Perturbing a Neural Network to Infer Effective Connectivity: Evidence from Synthetic EEG Data

Peizhen Yang[1*], **Xinke Shen**[1,*], **Zongsheng Li**[1,2], **Zixiang Luo**[1], **Kexin Lou**[1,3], **Quanying Liu**[1†]

[1]Department of Biomedical Engineering, Southern University of Science and Technology
[2]Department of Computer Science, University of Macau
[3]School of Information Technology and Electrical Engineering, University of Queensland
pyang11@u.rochester.edu, shenxk@sustech.edu.cn, 591393178@qq.com,
12032934@mail.sustech.edu.cn, k.lou@uq.edu.au, liuqy@sustech.edu.cn

Abstract

Identifying causal relationships among distinct brain areas, known as effective connectivity, holds key insights into the brain's information processing and cognitive functions. Electroencephalogram (EEG) signals exhibit intricate dynamics and inter-areal interactions within the brain. However, methods for characterizing nonlinear causal interactions among multiple brain regions remain relatively underdeveloped. In this study, we proposed a data-driven framework to infer effective connectivity by perturbing the trained neural networks. Specifically, we trained neural networks (*i.e.*, CNN, vanilla RNN, GRU, LSTM, and Transformer) to predict future EEG signals according to historical data and perturbed the networks' input to obtain effective connectivity (EC) between the perturbed EEG channel and the rest of the channels. The EC reflects the causal impact of perturbing one node on others. The performance was tested on the synthetic EEG generated by a biological-plausible Jansen-Rit model. CNN and Transformer obtained the best performance on both 3-channel and 90-channel synthetic EEG data, outperforming the classical Granger causality method. Our work demonstrated the potential of perturbing an artificial neural network, learned to predict future system dynamics, to uncover the underlying causal structure.

1 Introduction

The brain is a profoundly intricate, interwoven network characterized by causal influences among its various regions [Sporns *et al.*, 2004; Van Den Heuvel and Pol, 2010; Liu *et al.*, 2017]. From brain recordings like functional magnetic resonance imaging (fMRI) or electroencephalogram (EEG), one can easily compute the correlation of signal dynamics across different regions, which is typically referred to as functional connectivity (FC) [Park and Friston, 2013; Van Den Heuvel and Pol, 2010; Samogin *et al.*, 2019]. However, FC cannot characterize the *directionality* and the *sign* of connectivity [Buckner *et al.*, 2013]. Instead, effective connectivity (EC) can reflect the underlying directional causal influence from one brain region to another with the strength, directionality, and sign [Luo *et al.*, 2022; Kim *et al.*, 2023]. It characterizes the information flow across brain regions and is critical for understanding how information is integrated or segregated during the cognitive process. Therefore, developing methods to reliably estimate EC from the recorded neural data stands as a significant endeavor in the field of neuroscience [Reid *et al.*, 2019].

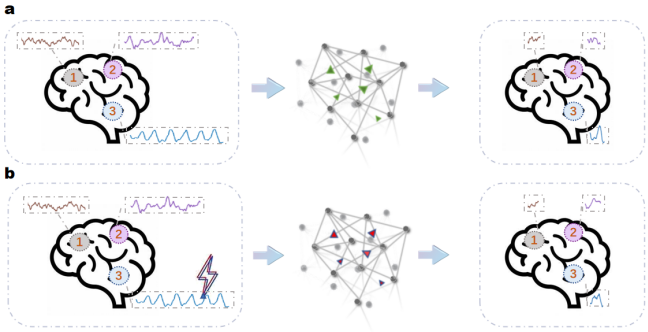

Figure 1: **Framework of perturbation-based EC inference on EEG data. a**, Training artificial neural networks (ANNs) for EEG signal prediction. ANNs are trained to predict EEG signals of the following time steps from multiple previous steps. **b**, Perturbing the trained ANNs to infer EC. The trained ANNs are used as surrogate brains. EC is estimated by sequentially perturbing each region of the surrogate brain and measuring the stimulation-induced responses.

A number of computational methods have been proposed to investigate effective connectivity among multiple brain regions based on time series of neural signals, including Granger causality(GC) [Granger, 1969] and dynamic causal modeling (DCM) [Friston *et al.*, 2003]. However, these methods are limited in their capacity to accurately encapsulate complex, non-linear interactions. For example, Granger causality operates on the premise that if the past activities of brain region A can predict the activities of another region B, there should be a causal interaction from region A to B.

*Co-first authors.
†The corresponding author.

Granger causality typically relies on the assumptions of linear dependency and is sensitive to the effect of noise [Friston *et al.*, 2014]. On the other hand, DCM employs a Bayesian framework to identify the nonlinear input–state–output systems from observed data [Friston *et al.*, 2003; Penny *et al.*, 2004]. Despite its popularity in neuroscience, DCM is limited by the pre-designed system dynamic model and its computational demand with an increasing number of nodes [Lohmann *et al.*, 2012; Daunizeau *et al.*, 2011].

Rather than identified with computational methods, effective connectivity can be directly mapped with in-vivo experiments by applying electrical or optical stimulations to a specific brain region and then evaluating the resultant impact on other regions [Kim *et al.*, 2023]. Although such perturbation-based experimental approaches are straightforward, they are typically unfeasible in human subjects due to ethical reasons and technical limitations. To address this challenge, we developed a data-driven framework that leverages a surrogate brain model to capture brain dynamics. Within this framework, we apply in-silico perturbations to various brain regions and observe the subsequent influence on other regions. This conceptually simple approach allows us to investigate effective connectivity by virtual perturbations.

Here, we employed artificial neural networks (ANNs) to characterize the nonlinear interactions among brain regions. By training a neural network to predict the future dynamics of neural signals, we create a surrogate brain that could be perturbed to uncover causal interactions (Fig. 1. This approach enables us to leverage state-of-the-art ANN architectures in time series forecasting and comprehensively investigate the optimal characterization of intricate nonlinear relationships among brain regions.

In our experiment, we used synthetic EEG data generated by a biological-plausible Jansen-Rit (J-R) mmodel [Coronel-Oliveros *et al.*, 2021], which can capture the fast-changing nonlinear dynamics of EEG. As the ground-truth effective connectivity is unknown in real EEG data, we developed a testbed for the framework using synthetic data. The ground-truth effective connectivity can be obtained from the synthetic data by perturbing the hidden variables during data generation. We generated two datasets: 1) A simple J-R synthesized dataset with 3 regions and pre-defined connections as a proof-of-concept; 2) A J-R synthesized dataset with 90 regions, in which the connections were real structural connectivity measured from diffusion tensor imaging (DTI).

The contributions of this paper are two-fold:

- We presented a testbed for verifying the data-driven EC inference framework on fast-changing synthetic EEG data with known real EC. Various neural network models were tested in this testbed.
- We validated the effectiveness of the neural perturbational inference framework in comparison to classical EC estimation methods. The results underline the importance of selecting a proper model to serve as a surrogate brain.

2 Problem statement

The primary aim of this work is to estimate the causal influence of one brain region on others. To achieve this end, we implement virtual perturbation on the trained neural networks that can predict the dynamics of neural signals. The prediction model can be represented as

$$\hat{\mathbf{x}}_{t+1:t+T'} = f(\mathbf{x}_{t-T:t}, \theta), \tag{1}$$

which means predicting neural activities $\hat{\mathbf{x}}_{t+1:t+T'}$ with a nonlinear model f based on previous activities $\mathbf{x}_{t-T:t}$. After the model is trained, we perturb one region at a time and see the changes in the predicted signals of other regions, which can be formulated as

$$\delta_{A\to B}(t+t') = \mathbf{E}[(B_{t+t'} \mid A_t + \Delta) - (B_{t+t'} \mid A_t)], \quad t' = 1, 2, ..., T' \tag{2}$$

Here, we add perturbation Δ on region A at time t and see the expected changes in the region B at time $t+1$ to $t+T'$. We perturb every region in a loop and see the causal influence between any two region pairs in this way.

3 Related Work

3.1 Classical EC estimation methods

The classical computational methods for estimating effective connectivity from neural data can be broadly categorized based on two aspects: i) linearity/nonlinearity and ii) bivariate/multivariate analysis. For instance, Granger causality, as a linear bivariate method, usually focuses on the linear interactions between two brain regions [Friston *et al.*, 2014]. Transfer entropy, as a nonlinear bivariate method, quantifies the directed transfer of information between two regions [Yang *et al.*, 2012]. On the other hand, multivariate models take into account the interactions among multiple regions concurrently. These multivariate models have the advantage of mitigating spurious connections that may arise in bivariate models. For example, the partially directed coherence [Baccalá and Sameshima, 2001] and directed transfer function [Wilke *et al.*, 2008] derive causal interactions from the Fourier transform of multivariate autoregressive parameters. The capability of these methods to simultaneously identify nonlinear multivariate interactions are underdeveloped. Recently, deep neural networks have shown their great expressive power for multivariate time-series prediction [Wang *et al.*, 2019; Bianchi *et al.*, 2020; Liang *et al.*, 2022]. However, whether their expressive power can transfer to accurately capture the complex and nonlinear interactions in the multivariate data requires further investigation.

3.2 Neural perturbational inference

The data-driven framework of Neural Perturbational Inference (NPI) was proposed by Luo et al. [Luo *et al.*, 2022]. NPI uses an artificial neural network (ANN) that learns neural dynamics as a surrogate brain. Perturbing the surrogate brain (*i.e.*, the trained ANN), region by region, and simultaneously observing the evoked neural response at all unperturbed regions provides the whole-brain effective connectivity. The ANN in NPI is instantiated with a four-layer perceptron and is trained using a one-step-ahead prediction error,

where the next state of fMRI is predicted given the current state. After ANN is trained, each region of ANN is sequentially perturbed, realized as a small increase or decrease of neural signal in the perturbed region. The EC is computed by the difference between the one-step neural responses with and without perturbation.

The NPI framework has demonstrated its ability to infer EC from fMRI signals [Luo *et al.*, 2022]. Owing to the long timescale and slow dynamics in fMRI signals, the current state contains most of the useful information to effectively forecast the next state [Nozari *et al.*, 2021]. Therefore, the ANN in NPI is simply realized with a multi-layer perceptron trained using one-step-ahead prediction error [Luo *et al.*, 2022]. However, this one-step prediction approach may fall short for capturing EEG dynamics. As the nature of EEG dynamics is highly nonlinear and complex, with rich information for predicting the next EEG signal is contained in many previous steps. Therefore, NPI developed for fMRI data cannot be directly applied to EEG. Here, we extended the original NPI framework with two factors: i) the ANN in the NPI framework to predict EEG dynamics is replaced with the state-of-the-art time-series prediction models (*e.g.*, RNN, GRU, LSTM, CNN, and Transformer), ii) the EC is estimated with the multi-step response after perturbation, rather than one-step transient response.

4 Methods

4.1 The framework

The framework of our method is shown in Fig. 1. To learn the system dynamics from EEG signals, we trained a time series forecasting model to predict the subsequent signals using previous n steps of EEG data (Fig. 1(a)). The virtual perturbation was applied to a region at the 76^{th} step. (Fig. 2(b) left), and the responses across all brain regions at future 77 to 99 steps were predicted by the trained models (Fig.2(b) right). The estimated EC was calculated as the difference between the expected signals with and without disturbance. To obtain the whole-brain causal connection between any two regions, we individually make an impulse perturbation (unit=0.1) into each region.

4.2 The time series forecasting models

Following the instruction of the NPI framework, the EC inference largely relies on the time series forecasting model. In this work, we realized five artificial neural network models (CNN, vanilla RNN, LSTM, GRU and transform) as the EEG series forecasting model. The models are detailed in the Appendix.

4.3 Synthetic EEG data

We used the biologically plausible Jansen-Rit model, as shown in Eq.(3), to generate EEG data. The Jansen-Rit model is a mathematical model used to simulate the macroscopic electrical behavior observed in EEG signals. The simulated data mimic the real EEG data in nonlinear dynamics and complex inter-regional interactions. For each brain region, it assumes three populations of neurons: pyramidal neurons, excitatory interneurons, and inhibitory interneurons. Pyramidal neurons have projections to the other two populations. Excitatory and inhibitory interneurons project back to pyramidal neurons. The pyramidal neurons also have long-range excitatory projections to other brain regions. The dynamics of each region are represented as follows:

$$
\begin{aligned}
\dot{x}_{0,i}(t) &= y_{0,i}(t) \\
\dot{y}_{0,i}(t) &= Aa\left[S\left(C_2 x_{1,i}(t) - C_4 x_{2,i}(t) + C\alpha z_i(t), r_0\right)\right] \\
&\quad - 2a y_{0,i}(t) - a^2 x_{0,i}(t) \\
\dot{x}_{1,i}(t) &= y_{1,i}(t) \\
\dot{y}_{1,i}(t) &= Aa\left[p(t) + S\left(C_1 x_{0,i}(t) - C\beta x_{2,i}, r_1\right)\right] \\
&\quad - 2a y_{1,i}(t) - a^2 x_{1,i}(t) \\
\dot{x}_{2,i}(t) &= y_{2,i}(t) \\
\dot{y}_{2,i}(t) &= Bb\left[S\left(C_3 x_{0,i}(t), r_2\right)\right] - 2b y_{2,i}(t) - b^2 x_{2,i}(t) \\
\dot{x}_{3,i}(t) &= y_{3,i}(t) \\
\dot{y}_{3,i}(t) &= A\bar{a}\left[S\left(C_2 x_{1,i}(t) - C_4 x_{2,i}(t) + C\alpha z_i(t), r_0\right)\right] \\
&\quad - 2\bar{a} y_{3,i}(t) - \bar{a}_i^2 x_{3,i}(t)
\end{aligned}
\tag{3}
$$

where x_0, x_1 and x_2 represent the output of the pyramidal neurons, excitatory interneurons, and inhibitory interneurons, respectively. x_3 represents the long-range output of the pyramidal neurons to other regions. S is a sigmoid function:

$$
S(v, r) = \frac{\zeta_{\max}}{1 + e^{r(\theta - v)}} \tag{4}
$$

z_i is the overall input from other regions to region i:

$$
z_i(t) = \sum_{j=1, j\neq i}^{n} \widetilde{M}_{ij} x_{3,j}(t) \tag{5}
$$

where $\widetilde{M}_{ij}$ is the normalized structural connectivity matrix:

$$
\tilde{M}_{ij} = \frac{M_{ij}}{\sum_{j=1, j\neq i}^{n} M_{ij}} \tag{6}
$$

M_{ij} represents the underlying structural connectivity from region j to region i. The EEG-like signal is calculated as:

$$
v_i(t) = C_2 x_{1,i}(t) - C_4 x_{2,i}(t) + C\alpha z_i(t) \tag{7}
$$

which represents the postsynaptic potentials of pyramidal neurons in region i.

We generated two versions of synthetic data: 1) A toy example with 3 nodes. The structural connectivity among the 3 nodes was set manually, with node 0 exerting directed connections to node 1 and node 2 (Fig. 2a). 2) A whole-brain model with 90 nodes. Connectivity among the nodes was determined by real structural connectivity measured from DTI. The brain was parcellated into 90 regions with Anatomical Automatic Labeling (AAL) atlas. The structural connectivity was calculated from the average of 245 subjects in the Human Connectome Project (https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release) (Fig. 3(a,b)).

Hyperparameter settings of Jasen-Rit model. In this study, the excitatory gain α and the inhibitory gain β were set as 0.71 and 0.4, respectively, to generate reasonable power spectra and functional connectivity properties in synthetic data. All other hyperparameters were set the same as in [Coronel-Oliveros *et al.*, 2021]. The neural dynamic described in Eq.(3) was transformed into a discretized formula by the forward-Euler method and evolved with a time step of 0.001 seconds. Then the generated EEG signals were downsampled to 100 Hz, with a time interval of 0.01 seconds between two consecutive time steps.

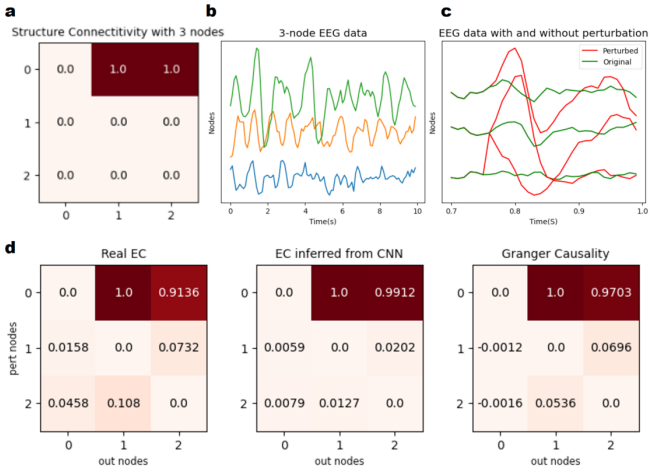

Figure 2: **Data and results visualization of the 3-channel synthetic EEG**. **a**, The setting for 3-channel structural connectivity matrix. **b**, An example of 3-channel EEG data. **c**, The time course of the original EEG data (the green lines) and the perturbed EEG data (the red lines). **d**, the comparison among the real EC (left), the EC inferred by perturbing a CNN model (middle), and EC inferred by Granger causality (right). All the EC matrices were re-scaled to the range 0-1 for visualization.

4.4 Implementation details of ANN models

We implemented five ANN models, including temporal CNN, vanilla RNN, LSTM, GRU, and Transformer models. Each ANN model consists of two hidden layers with 8, 32, 128, and 512 units, respectively, thus we can examine the impact of the model complexity on the performance of data prediction and EC inference. A linear readout layer was employed to predict future EEG dynamics. We applied Adam optimizer and ReduceLROnPlateau scheduler while training. The initial learning rate was $1e^{-4}$ and the batch size was 30. The number of training epochs was determined according to the minimum validation loss of each ANN model.

Training and testing datasets. For training and validating the forecasting model, an EEG signal with 900,000 time points was generated. The first 70% of the generated time series was used in training and the remaining 30% was used in validation. Each training or validation sample contains signals of 100 time points (i.e., 1 second). Adjacent samples do not overlap. The model needs to predict the following 24-step signals based on the previous 76-time steps. We reported the validation mean squared error as the model's prediction performance.

4.5 Virtual perturbation of ANN models

For model perturbation, we generated another time series of 100,000 time points, which formed 1,000 samples. During the generation of synthetic data, we added a perturbation with a value of 0.1 to the excitatory interneurons x_1 at the 76^{th} step of each sample and recorded the changes in the following steps. The perturbation was applied to one region and will affect other regions. The real EC was calculated as the average difference between the following generated data with and without perturbation. For the trained ANNs, we input the perturbed data (time steps 1-76 of each sample) and obtained the predicted signals of the following time steps. The estimated EC was calculated from the average difference between the predicted data with and without perturbation in the input.

Table 1: **Related statistics of 3-channel EEG data prediction.** Five ANN models, including CNN, RNN, GRU, LSTM, and Transformer with 8, 32, 128, and 512 hidden units, were trained to predict EEG signals, respectively. The prediction error and the correlation between the real EC and the NPI-EC were calculated with the test data.

ANN model	Hidden Units	Prediction Error↓	EC Correlation↑
CNN	8	5.4477	0.7442
	32	5.4032	0.7274
	128	**5.3920**	**0.8785**
	512	5.4202	0.7451
RNN	8	**5.3026**	-0.1080
	32	5.3057	-0.0890
	128	5.3047	**0.2636**
	512	5.3514	0.1820
LSTM	8	5.6079	0.2593
	32	**5.4021**	0.5353
	128	5.5618	0.2984
	512	5.6119	**0.6430**
GRU	8	5.5250	0.3675
	32	**5.3751**	0.6186
	128	5.3816	**0.6436**
	512	5.5334	0.4198
Transformer	8	5.4221	0.7680
	32	5.4861	0.6780
	128	**5.3803**	**0.8110**
	512	5.4132	0.8022

5 Results

5.1 Results on 3-channel synthetic EEG

We first examined the model performance for EC estimation with 3-channel synthetic EEG data. The model performance was evaluated based on two metrics: time series prediction error and correlation between the real EC and the predicted EC on testing data, as shown in Table 1. The time series prediction error is calculated as the Mean Square Error (MSE)

between the real signal and the predicted signal of the last 24 time steps in each sample. EC correlation is defined as the Pearson correlation coefficient between the real EC and the NPI-EC without considering matrix diagonals. We reported the performance of CNN, RNN, LSTM, GRU, and Transformer models with different hidden dimensions. CNN model with 128 hidden dimensions achieved the best EC correlation of 0.8785. The transformer model follows with an EC correlation of 0.8110. For these two models, the hidden dimension with a higher EC correlation also accompanies a lower time series prediction error. LSTM and GRU obtained an inferior EC correlation of 0.6430 and 0.6436, respectively. Vanilla RNN obtained the worst EC correlation of 0.2636. The results indicated that recurrent neural networks are worse than CNN and Transformer in recovering the underlying causal interactions from the synthetic EEG data, although they all achieved comparable time series prediction errors.

We visualized the real EC and the EC inferred from CNN at time step 3 (30 ms) after perturbation (Fig. 2d, left and middle columns). The inferred EC faithfully recovers the causal interaction from node 0 to nodes 1 and 2. We also visualized the EC estimated by Granger causality (Fig. 2d, right column). We used multivariate GC to calculate the direct connection between two channels and choose 12 as the input for maxlag based on the minimum value of bic. There are small false positive connections between node 1 and node 2 in EC estimated by GC, which is better suppressed in the NPI-EC.

5.2 Results on whole-brain synthetic EEG

For the whole-brain synthetic EEG with 90 regions, CNN obtained the highest correlation between the real EC and the NPI-EC ($R = 0.3340$), compared with Transformer ($R = 0.3245$), LSTM ($R = 0.2055$), GRU ($R = -0.0096$) and RNN ($R = -0.0006$). This trend is similar to that of 3-channel synthetic data. GRU and RNN wrongly estimate the causality. GRU's failure may be due to the simpler design of the gating mechanism in contrast to LSTM.

We visualized the real EC, the EC inferred by CNN and Granger causality at time step 16 (160 ms) after perturbation (Fig. 3c, from left to right). The inferred EC can recover the real EC faithfully, with a high EC correlation ($R = 0.7081$) (for this time step). In contrast, the Granger causality inferred EC is much worse ($R = 0.3136$).

To exhibit the effect of perturbing one region on the others, we show the spatial distribution of signal changes resulting from perturbing a specific seed region in the J-R model (i.e., real EC) in Fig. 3d, as well as the NPI-EC of CNN model in Fig. 3e. The NPI-EC recovers the general distribution of real EC. To show the temporal evolution of the signals after perturbation, we also compare the real event-related potentials (ERP) and the predicted ERP under perturbation in Fig. 3f. After the perturbation was given at time point 0, the predicted ERP and the real ERP show a similar trend of change, although their exact values were different.

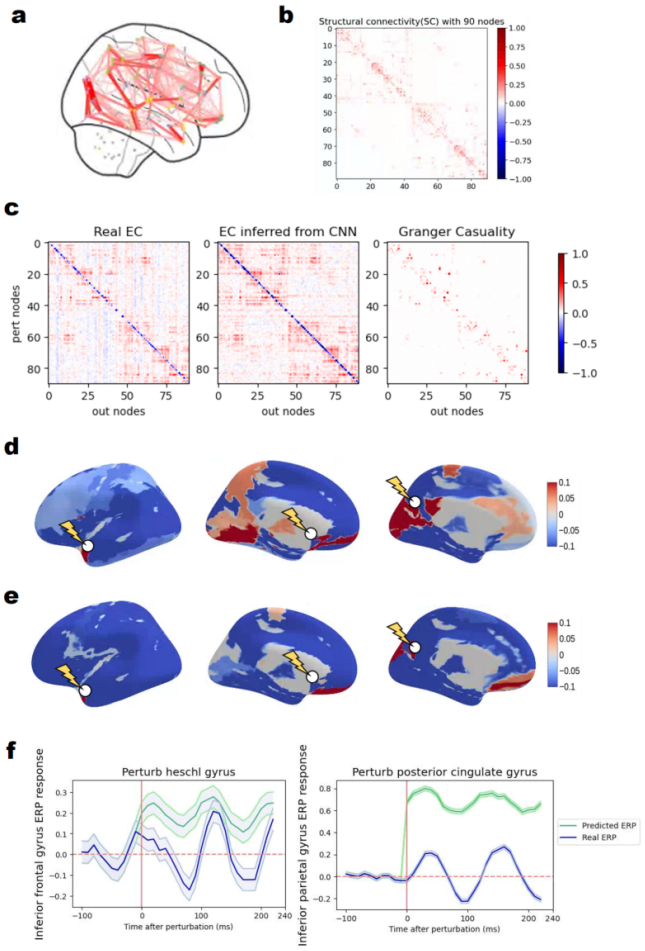

Figure 3: **Data and results visualization of 90-channel synthetic EEG**. **a**, The real structural connectivity measured from diffusion tensor imaging (mapped on 90 brain regions). **b**, The structural connectivity matrix. **c**, The comparison among the real EC (left), the EC inferred by perturbing the CNN model (middle), and EC inferred by Granger causality (right). All the EC matrices were re-scaled to the range 0-1 for visualization. **d**, Spatial distribution of the real EC (i.e., neural responses) by perturbing left amygdala (left), left rectus (middle), and left cuneus (right). **e**, Spatial distribution of the NPI-EC by perturbing the three regions same as in **d**. The perturbed region is indicated with an arrow in each panel. **f**, Sample event-related potential (ERP) of the stimulus-evoked neural responses. ERP is the average response to 1000 times perturbation. Perturbation is given at time point 0. We show the predicted and real response of the inferior frontal gyrus to heschl gyrus perturbation (left) and the response of the inferior parietal gyrus to the posterior cingulate gyrus perturbation (right).

6 Discussion

In this study, we presented a testbed for perturbation-based EC estimation methods with synthetic EEG data. Our results validated that by perturbing specific types of ANN prediction models (i.e., CNN and Transformer), we can estimate the underlying causal interactions among different nodes effectively.

ANN models have been widely used in time series forecasting and achieved SOTA performance. However, it is

unclear whether the models can reveal the causal interactions among different variates. Our experiments showed that specific types of models can encapsulate the underlying causal interactions of synthetic EEG data. CNN and Transformer achieved higher performance here, probably due to they can capture the oscillation characteristics in synthetic EEG data [Lawhern *et al.*, 2018; Song *et al.*, 2022].

In future studies, several important questions remain to be investigated. Firstly, what are the effects of different types of perturbation? Some perturbations may cause the signals to be outside the manifold of natural signals, while others may not [Shenoy and Kao, 2021]. Specific forms of perturbations may resemble those in real brain stimulation. It is critical to investigate these different types of perturbations on EC estimation. Secondly, it still lacks a clear explanation of why CNN and Transformer work better on EC estimation than RNN-series models. Do they also work well on real EEG data? How to choose the proper model for different types of data? These questions need to be further investigated in the future.

Acknowledgements

We thank Mr. Zhichao Liang for sharing some code, and Mr. Song Wang, and Mr. Kaining Peng for their useful discussions. This work was funded in part by Shenzhen Science and Technology Innovation Committee (2022410129, 20200925155957004, KCXFZ2020122117340001, JCYJ20220818100213029, SGDX2020110309280100), Guangdong Provincial Key Laboratory of Advanced Biomaterials (2022B1212010003). 0

Appendix

Convolutional Neural Network (CNN). The Convolutional Neural Network is a feedforward multilayered hierarchical network, which is a widely used ANN model. It uses a combination of convolutional layers, nonlinear processing units, and subsampling layers to automatically extract features from the raw pixel data of the image for improved categorization with 2-dimensional data. It can be applied to 1-dimensional time series data by temporal convolution.

Vanilla RNN. Vanilla RNN was used as an example of a simple nonlinear forecasting model. The iteration of the hidden state is represented as

$$h_t = tanh(x_t W_{ih}^T + b_{ih} + h_{t-1} W_{hh}^T + b_{hh}), \qquad (8)$$

where x_t is the input and h_t is the hidden state.

Long Short-Term Memory (LSTM). LSTM adds gate design to vanilla RNN to capture long-term dependencies in the time series. The detailed computation in an LSTM unit is shown below:

$$\begin{aligned} i_t &= \sigma\left(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}\right) \\ f_t &= \sigma\left(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}\right) \\ g_t &= \tanh\left(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}\right) \\ o_t &= \sigma\left(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}\right) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tanh\left(c_t\right) \end{aligned} \qquad (9)$$

where x_t is the input and h_t is the hidden state. i_t, f_t, and o_t represent the output of the input gate, the forget gate, and the output gate, respectively. g_t is the candidate cell state and c_t is the cell state.

Gated Recurrent Unit (GRU). GRU simplified the gate design in LSTM to improve the computational efficiency:

$$\begin{aligned} r_t &= \sigma\left(W_{ir}x_t + b_{ir} + W_{hr}h_{(t-1)} + b_{hr}\right) \\ z_t &= \sigma\left(W_{iz}x_t + b_{iz} + W_{hz}h_{(t-1)} + b_{hz}\right) \\ n_t &= \tanh\left(W_{in}x_t + b_{in} + r_t * \left(W_{hn}h_{(t-1)} + b_{hn}\right)\right) \\ h_t &= (1 - z_t) * n_t + z_t * h_{(t-1)} \end{aligned} \qquad (10)$$

where x_t is the input and h_t is the hidden state. r_t and z_t represent the output of the reset gate and the update gate, respectively. n_t is the candidate's hidden state.

Transformer. The Transformer model employs a self-attention mechanism. By encoding the input time series into a set of vectors and applying self-attention across all time steps, the Transformer model captures both local and global dependencies, enabling accurate predictions. The attention weights are obtained through a softmax function applied to the scaled dot-product of query, key, and value embeddings. It computes the attention function on a set of queries simultaneously, packed together into a matrix Q. The keys and values are also packed together into matrices K and V. The output is computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \qquad (11)$$

where d_k is the dimension of queries and keys. By iteratively updating the hidden states through the self-attention layers, the Transformer model learns to capture complex temporal dependencies, facilitating accurate prediction of future values in the time series.

where $\mathbf{y}_t$ is a vector of observed variables at time t, $\mathbf{c}$ is a constant vector, $\mathbf{A}i$ is the coefficient matrix associated with the i-th lag, $\mathbf{y}t - i$ represents the vector of lagged variables, and $\mathbf{e}_t$ is a vector of error terms assumed to be white noise.

References

[Baccalá and Sameshima, 2001] Luiz A Baccalá and Koichi Sameshima. Partial directed coherence: a new concept in neural structure determination. *Biological cybernetics*, 84(6):463–474, 2001.

[Bianchi *et al.*, 2020] Filippo Maria Bianchi, Simone Scardapane, Sigurd Løkse, and Robert Jenssen. Reservoir computing approaches for representation and classification of multivariate time series. *IEEE transactions on neural networks and learning systems*, 32(5):2169–2179, 2020.

[Buckner *et al.*, 2013] Randy L Buckner, Fenna M Krienen, and BT Thomas Yeo. Opportunities and limitations of intrinsic functional connectivity mri. *Nature neuroscience*, 16(7):832–837, 2013.

[Coronel-Oliveros *et al.*, 2021] Carlos Coronel-Oliveros, Rodrigo Cofré, and Patricio Orio. Cholinergic neuromodulation of inhibitory interneurons facilitates functional

integration in whole-brain models. *PLoS Computational Biology*, 17(2):e1008737, 2021.

[Daunizeau *et al.*, 2011] Jean Daunizeau, Olivier David, and Klaas E Stephan. Dynamic causal modelling: a critical review of the biophysical and statistical foundations. *Neuroimage*, 58(2):312–322, 2011.

[Friston *et al.*, 2003] Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19(4):1273–1302, 2003.

[Friston *et al.*, 2014] Karl J Friston, André M Bastos, Ashwini Oswal, Bernadette van Wijk, Craig Richter, and Vladimir Litvak. Granger causality revisited. *Neuroimage*, 101:796–808, 2014.

[Granger, 1969] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[Kim *et al.*, 2023] Seonghoon Kim, Hyun Seok Moon, Thanh Tan Vo, Chang-Ho Kim, Geun Ho Im, Sungho Lee, Myunghwan Choi, and Seong-Gi Kim. Whole-brain mapping of effective connectivity by fmri with cortex-wide patterned optogenetics. *Neuron*, page S0896627323001708, March 2023.

[Lawhern *et al.*, 2018] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018.

[Liang *et al.*, 2022] Zhichao Liang, Zixiang Luo, Keyin Liu, Jingwei Qiu, and Quanying Liu. Online learning koopman operator for closed-loop electrical neurostimulation in epilepsy. *IEEE Journal of Biomedical and Health Informatics*, 2022.

[Liu *et al.*, 2017] Quanying Liu, Seyedehrezvan Farahibozorg, Camillo Porcaro, Nicole Wenderoth, and Dante Mantini. Detecting large-scale networks in the human brain using high-density electroencephalography. *Human brain mapping*, 38(9):4631–4643, 2017.

[Lohmann *et al.*, 2012] Gabriele Lohmann, Kerstin Erfurth, Karsten Müller, and Robert Turner. Critical comments on dynamic causal modelling. *Neuroimage*, 59(3):2322–2329, 2012.

[Luo *et al.*, 2022] Zixiang Luo, Zhichao Liang, Chenyu Xu, Changsong Zhou, and Quanying Liu. Effective brain connectome: the whole-brain effective connectivity from neural perturbational inference. *arXiv preprint arXiv:2301.00148*, 2022.

[Nozari *et al.*, 2021] Erfan Nozari, Maxwell A. Bertolero, Jennifer Stiso, Lorenzo Caciagli, Eli J. Cornblath, Xiaosong He, Arun S. Mahadevan, George J. Pappas, and Dani Smith Bassett. Is the brain macroscopically linear? a system identification of resting state dynamics, August 2021.

[Park and Friston, 2013] Hae-Jeong Park and Karl Friston. Structural and functional brain networks: from connections to cognition. *Science*, 342(6158):1238411, 2013.

[Penny *et al.*, 2004] William D Penny, Klaas E Stephan, Andrea Mechelli, and Karl J Friston. Comparing dynamic causal models. *Neuroimage*, 22(3):1157–1172, 2004.

[Reid *et al.*, 2019] Andrew T. Reid, Drew B. Headley, Ravi D. Mill, Ruben Sanchez-Romero, Lucina Q. Uddin, Daniele Marinazzo, Daniel J. Lurie, Pedro A. Valdés-Sosa, Stephen José Hanson, Bharat B. Biswal, Vince Calhoun, Russell A. Poldrack, and Michael W. Cole. Advancing functional connectivity research from association to causation. *Nature Neuroscience*, 22(11):1751–1760, November 2019.

[Samogin *et al.*, 2019] Jessica Samogin, Quanying Liu, Marco Marino, Nicole Wenderoth, and Dante Mantini. Shared and connection-specific intrinsic interactions in the default mode network. *Neuroimage*, 200:474–481, 2019.

[Shenoy and Kao, 2021] Krishna V Shenoy and Jonathan C Kao. Measurement, manipulation and modeling of brain-wide neural population dynamics. *Nature communications*, 12(1):633, 2021.

[Song *et al.*, 2022] Yonghao Song, Qingqing Zheng, Bingchuan Liu, and Xiaorong Gao. Eeg conformer: Convolutional transformer for eeg decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2022.

[Sporns *et al.*, 2004] Olaf Sporns, Dante R Chialvo, Marcus Kaiser, and Claus C Hilgetag. Organization, development and function of complex brain networks. *Trends in cognitive sciences*, 8(9):418–425, 2004.

[Van Den Heuvel and Pol, 2010] Martijn P Van Den Heuvel and Hilleke E Hulshoff Pol. Exploring the brain network: a review on resting-state fmri functional connectivity. *European neuropsychopharmacology*, 20(8):519–534, 2010.

[Wang *et al.*, 2019] Kang Wang, Kenli Li, Liqian Zhou, Yikun Hu, Zhongyao Cheng, Jing Liu, and Cen Chen. Multiple convolutional neural networks for multivariate time series prediction. *Neurocomputing*, 360:107–119, 2019.

[Wilke *et al.*, 2008] Christopher Wilke, Lei Ding, and Bin He. Estimation of time-varying connectivity patterns through the use of an adaptive directed transfer function. *IEEE transactions on biomedical engineering*, 55(11):2557–2564, 2008.

[Yang *et al.*, 2012] Chunfeng Yang, Régine Le Bouquin Jeannès, Jean-Jacques Bellanger, and Huazhong Shu. A new strategy for model order identification and its application to transfer entropy for eeg signals analysis. *IEEE Transactions on Biomedical Engineering*, 60(5):1318–1327, 2012.