occurs for the given parameters in **Figure 3**. **Figure 4** further shows the image of the signal-to-noise ratio on the two-parameter plane of noise intensity and global feedback strength. From this figure, it can be seen that, for fixed feedback strength, the existence of a sharp peak indicates stochastic resonance in the global feedback network, while for fixed noise intensity, the signal-to-noise ratio is a growing function of the feedback strength, which suggests the larger feedback strength is beneficial for resonant effect. Here we emphasize that the effect of the inhibitory feedback on the weak signal amplification is different from its effect on the intrinsic oscillation measure in Lindner et al. (2005) since these are two kinds of different synchronization. Phenomenologically, the former is the synchronization behavior of the external weak signal and the firing activity caused by noise, while the latter is the synchrony among the population neurons, and the difference in quantifying indexes directly leads to distinct observation. Thus, from the viewpoint of weak signal detection, one can say that the excitatory neural feedback is better than the inhibitory neural feedback.

Note that, in real neural activities, the spiking threshold may vary following the changing circumstance (Destexhe, 1998; Taillefumier and Magnasco, 2013), so it makes sense to consider the effect of the threshold on the population activity. By Equation (7), one has

$$\frac{\partial SNR}{\partial V_T} = S_0^{-2}N\pi\varepsilon^2\left(2\mathrm{Re}\left(A^*\frac{\partial A}{\partial V_T}\right)S_0 - |A|^2\frac{\partial S_0}{\partial V_T}\right), \quad (8)$$

where

$$\frac{\partial A}{\partial V_T} = \frac{i\omega}{\sqrt{D}(i\omega-1)}\left(\frac{\frac{\partial r}{\partial V_T}\tilde{D}_{i\omega-1}\left(\frac{\mu-V_T}{\sqrt{D}}\right)-e^{\gamma}\tilde{D}_{i\omega-1}\left(\frac{\mu-V_R}{\sqrt{D}}\right)}{\tilde{D}_{i\omega}\left(\frac{\mu-V_T}{\sqrt{D}}\right)-e^{\gamma}e^{i\omega\tau_R}\tilde{D}_{i\omega}\left(\frac{\mu-V_R}{\sqrt{D}}\right)} + r\frac{-\frac{1}{\sqrt{D}}\frac{\partial\tilde{D}_{i\omega-1}}{\partial V_T}\Big|_{\frac{\mu-V_T}{\sqrt{D}}}-e^{\gamma}\frac{(\mu-V_T)}{2D}\tilde{D}_{i\omega-1}\left(\frac{\mu-V_R}{\sqrt{D}}\right)}{\tilde{D}_{i\omega}\left(\frac{\mu-V_T}{\sqrt{D}}\right)-e^{\gamma}e^{i\omega\tau_R}\tilde{D}_{i\omega}\left(\frac{\mu-V_R}{\sqrt{D}}\right)}\right)$$

$$+\frac{i\omega}{\sqrt{D}(i\omega-1)}\left(r\frac{\left(-\frac{1}{\sqrt{D}}\frac{\partial\tilde{D}_{i\omega}}{\partial V_T}\Big|_{\frac{\mu-V_T}{\sqrt{D}}}-e^{\gamma}e^{i\omega\tau_R}\frac{(\mu-V_T)}{2D}\tilde{D}_{i\omega-1}\left(\frac{\mu-V_R}{\sqrt{D}}\right)\right)\left(\tilde{D}_{i\omega-1}\left(\frac{\mu-V_T}{\sqrt{D}}\right)-e^{\gamma}\tilde{D}_{i\omega-1}\left(\frac{\mu-V_R}{\sqrt{D}}\right)\right)}{\left(\tilde{D}_{i\omega}\left(\frac{\mu-V_T}{\sqrt{D}}\right)-e^{\gamma}e^{i\omega\tau_R}\tilde{D}_{i\omega}\left(\frac{\mu-V_R}{\sqrt{D}}\right)\right)^2}\right)$$

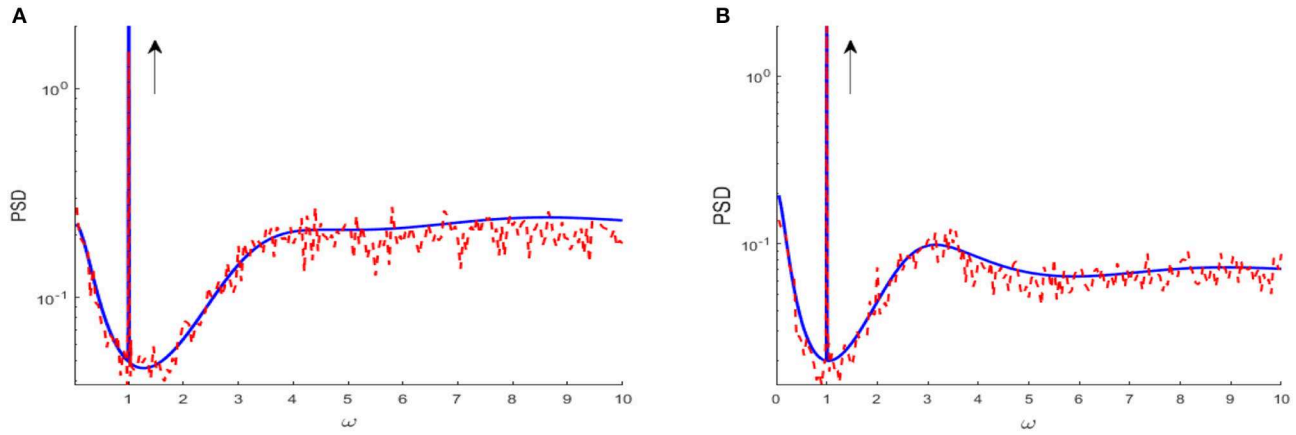


**FIGURE 2 |** Power spectrum density obtained from linear approximation (blue solid curve), compared to simulation (red dash curve) for $D = 0.01$ **(A)** and $D = 0.1$ **(B)**, respectively. The parameters are set as $\mu = 0.8$, $V_T = 1$, $V_R = 0$, $G = 0.5$, $N = 3$, $\varepsilon = 0.1$, $\Omega = 1$, $\tau_S = 0.5$, $\tau_D = 1$, $\tau_{ref} = 0$, and $\tau = 1$. The black arrow indicates the spike spectral line at the driving frequency, and the remaining part characterizes the spectral density of environmental fluctuations. Clearly, the spectral line of the driving signal is riding over the fluctuation spectral density. The explanation of figure and derivation of power spectrum density are displayed in **Appendix**.

and

$$\frac{\partial S_0}{\partial V_T} = \frac{\partial r}{\partial V_T}\frac{|\tilde{D}_{i\omega}(\frac{\mu-v_T}{\sqrt{D}})|^2 - e^{2\gamma}|\tilde{D}_{i\omega}(\frac{\mu-v_R}{\sqrt{D}})|^2}{|\tilde{D}_{i\omega}(\frac{\mu-v_T}{\sqrt{D}}) - e^{\gamma}e^{i\omega\tau_R}\tilde{D}_{i\omega}(\frac{\mu-v_R}{\sqrt{D}})|^2} + r\frac{2\,\mathrm{Re}\left(-\frac{1}{\sqrt{D}}\left(\tilde{D}_{i\omega}(\frac{\mu-v_T}{\sqrt{D}})\right)^*\frac{\partial\tilde{D}_{i\omega}}{\partial V_T}\Big|_{\frac{\mu-v_T}{\sqrt{D}}}\right) - e^{2\gamma}\frac{\mu-V_T}{D}|\tilde{D}_{i\omega}(\frac{\mu-v_R}{\sqrt{D}})|^2}{|\tilde{D}_{i\omega}(\frac{\mu-v_T}{\sqrt{D}}) - e^{\gamma}e^{i\omega\tau_R}\tilde{D}_{i\omega}(\frac{\mu-v_R}{\sqrt{D}})|^2}$$
$$+ r\frac{2\,\mathrm{Re}\left(\left(\tilde{D}_{i\omega}(\frac{\mu-v_T}{\sqrt{D}}) - e^{\gamma}e^{i\omega\tau_R}\tilde{D}_{i\omega}(\frac{\mu-v_R}{\sqrt{D}})\right)^*\left(\tilde{D}_{i\omega}(\frac{\mu-v_T}{\sqrt{D}})\cdot\frac{\partial\tilde{D}_{i\omega}}{\partial V_T}\Big|_{\frac{\mu-v_T}{\sqrt{D}}}\cdot(-\frac{1}{\sqrt{D}}) - e^{\gamma}e^{i\omega\tau_R}\tilde{D}_{i\omega}(\frac{\mu-v_R}{\sqrt{D}})\frac{\mu-V_T}{2D}\right)\right)}{\left|\tilde{D}_{i\omega}(\frac{\mu-v_T}{\sqrt{D}}) - e^{\gamma}e^{i\omega\tau_R}\tilde{D}_{i\omega}(\frac{\mu-v_R}{\sqrt{D}})\right|^4}$$

with Re( · ) being the real part of a complex value. Here, the Whittaker notation $\tilde{D}_a$(Abramovitz and Stegun, 1964) is used for the parabolic cylinder function, with the recursion property $\tilde{D'}_a(x) + \frac{1}{2}x\tilde{D}_a(x) - a\tilde{D}_{a-1}(x) = 0$ and

$$\frac{\partial r}{\partial V_T} = \frac{-\frac{r^2\sqrt{\pi}}{\sqrt{2D}}\cdot\exp\left(\left(\frac{\mu+Gr-V_T}{\sqrt{2D}}\right)^2\right)erfc\left(\frac{\mu+Gr-V_T}{\sqrt{2D}}\right)}{1+\frac{Gr^2\sqrt{\pi}}{\sqrt{2D}}\left(\exp\left(\left(\frac{\mu+Gr-V_R}{\sqrt{2D}}\right)^2\right)erfc\left(\frac{\mu+Gr-V_R}{\sqrt{2D}}\right) - \exp\left(\left(\frac{\mu+Gr-V_T}{\sqrt{2D}}\right)^2\right)erfc\left(\frac{\mu+Gr-V_T}{\sqrt{2D}}\right)\right)}$$

The evolution of the signal-to-noise ratio [Equation (7)] and its partial derivative [Equation (8)] obtained *via* the threshold is shown in **Figures 5A**,**B**, respectively. The monotonical decrease in the signal-to-noise ratio suggests that a smaller threshold is better for weak signal detection. Moreover, from these figures, one can also see that an increasing distance between the base current and the firing threshold will lead to a reduced signal-to-noise ratio, as disclosed by Kang et al. (2005). As a result, the minimum distance between the base current and the firing threshold should be an important reference in designing visual perception applications of the global feedback network.

## STOCHASTIC RESONANCE BASED IMAGE PERCEPTION

We have systematically disclosed the phenomenon of stochastic resonance from the viewpoint of model investigation, and in this section, we wish to propose an algorithm for visual perception under the guidance of the above theoretical results. In fact, it is the theoretical evidence of SR in the integrate-and-fire neuron network in section stochastic resonance in an integrate-and-fire neuronal network that motivates us to do the application exploration. If noise at a certain level can amplify a weak harmonic signal *via* stochastic resonance, then noise of suitable amount can very likely enhance a more realistic weak signal such as the image of low contrast *via* aperiodic stochastic resonance.

In stochastic resonance, since the external weak signal is harmonic, one can use the spectral amplification factor or the output signal-to-noise ratio as quantifying index through frequency matching, while in aperiodic stochastic resonance, the external weak signal is aperiodic, so one has to resort to some coherence measure to describe the involved shape matching, as confirmed in neural information coding (Parmananda et al., 2005), hearing enhancement (Zeng et al., 2000). For the picture

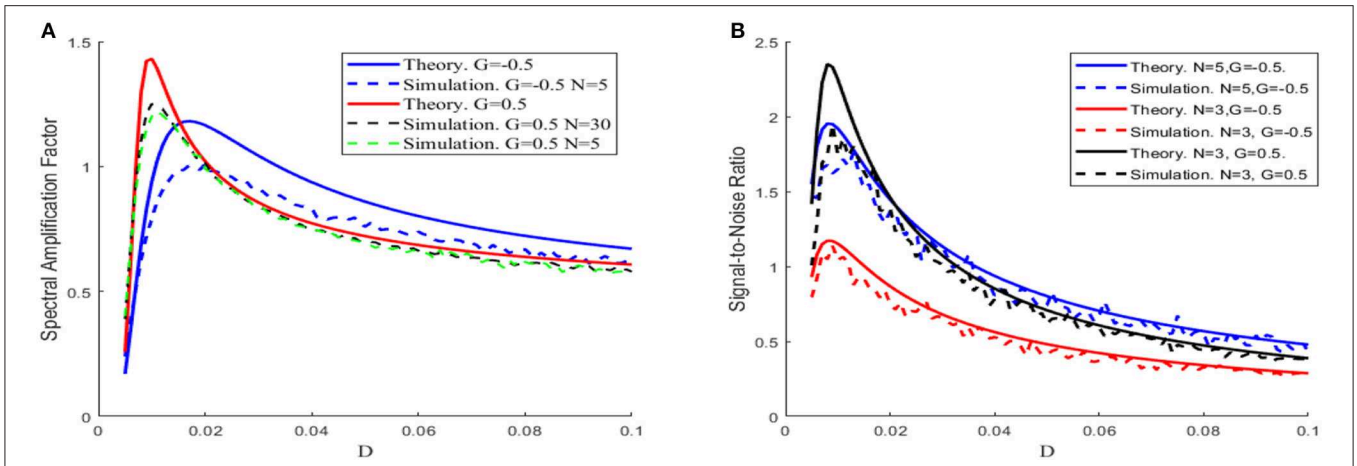


**FIGURE 3 |** The theoretical (solid) and simulation (dash) results of spectral amplification factor **(A)** and signal-to-noise ratio **(B)** vs. noise intensity, with $\mu = 0.8, V_T = 1, V_R = 0, \varepsilon = 0.1, \Omega = 1, \tau_S = 0.5, \tau_D = 1, \tau_{ref} = 0$, and $\tau = 1$, under different sizes of the network $N$ and different feedback strengths $G$, respectively.

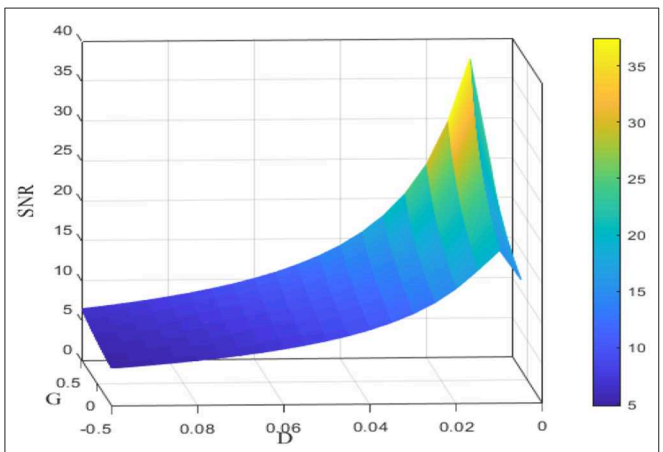


**FIGURE 4 |** Signal-to-noise ratio for different noise intensities $D$ and feedback strengths $G$. Data were obtained by the numerical simulation of a network, with $\mu = 0.8$, $V_T = 1$, $V_R = 0$, $N = 50$, $\varepsilon = 0.1$, $\Omega = 1$, $\tau_S = 0.5$, $\tau_D = 1$, $\tau_{ref} = 0$, and $\tau = 1$. It is clear that for fixed $G$, the signal-to-noise ratio shows a rise before fall as a function of noise intensity $D$; for fixed $D$, the signal-to-noise ratio is an increasing function of the feedback strength $G$.

of low contrast, its contrast can be changed by noise and will attain to a maximum when the phenomenon of aperiodic stochastic resonance occurs; thus, we use the variance of image as a quantifying index as explained below. Even though a difference exists in quantifying index between stochastic resonance and aperiodic stochastic resonance, we can still use the results obtained from the model investigation as guidance. The numerical results in section stochastic resonance in an integrate-and-fire neuronal network show that positive feedback strength and low threshold are beneficial factors for observing the effect of stochastic resonance, and therefore we will take the two factors into account in the following algorithm design.

With the theoretical guidance in mind, we now start to present the algorithm for enhancing the image of low contrast. By the term dark image or image of low contrast, we mean that the picture is taken in a dark surrounding and cannot be detected at first sight. We put the new algorithm under the frame of the fundamental process for visual formation (Purves, 2011; Li, 2019): the photoreceptors in the retina receive the light and convert it into electrical signals, which is called encoding process, and then the signals are processed ultimately in the visual cortex, which is called decoding and integration process. Our algorithm is expounded into three steps, as shown in the flow chart in **Figure 6**.

## Step 1. Encoding

When light enters the eye, the retina will convert the optical signal into electrical signal first. There are two kinds of photoreceptors in the retina, which are called rods and cones, respectively. The cones are active at bright light conditions and capable of color vision, while the rods are responsible for scotopic vision but cannot perceive color. As a result, human can capture the shape of the object in dim surroundings. We use the global feedback network [Equation (5)] of $K$ integrate-and-fire neurons to simulate the perceptive process for rod cells. The membrane potential $V_i^{m,n}$ for each neuron is governed by

$$\frac{dV_i^{m,n}(t)}{dt} = -\frac{1}{\tau}V_i^{m,n}(t) + U(m,n) + \sqrt{2D}\eta_i^{m,n}(t) + f^{m,n}(t), 1 \le i \le K \quad (9)$$

where the superscript corresponds to the pixels of the image and the subscript corresponds to the neurons, $U(m,n) \in [0, 1]$ denotes the brightness of the input image, the Gaussian white noise $\eta_i^{m,n}(t)$ satisfying $\left\langle \eta_i^{m,n}(t+s)\eta_j^{m,n}(t) \right\rangle = \delta(s)\delta(i-j)$ is assumed to describe the fluctuation arising from the rhythms and the distribution of the rod cells along the retina, and $f^{m,n}(t)$ is the same global feedback function as in Equation (3). Upon $V_i^{m,n}$ reaching the threshold $V_{th}$ from below, the $i$th neuron will emit an action potential at once and then the membrane potential is immediately reset to $V_r$.

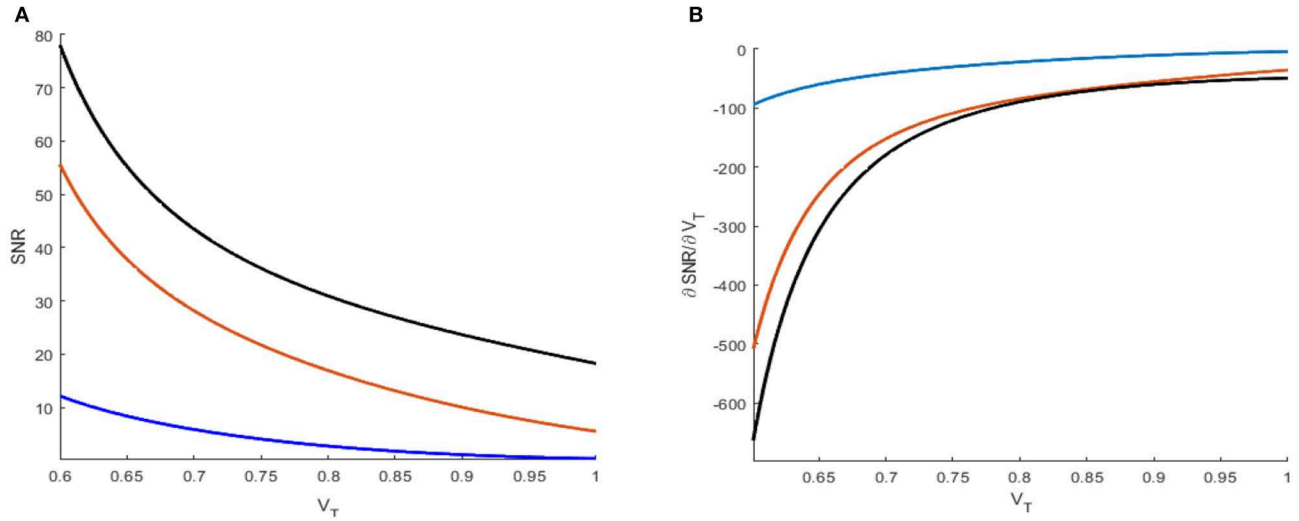


**FIGURE 5 |** Signal-to-noise ratio **(A)** and its partial derivative with respect to threshold $V_T$ **(B)** under different reference currents $\mu = 0.2$ (blue), 0.5 (orange), and 0.8 (black). The network size is $N = 50$.

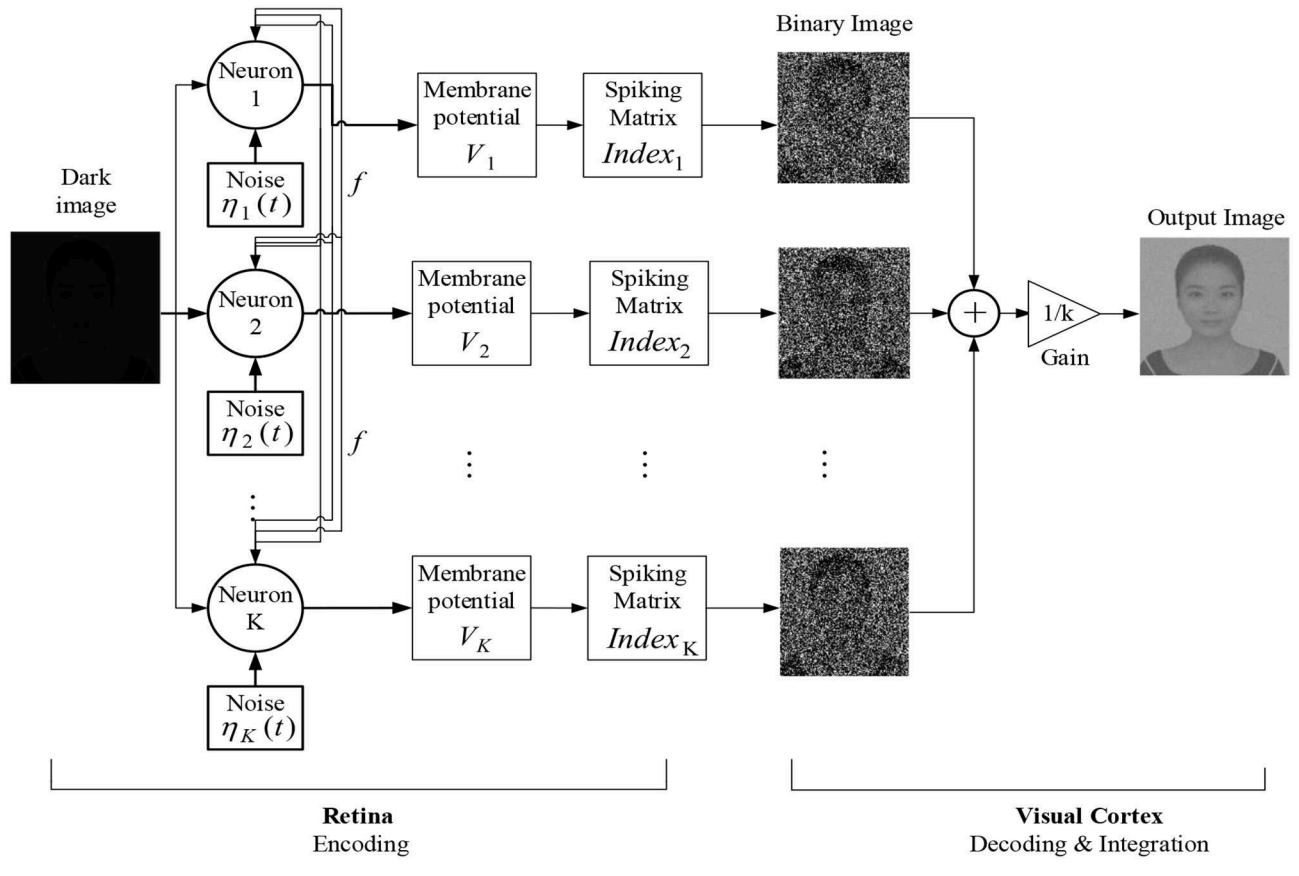


**FIGURE 6 |** Schematic diagram of the dark image enhancement algorithm based on the global feedback integrate-and-fire network.

## Step 2. Decoding and Integration

The coming information from the rod cells is decoded into a binary image within the visual cortex. We explain it from two aspects. Firstly, the carrier of neural information transmission is spike impulse, so the encoded information should be in the form of a spike train instead of the continuous membrane potential. Secondly, note that rod cells play a minor role in color vision, which actually leads to loss of color in dim light (Purves, 2011; Owsley et al., 2016), so it is reasonable to assume that all the receiving spike trains can be transformed into a binary image. Let matrix $(Index_i)_{M\times N}$ store the spiking information of the $i$th neuron at the encoding stage. Then, the corresponding binary image matrix $(Pic_i)_{M\times N}$ decoded by the $i$th neuron can be written as

$$Pic_i(m,n) = \begin{cases} 0, & Index_i(m,n) = 0; \\ 255, & Index_i(m,n) = 1. \end{cases} \tag{10}$$

With the decoded information from each neuron available, the visual cortex, as command center, will integrate all the information to form an overall gray image, which should be the picture we finally see in the dark surrounding. The idea of integration is inspired by boosting (Friedman, 2002). If each binary image is regarded as the output of the weak learner, the combination of the weak learners will be a strong learner and produce the gray image. We assume that the integration is in the way of linear superposition, namely,

$$Pic(m,n) = \frac{1}{N}\sum_{i=1}^{k} Pic_i(m,n) \tag{11}$$

where $(Pic)_{M\times N}$ represents the integrated image.

We wish to put more emphasis on the validity of using the principle of stochastic resonance in our perception algorithm. It is well-known that noise is prevalent at the cellular level, and the level of the fluctuation in a neural system can be self-adjusted (Faisal and Selen, 2008; Durrant et al., 2011). What is more, distinct biophysical experiments (Douglass et al., 1993; Collins et al., 1996; Cordo et al., 1996; Levin and Miller, 1996; Pei et al., 1996; Borel and Ribot-Ciscar, 2016; Itzcovich et al., 2017; van der Groen et al., 2018) have shown that the benefit of noise can be utilized by biology. Thus, we assume that the human brain can select the perceived image of maximal contrast by means of the principle of stochastic resonance. The perceptive function of the brain is realized by neuron population, while the effect of stochastic resonance can be enhanced by uncoupled array or coupled ensemble; thus, our visual perception algorithm should be of some biological rationality.

The procedure of the new algorithm is carried out in one unit of time by Euler integration with a step length of 0.01 time unit for all the detection experiments. The dark-input images were photos directly taken in a dark environment, such as that in **Figure 7A**, or artificially designed by compressing the original bright images into dark inputs, as shown in **Figures 7D, G, J**. The recognized images of the best quality, namely, the best enhanced images, are shown in the second column. During the experiments, it was found that some subtle key details, such as the quantifying index, the firing threshold, and the global feedback strength, need to be further explained.

### Quantifying Index

To evaluate the quality of an image, in the image processing literature, the most frequently used indexes are the peak signal-to-noise ratio and the mean-square error, where some known reference images are required. The perceptual quality metric (PQM) (Wang et al., 2002), another quantifying index used in visual perception, can skillfully evade the reference images. The more that PQM is close to 10, the better the quality of the image is (Susstrunk and Winkler, 2003), but it tends to become flat near the optimal value, as shown in **Figure 7**. Since the flatness is not favorable for picking out the optimal noise intensity to get the best enhanced image, the objective here is to find a better quantifying index to assess the perceptual quality. The new index is found to be the variance of image. For a given image $U_{M\times N}$, the variance is defined by

$$Var\,(U) = \frac{1}{(M\times N)^2}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(U\,(m,n) - \bar{U}\right)^2,$$

where $\bar{U}$ is the mean of the pixel matrix $U_{M\times N}$. The reason lies in the fact that this variance can reflect the heterogeneity among all the pixels. Intuitively, for a low-contrast image, the value of the variance will be quite low, but for a high-contrast image, the variance should take a much higher value. **Figure 7** indeed verifies this reasoning. First of all, when the PQM is closer to 10, the variance curve will be nearer its peak. That is, the variance has the same capacity to identify which picture is the best in this task. Secondly, there is a sharp peak in the variance vs. the noise intensity curve so that one can easily detect an image with the best quality, namely, the best enhanced image. This is an advantage of the variance measure for the perceptual quality over the PQM measure, as shown in the third column of **Figure 7**. In addition, we note that the mean of the image is not suitable to be used as quantifying index. In fact, the mean of the image measures the luminance of an image, and it takes different values from the dark input and the best enhanced image to the blurred image due to excessive noise, but its value monotonically grows as noise intensity increases, as shown in **Figure 8**; thus, the mean is incapable of identifying the image with the best contrast as well. Undoubtedly, the comparison further emphasizes the applicability of the variance in visual perception.

### Firing Threshold

In real cortical activities, neurons can adopt a self-adaptive threshold strategy dependent on varying environments (Destexhe, 1998; Taillefumier and Magnasco, 2013) since the threshold has a direct impact on the neural electronic activity. We find that the threshold also has a large impact on the performance of the visual perception algorithm in **Figure 9**. The picture clearly shows that the choice of a suitable firing threshold is vital for the quality of the perceived image. Here the threshold is chosen according to the following rule. Firstly, find the frequency histogram of the dark image and denote the maximum pixel of the normalized histogram as max ( $U$). Then, define the

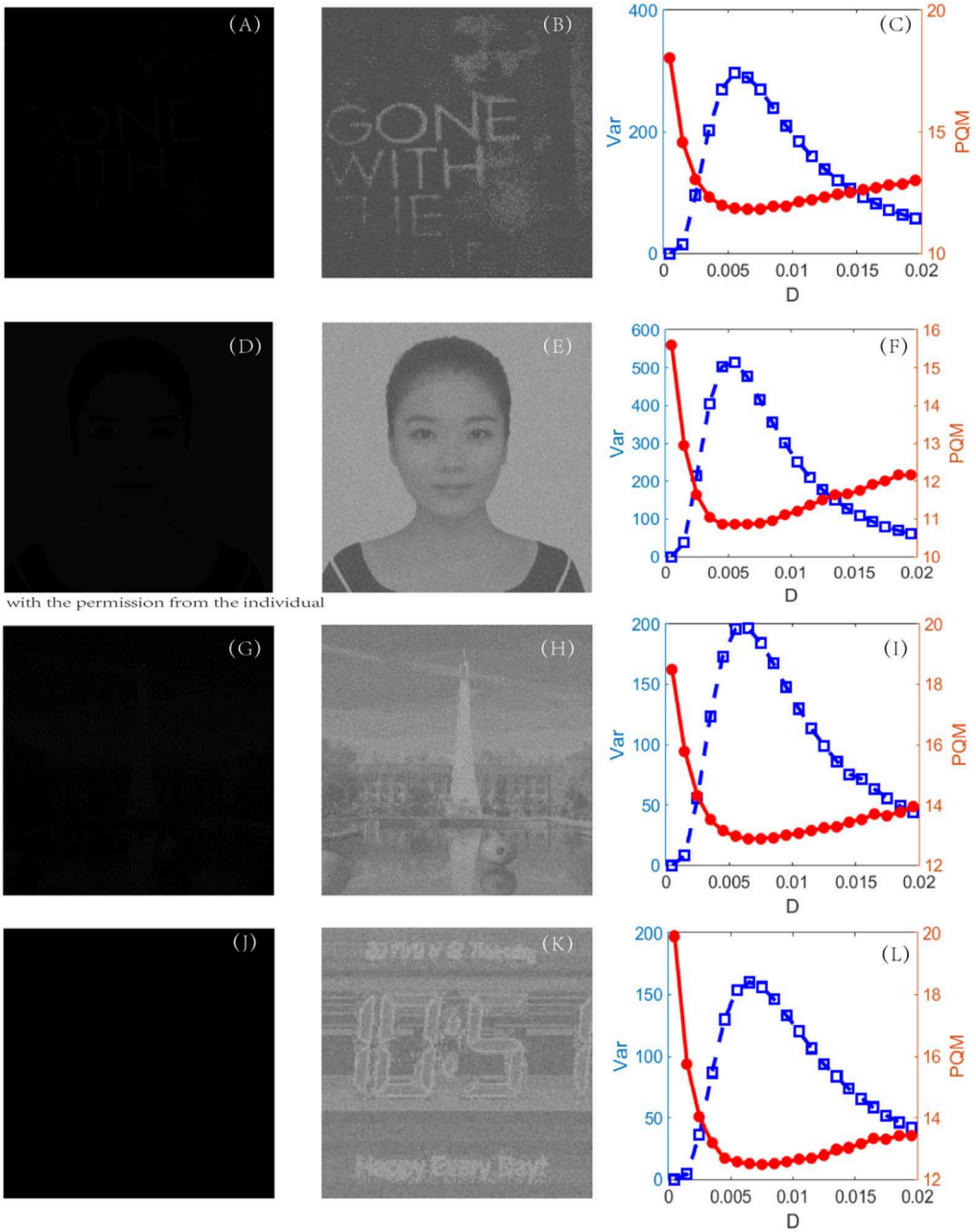


**FIGURE 7 |** First column (**A**, **D**, **G** and **J**): original dark-input images; second column (**B**, **E**, **H** and **K**): enhanced images with best quality; and third column (**C**, **F**, **I** and **L**): dependence of variance (blue, square) and PQM (red, dot) on noise intensity, for each experiment. The parameters are set as $k = 1{,}000$, $V_T = 0.1$, $V_R = 0$, $G = 0.12$, $\tau_s = 0.05$, $\tau_D = 0.01$, and $\tau = 1$. For each experiment, the location of the peak of variance is always near the location of the bottom of the PQM, indicating that variance helps in recognizing the best-quality image.

threshold by $V_{th} = 10^{-1} ceil\left(10 \max(U)\right)$, where $ceil(\cdot)$ is the rounding function toward positive infinity. For example, the maximum pixel of the image in **Figure 7D** is $\max(U) = 0.05$, as seen from **Figure 8A1**; accordingly, the threshold is taken as 0.1. It is worthy to remark that this kind of choice can guarantee that the distance between the base current and the firing threshold is minimized as far as possible, as suggested by the discussion following **Figure 5**.

## Feedback Strength

In section stochastic resonance based image perception, it was demonstrated that, when the global feedback changes from the inhibitory type into the excitatory type, the peak of the signal-to-noise ratio can be improved as shown in **Figure 4**. This theoretical observation encourages us to check the influence of the feedback strength of the encoding stage on the enhanced images as illustrated in **Figure 10**. Evidently, the excitatory

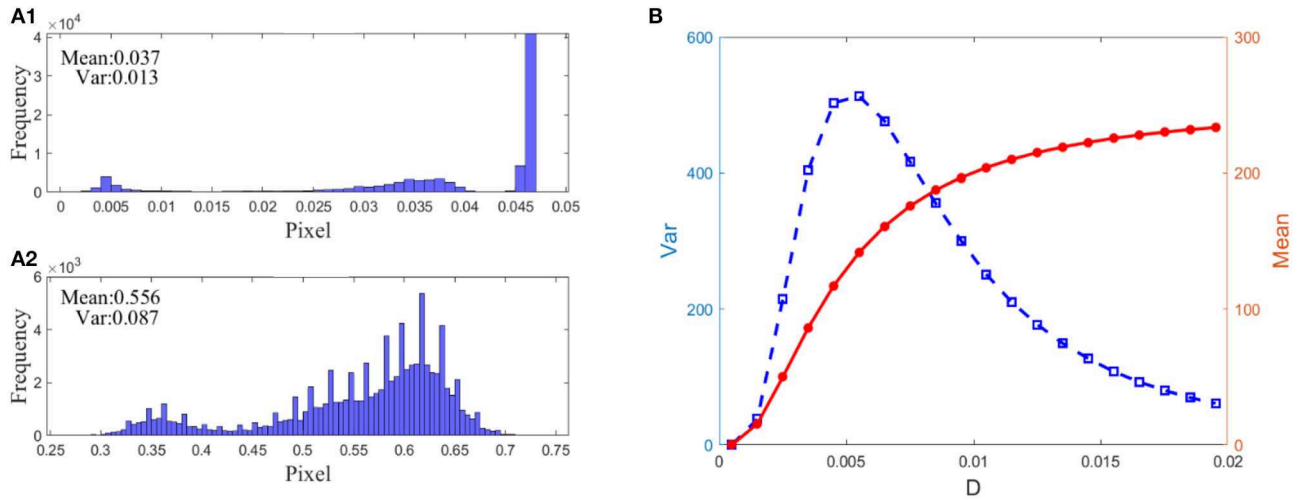


**FIGURE 8 |** Demonstration of the advantage of variance over mean: **(A)** 1: normalized histogram of the dark input image in **Figure 7D**, **(A)** 2: normalized histogram of the best enhanced image in **Figure 7E**, **(B)** Dependence of the variance (blue square) and the mean (red dot) on noise intensity. As seen from **(A)** 1 and 2, the dark-input image and the best enhanced image differ in the histogram. Nevertheless, it is not applicable for picking up the image of the best contrast since the mean grows monotonically **(B)** as noise intensity increases, even I when the contrast of the detected image deteriorates again. By contrast, the bell-shaped change of variance is suitable.

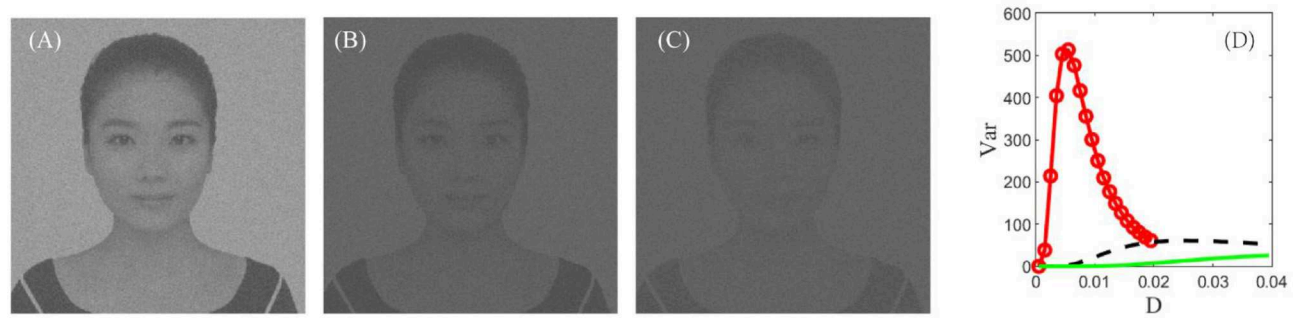


**FIGURE 9 |** The best enhanced image under $V_T = 0.1$ **(A)**, $V_T = 0.2$ **(B)**, and $V_T = 0.3$ **(C)**, respectively. **(D)** Dependence of variance on the noise intensity under $V_T = 0.1$ (red, circle), $V_T = 0.2$ (black, dash), and $V_T = 0.3$ (green, solid). Other parameters are the same with **Figure 7**. This figure demonstrates that the best enhanced image is dependent on the spiking threshold, and thus choosing a suitable threshold is vital for image detection.

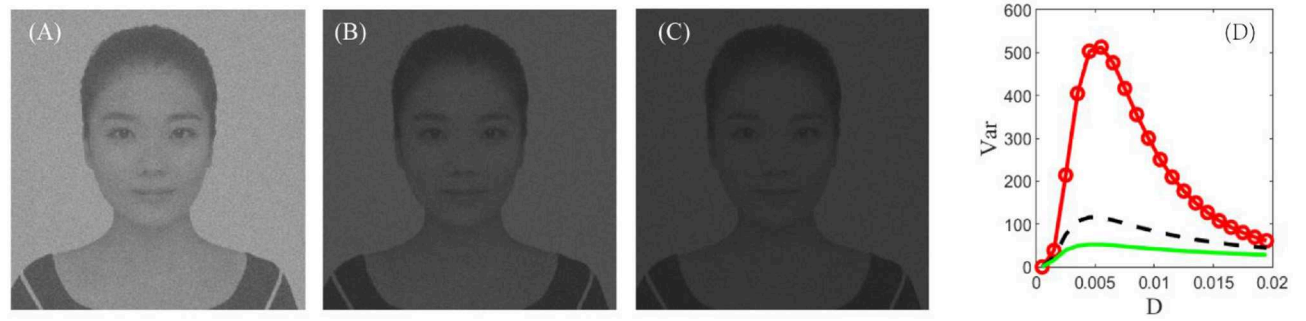


**FIGURE 10 |** The best enhanced image under $G = 0.12$ **(A)**, $G = 0$ **(B)**, and $G = -0.12$ **(C)**, respectively. **(D)** Dependence of variance on the noise intensity under $G = 0.12$ (red, circle), $G = 0$ (black, dash), and $G = -0.12$ (green, solid). Other parameters are the same with **Figure 7**. Clearly, an excitatory feedback is the best among all the types of global feedback, and this implies that different rods should cooperate with each other when facing the same task.

feedback leads to the best enhancement among all the cases, and thus one can fix the feedback strength to be positive as shown in **Figure 7**. We emphasize that this finding does not deny that inhibition plays an important role in visual perception (Roska et al., 2006). As we know, both excitation and inhibition exist in the retina (Rizzolatti et al., 1974). We assume that excitation is reflected by step 1 of our algorithm. That is, different neurons in the retina help each other in detecting the same target and exhibit the cooperative effect in a general homogenous network at the encoding stage. This cooperative effect helps the individuals of the network spike regularly, and certainly this effect is consistent with the description in Brunel (2000) which states that the neurons exhibit a regular state when excitation dominates inhibition.

## CONCLUSION

We have proposed a visual perception algorithm by combining the stochastic resonance principle of a global feedback network of integrate-and-fire neurons with the biophysical process for visual formation. The results can be summarized from the two closely related aspects. From the aspect of model investigation, we applied the technique of linear approximation and direct simulation to disclose the phenomenon of stochastic resonance in a global feedback network of integrate-and-fire neurons. It is demonstrated that both the spectral amplification factor and the output signal-to-noise ratio obtained from linear approximation are accurate when the size of the network is sufficiently large. Then, using the results derived from linear approximation, we found that positive feedback strength is beneficial for boosting the output signal-to-noise ratio, while a decreasing distance between the base current and the firing threshold can enhance the resonance effect. The theoretical observations are new, and they are also helpful for us to understand the working mechanism in rod neurons.

From the aspect of algorithm design, by applying the global feedback network (5) of integrate-and-fire neurons to simulate the perceptive process for rod cells, we have developed a novel visual perception algorithm. In the algorithm, the firing threshold is so critical that an inappropriate choice will lead to inefficiency in image enhancement. Under the inspiration of the theoretical finding that a decreasing distance between the base current and the firing threshold is favorable for stochastic resonance, we have proposed an explicit expression of a suitable firing threshold by referring to the histogram of the dark images. Moreover, we creatively introduced the variance of image rather than the perceptual quality metric as a more effective measure to examine the quality of the enhanced images. Massively numerical tests have shown that the biologically inspired algorithm is effective and powerful. We emphasize that the visual perception algorithm is a dynamical system based algorithm. We hope that it can be applied to relevant fields such as medical diagnosis, flight security, and cosmic exploration, where dark images are common. The algorithm also offers a good example of how the dynamical system research guides the neural engineering application. Following the success of this research, we will start to explore more interesting and important problems, such as the recovery of incomplete images, in the near future.

## DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

## AUTHOR CONTRIBUTIONS

YK guided and sponsored the research. YF did the simulation and algorithm implementation. YF worked out the initial draft, YK rewrote it, and GC made contribution in language polishing and general guide. The contributions from all the authors are important.

## FUNDING

This work was financially supported by the National Natural Science Foundation under Grant No. 11772241.

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fncom.2020.00024/full#supplementary-material

## REFERENCES

Abramovitz, M., and Stegun, I. A. (1964). *Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables*. U.S. Department of Commerce, NIST.

Benzi, R., Sutera, A., and Vulpiani, A. (1981). The mechanism of stochastic resonance. *J. Phys. A*. 14, L453–L457. doi: 10.1088/0305-4470/14/11/006

Borel, L., and Ribot-Ciscar, E. (2016). Improving postural control by applying mechanical noise to ankle muscle tendons. *Exp. Brain Res*. 234, 2305–2314. doi: 10.1007/s00221-016-4636-2

Brunel, N. (2000). Dynamics of sparsely connected networks of excitatory and inhibitory spiking neurons. *J. Comput. Neurosci*. 8, 183–208. doi: 10.1023/A:1008925309027

Chouhan, R., Kumar, C. P., Kumar, R., and Jha, R. K. (2013). Contrast enhancement of dark images using stochastic resonance in wavelet domain. *Int. J. Mach. Learn. Comput*. 2, 711–715. doi: 10.7763/IJMLC.2012.V2.220

Collins, J. J., Imhoff, T. T., and Grigg, P. (1996). Noise enhanced information transmission in rat SA1 cutaneous mechanoreceptors via a periodic stochastic resonance. *J. Neurophysiol.* 76, 642–645. doi: 10.1152/jn.1996.76.1.642

Cordo, P., Inglis, J. T., Verschueren, S., and Collins, J. J., et. al. (1996). Noise in human muscle spindles. *Nature* 383, 769–770. doi: 10.1038/383769a0

Destexhe, A. (1998). Spike-and-wave oscillations based on the properties of GABA(B) receptors. *J. Neurosci*. 18, 9099–9111. doi: 10.1523/JNEUROSCI.18-21-09099.1998

Ditzinger, T., Stadler, M., Strüber, D., and Kelso, J. A. S. (2000). Noise improves three-dimensional perception: stochastic resonance and other impacts of noise to the perception of autostereograms. *Phys. Rev. E*. 62, 2566–2575. doi: 10.1103/PhysRevE.62.2566

Douglass, J. K., Wilkens, L., Pantazelou, E., and Moss, F. (1993). Noise enhancement of information transfer in crayfish mechanoreceptors by stochastic resonance. *Nature* 365, 337–340. doi: 10.1038/365337a0

Durrant, S., Kang, Y., Stocks, N., and Feng, J. (2011). Suprathreshold stochastic resonance in neural processing tuned by correlation. *Phys. Rev. E.* 84:011923. doi: 10.1103/PhysRevE.84.011923

Dylov, D. V., and Fleischer, J. W. (2010). Nonlinear self-filtering of noisy images via dynamical stochastic resonance. *Nat. Photonics* 4, 323–328. doi: 10.1038/nphoton.2010.31

Faisal, A., and Selen, L. (2008). Noise in the nervous system. *Nat. Rev. Neurosci*. 9, 292–303. doi: 10.1038/nrn2258

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat. Data Anal*. 38, 367–378. doi: 10.1016/S0167-9473(01)00065-2

Gu, Q. L., Li, S., Dai, W. P., Zhou, D., and Cai, D. (2019). Balanced active core in heterogeneous neuronal networks. *Front. Comput. Neurosci*. 12:109. doi: 10.3389/fncom.2018.00109

Itzcovich, E., Riani, M., and Sannita, W. G. (2017). Stochastic resonance improves vision in the severely impaired. *Sci. Rep*. 7:12840. doi: 10.1038/s41598-017-12906-2

Kang, Y., Xu, J., and Xie, Y. (2005). Signal-to-noise ratio gain of a noisy neuron that transmits subthreshold periodic spike trains. *Phys. Rev. E* 72:021902. doi: 10.1103/PhysRevE.72.021902

Levin, J. E., and Miller, J. P. (1996). Broadband neural encoding in the cricket cercal sensory system enhanced by stochastic resonance. *Nature* 380, 165–168. doi: 10.1038/380165a0

Li, Z. P. (2019). A new framework for understanding vision from the perspective of the primary visual cortex. *Curr. Opin. Neurobiol*. 58, 1–10. doi: 10.1016/j.conb.2019.06.001

Lindner, B., Doiron, B., and Longtin, A. (2005). Theory of oscillatory firing induced by spatially correlated noise and delayed inhibitory feedback. *Phys. Rev. E.* 72:061919. doi: 10.1103/PhysRevE.72.061919

Lindner, B., and Schimansky-Geier, L. (2001). Transmission of noise coded versus additive signals through a neuronal ensemble. *Phys. Rev. Lett*. 86, 2934–2937. doi: 10.1103/PhysRevLett.86.2934

Liu, J., Hu, B., and Wang, Y. (2019). Optimum adaptive array stochastic resonance in noisy grayscale image restoration. *Phys. Lett. A*. 383, 1457–1465. doi: 10.1016/j.physleta.2019.02.006

Liu, R. N., and Kang, Y. M. (2018). Stochastic resonance in underdamped periodic potential systems with alpha stable Lévy noise. *Phys. Lett. A*. 382, 1656–1664. doi: 10.1016/j.physleta.2018.03.054

Nakamura, O., and Tateno, K. (2019). Random pulse induced synchronization and resonance in uncoupled non-identical neuron models. *Cogn. Neurodyn*. 13, 303–312. doi: 10.1007/s11571-018-09518-5

Owsley, C., McGwin, G., Clark, M. E., Jackson, G. R., Callahan, M. A., Kline, L. B., et al. (2016). Delayed rod-mediated dark adaptation is a functional biomarker for incident early age-related macular degeneration. *Ophthalmology* 123, 344–351. doi: 10.1016/j.ophtha.2015.09.041

Parmananda, P., Santos, G. J. E., Rivera, M., and Showalter, K. (2005). Stochastic resonance of electrochemical aperiodic spike trains. *Phys. Rev. E.* 71:031110. doi: 10.1103/PhysRevE.71.031110

Patel, A., and Kosko, B. (2011). Noise benefits in quantizer-array correlation detection and watermark decoding. *IEEE Trans. Signal Process* 59, 488–505. doi: 10.1109/TSP.2010.2091409

Pei, X., Wilkens, L. A., and Moss, F. (1996). Light enhances hydrodynamic signaling in the multimodal caudal photoreceptor interneurons of the crayfish. *J. Neurophysiol*. 76, 3002–3011. doi: 10.1152/jn.1996.76.5.3002

Pernice, V., Staude, B., Cardanobile, S., and Rotter, S. (2011). How structure determines correlations in neuronal networks. *PLoS Comput. Biol*. 7:e1002059. doi: 10.1371/journal.pcbi.1002059

Purves, D. (2011). *Brains: How They Seem to Work*. New Jersey, NJ: Financial Times Press Science.

Rizzolatti, G., Camarda, R., Grupp, L. A., and Pisa, M. (1974). Inhibitory effect of remote visual stimuli on visual response of cat superior colliculus: spatial and temporal factors. *J. Neurophysiol*. 37, 1262–1275. doi: 10.1152/jn.1974.37.6.1262

Roska, B., Molnar, A., and Werblin, F. S. (2006). Parallel processing in retinal ganglion cells: how integration of space-time patterns of excitation and inhibition form the spiking output. *J. Neurophysiol*. 95, 3810–3822. doi: 10.1152/jn.00113.2006

Sasaki, H., Sakane, S., Ishida, T., Todorokihara, M., Kitamura, T., and Aoki, R. (2008). Suprathreshold stochastic resonance in visual signal detection. *Behav. Brain Res*. 193, 152–155. doi: 10.1016/j.bbr.2008.05.003

Simonotto, E., Riani, M., Seife, C., Roberts, M., Twitty, J., and Moss, F. (1997). Visual perception of stochastic resonance. *Phys. Rev. Lett*. 78, 1186–1189. doi: 10.1103/PhysRevLett.78.1186

Susstrunk, S. E., and Winkler, S. (2003). Color image quality on the Internet. *SPIE Electron. Imaging* 5304, 118–131. doi: 10.1117/12.537804

Sutherland, C., Doiron, B., and Longtin, A. (2009). Feedback-induced gain control in stochastic spiking networks. *Biol. Cybern*. 100, 475–489. doi: 10.1007/s00422-009-0298-5

Taillefumier, T., and Magnasco, M. O. (2013). A phase transition in the first passage of a Brownian process through a fluctuating boundary with implications for neural coding. *Proc. Natl. Acad. Sci. U.S.A*. 110, 1438–1443. doi: 10.1073/pnas.1212479110

Trousdale, J., Hu, Y., Shea-Brown, E., and Josić, K. (2012). Impact of network structure and cellular response on spike time correlations. *PLoS Comput. Biol.* 8:e1002408. doi: 10.1371/journal.pcbi.1002408

van der Groen, O., Tang, M. F., Wenderoth, N., and Mattingley, J. B. (2018). Stochastic resonance enhances the rate of evidence accumulation during combined brain stimulation and perceptual decision-making. *PLoS Comput. Biol*. 14:e1006301. doi: 10.1371/journal.pcbi.1006301

Wang, Z., Sheikh, H. R., and Bovik, A. C. (2002). No reference perceptual quality assessment of JPEG compressed images. *IEEE Int. Conf. Image Process* 1, 477–480. doi: 10.1109/ICIP.2002.1038064

Yang, T. (1998). Adaptively optimizing stochastic resonance in visual system. *Phys. Lett.* 245, 79–86. doi: 10.1016/S0375-9601(98)00351-X

Yu, T., Park, J., Joshi, S., Maier, C., and Cauwenberghs, G. (2012). "65K-neuron integrate-and-fire array transceiver with address-event reconfigurable synaptic routing," in *2012 IEEE Biomed. Circuits Syst. Conf. Intell. Biomed. Electron. Syst. Better Life Better Environ.* BioCAS 2012 - Conf. Publ., 21–24. doi: 10.1109/BioCAS.2012.6418479

Zeng, F. G., Fu, Q. J., and Morse, R. (2000). Human hearing enhanced by noise. *Brain Res*. 869, 251–255. doi: 10.1016/S0006-8993(00)02475-6

Zhang, Y., Liu, H., Huang, N., and Wang, Z. (2019). Discrete image recovery via stochastic resonance in optically induced photonic lattices. *Sci. Rep*. 9:11815. doi: 10.1038/s41598-019-48313-y

**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.

# 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