DESIGNING OPTIMAL PERTURBATION INPUTS FOR SYSTEM IDENTIFICATION IN NEUROSCIENCE

Mikito Ogino[1*], **Daiki Sekizawa**[1], **Jun Kitazono**[2], **and Masafumi Oizumi**[1†]

[1] Graduate School of Arts and Sciences, The University of Tokyo
[2] School of Data Science, Yokohama City University
*ogino-m@g.ecc.u-tokyo.ac.jp, †c-oizumi@g.ecc.u-tokyo.ac.jp

ABSTRACT

Investigating the dynamics of neural networks, which are governed by connectivity between neurons, is a fundamental challenge in neuroscience. Perturbation-based approaches allow the precise estimation of neural dynamic models and have been extensively applied in studies of brain functions and neural state transitions. However, the question of how optimal perturbations which most effectively identify dynamical models in neuroscience should be designed remains unclear. To address this, we propose a novel theoretical framework for estimating optimal perturbation inputs for system identification in linear time-invariant systems. The core theoretical insight underlying our approach is that perturbations reveal hidden dynamical modes that are otherwise obscured in the steady state, leading to improved accuracy in system identification. Guided by this insight, our framework derives an objective function to optimize perturbation inputs, which minimizes estimation errors in the model matrices. Building upon this, we further explore the relationship of this function with stimulation patterns commonly used in neuroscience, such as frequency, impulse, and step inputs. We then outline an iterative approach to perturbation input design. Our findings demonstrate that incorporating perturbation inputs significantly improves system identification accuracy as dictated by the objective function. Moreover, perturbation inputs tuned to a parameter related to the eigenvalues and a network structure of the intrinsic model enhance system identification. Through this iterative approach, the estimated model matrix gradually approaches the true matrix. As an application, we confirmed that the framework also contributes to the optimal control theory. This study highlights the potential of designing perturbation inputs to achieve the advanced identification of neural dynamics. By providing a framework for optimizing perturbations, our work facilitates deeper insights into brain functions and advances in the study of complex neural systems.

***K*eywords** control theory, electroencephalogram, frequency input, neural perturbation, system identification, TMS, tDCS, tACS

1 Introduction

Much recent interest has focused on how interactions between individual neurons and neuronal populations support cognitive functions and behavior, and on how the disruption of these interactions contributes to various neurological and psychiatric disorders [1–5]. This interaction—called neuronal connectivity [6, 7]—is often represented as a network structure or a connectivity matrix. The network functional connectivity is estimated using time-series neural activities measured through recording techniques such as neural spike recording techniques [8,9], electrocorticography (ECoG) [10, 11], functional magnetic resonance imaging (fMRI) [12, 13] and electroencephalography (EEG) [14, 15]. While various methods have been proposed and used to estimate these connections (see for example [2, 16] for a comprehensive review), one typical method is based on dynamic modeling, among which is a simple but widely used linear auto-regressive model (Fig 1 (a)). Connectivity is statistically estimated from the time-series data of neural activity (Fig 1 (b)) by fitting the model parameters (Fig 1 (c)). A neural connectivity matrix helps visualize the connections between different brain regions, and provides a detailed map of functional or structural interactions within the

brain. By comparing connectivity matrices across individuals or groups, researchers can deepen their understanding of neuronal architectures and communication across various disciplines, including neuroscience [17, 18], cognitive science [19, 20], and brain-computer interfaces (BCIs) [21, 22]. Nevertheless, while these methods provide significant insights, their reliance on static observations often limits the ability to accurately infer causal relationships.

To address this limitation, the application of perturbations to neural systems has emerged as a highly effective approach to enhancing the estimation of connectivity. Its effectiveness lies in its ability to actively manipulate the system, rather than merely observing it, and to thereby uncover dynamic and causal interactions that are otherwise hidden [23] (see Section 3.1 for an intuitive explanation). Recent studies using stimulation techniques such as optogenetics [24, 25] and transmagnetic stimulation (TMS) [26–28] have demonstrated that these interventions can significantly improve the detection of causal relationships within the system [29]. Such approaches have been applied to the study of cognitive processes and states of consciousness [20, 30], and the results have highlighted the potential of perturbation inputs in advancing our analysis of neural dynamics.

Although it is well-established that perturbations can have a beneficial impact on the estimation of dynamical models in neuroscience, an optimal design of these perturbations for system identification remains unclear. If an optimal perturbation could be determined, dynamics identification would become more reliable and efficient, which would in turn enhance the model's ability to capture and represent the underlying neural dynamics.

In this paper, we propose a framework for estimating the optimal perturbation input for system identification in neuroscience applications. To our knowledge, this framework has not been investigated in the field of neuroscience. First, we derive a theorem that enables us to estimate the optimal perturbation inputs for the identification of linear time-invariant (LTI) systems [31–33]. In this model, the state of the neurons is represented by $\boldsymbol{x}$, and the external inputs are denoted by $\boldsymbol{u}$. The neural connectivity is described by a connectivity matrix $\boldsymbol{A}$, which defines the interactions among the neurons, while perturbations are modulated through an input matrix $\boldsymbol{B}$. The theorem indicate that the estimation errors of these matrices asymptotically correspond to the trace of the inverse of the covariance matrix, which incorporates both $\boldsymbol{x}$ and $\boldsymbol{u}$. This relationship can be explained from the perspectives of asymptotic theory [34, 35] and the principles of design of experiments [36, 37]. Furthermore, we show an objective function that provides an intuitive understanding of the relationship between perturbation inputs and true dynamical models. This objective function serves as a guideline for designing perturbation inputs.

We then extend the derived objective function to methodologies commonly employed in neuroscience. We explore the association between true dynamical models and the optimal parameters of neurostimulation techniques, such as transcranial magnetic stimulation (TMS), transcranial direct current stimulation (tDCS), and transcranial alternating current stimulation (tACS) by considering impulse, step, and frequency inputs, respectively. We revealed that high input intensity and resonant frequencies aligned with the eigenvalues of true matrix $\boldsymbol{A}$ enhance system identification, with inputs targeting nodes with a highly weighted out-degree. Since the true matrix cannot be known from the outset in practical studies, we outline an iterative identification approach to address this issue. This approach gradually improves the accuracy of system identification, and enables the design of optimal perturbation inputs. While this study focuses on estimating a dynamical model of neural systems, we also explore how the estimated parameters contribute to optimal control theory in neuroscience [38–41].

Our research offers a comprehensive framework for system identification with perturbation in neuroscience, and paves the way for more precise and effective analysis of neural dynamics. By providing clear guidelines on the design and application of perturbations, this framework serves as a practical reference for experimental settings, helping researchers determine the most effective stimulation parameters for their studies. We anticipate that these contributions will improve the accuracy of investigations into neural connectivity and dynamics, and ultimately support advances in the precise control and modulation of neural states.

2 Background

In this section, we present a well-established framework in neuroscience that models brain dynamics using Linear Time-Invariant (LTI) systems. We begin by outlining the methods used to identify LTI systems, with an emphasis on parameter estimation techniques through maximum likelihood estimation. We then explore the theoretical foundations of parameter estimation, highlighting the asymptotic properties of the identified parameters, the constraints imposed by the Cramér-Rao bound, and representative evaluation criteria derived from the Fisher Information Matrix. Finally, we explore how the identified parameters impact optimal control theory, which are increasingly utilized in the field of neuroscience.

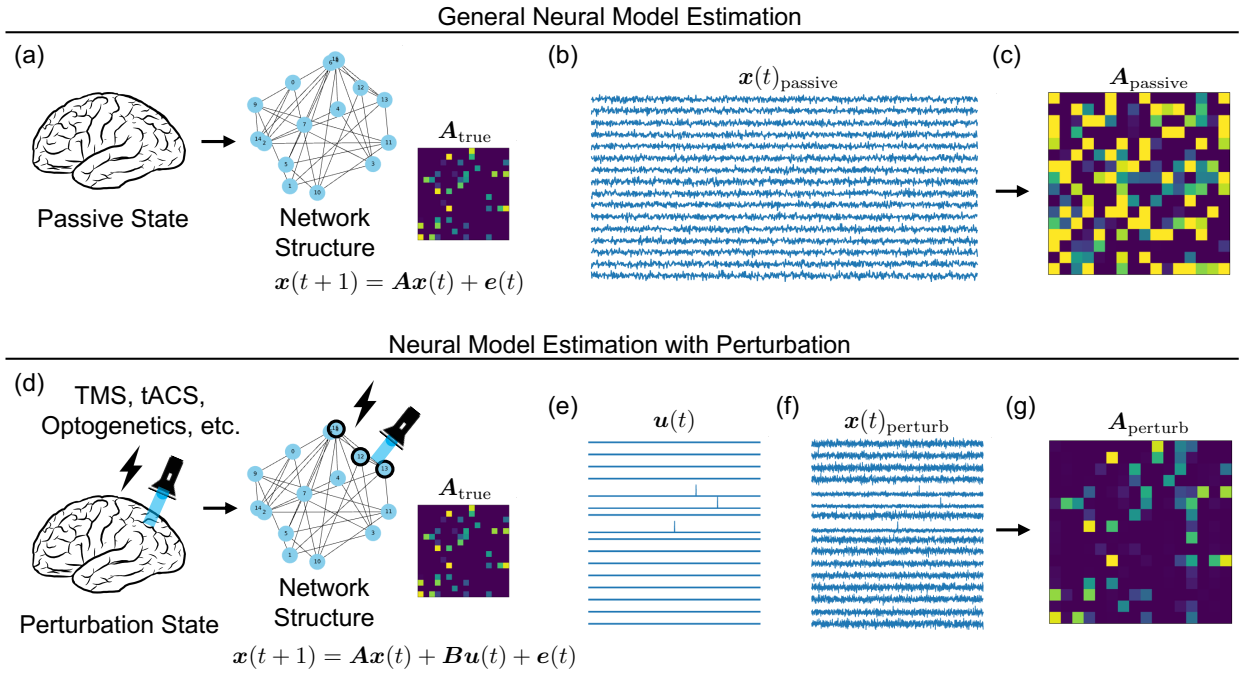

Fig 1: Passive and perturbation states and determined neural dynamic models. (a) The passive state of a neural network and the true model. (b) Measured neural activities without perturbation. (c) Estimated model without perturbation. (d) The perturbation state of a neural network and the true model. (e) Perturbation input for perturbation. (f) Measured neural activities with perturbation. (g) Estimated model with perturbation.

2.1 System Identification Method for LTI system

The LTI system can be represented as follows:

$$\boldsymbol{x}(t+1) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) + \boldsymbol{\xi}(t) \tag{1}$$

where $\boldsymbol{x}(t)$ is a state vector representing the neural activity at time t, with dimensions $n \times 1$. The matrix $\boldsymbol{A}$ $(n \times n)$ describes the interactions between state variables inherent to the system. The input coefficient matrix $\boldsymbol{B}$ $(n \times m)$ determines how external inputs influence the system, and $\boldsymbol{u}(t)$ $(m \times 1)$ represents these external inputs, such as stimuli or control signals. Finally, $\boldsymbol{\xi}(t)$ is a noise vector, typically following a multivariate normal distribution with mean $\mathbf{0}$ and covariance $\sigma^2\boldsymbol{I}$, with dimensions $n \times 1$. This model has been extensively utilized in neuroscience research to investigate the dynamics of neural systems and the effects of external inputs on neural activity [38,42,43]. Given a set of observed $\boldsymbol{x}(t)$ and $\boldsymbol{u}(t)$, we define the cost function for estimating $\boldsymbol{A}$ and $\boldsymbol{B}$ as:

$$\min_{\boldsymbol{A},\boldsymbol{B}} \left\| \boldsymbol{Y} - \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \end{pmatrix} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{U} \end{pmatrix} \right\|_F^2 \tag{2}$$

where $\boldsymbol{X}$, $\boldsymbol{U}$, and $\boldsymbol{Y}$ are the following matrices.

$$\boldsymbol{X} = [\boldsymbol{x}(0), \boldsymbol{x}(1), \ldots, \boldsymbol{x}(T-1)], \quad \boldsymbol{Y} = [\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)], \quad \boldsymbol{U} = [\boldsymbol{u}(0), \ldots, \boldsymbol{u}(T-1)] \tag{3}$$

Assuming that the number of data points T is sufficiently larger than the dimension of the state vector $\boldsymbol{x}$, this minimization problem is solved by Ordinary Least Squares (OLS):

$$\begin{pmatrix} \hat{\boldsymbol{A}} & \hat{\boldsymbol{B}} \end{pmatrix} = \boldsymbol{Y}\boldsymbol{Z}^{\dagger} \tag{4}$$

where the symbol $\hat{\cdot}$ denotes an estimated matrix. Here, we denote $(\boldsymbol{X}\ \boldsymbol{U})^{\top}$ as $\boldsymbol{Z}$. A matrix with a $\dagger$ symbol represents a pseudoinverse matrix.

To analyze the passive estimation (i.e., estimation without perturbation input), consider the case where the input $\boldsymbol{u}$ is the zero vector. The LTI system can be represented as a first-order vector autoregressive (VAR(1)) model, which is commonly used to describe time-series data in neuroscience [44],

$$\boldsymbol{x}(t+1) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{\xi}(t). \tag{5}$$

Passive estimation involves estimating the system matrix $\hat{\boldsymbol{A}}$ solely from the state observations $\boldsymbol{x}(t)$. Similarly to the case of estimation with perturbation, the conditional maximum likelihood estimation is to minimize the following cost function:

$$\min_{\boldsymbol{A}} \|\boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}\|_F^2 \tag{6}$$

The parameter matrix $\boldsymbol{A}$ can be obtained through OLS as follows:

$$\hat{\boldsymbol{A}} = \boldsymbol{Y}\boldsymbol{X}^{\dagger}. \tag{7}$$

The OLS estimation presented above provides a computationally tractable way to approximate neural dynamics as the LTI system. The estimated matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ enable systematic analysis of both autonomous network dynamics and of how external inputs drive the neural population responses.

2.2 Estimation Error in LTI System Identification

The estimation error of the OLS procedure is known to asymptotically follow a Gaussian distribution with a zero mean vector as the number of samples increases [34, 35]. In the following, we illustrate how this asymptotic variance theory is derived for the LTI system.

By considering multiple time steps and stacking the state and input variables over time, the system following Eq. (1) can be rewritten as:

$$\boldsymbol{Y} = (\boldsymbol{A} \quad \boldsymbol{B}) \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{U} \end{pmatrix} + \boldsymbol{\Xi} \tag{8}$$

$$= \boldsymbol{\Phi}\boldsymbol{Z} + \boldsymbol{\Xi} \tag{9}$$

where we denote $(\boldsymbol{A}\ \boldsymbol{B})$ as $\boldsymbol{\Phi}$, $(\boldsymbol{X}\ \boldsymbol{U})^\top$ as $\boldsymbol{Z}$, and $\boldsymbol{\Xi} = [\boldsymbol{\xi}(0), \ldots, \boldsymbol{\xi}(T-1)]$.

The matrix $\boldsymbol{\Phi}$ can be estimated using OLS, and the estimated $\hat{\boldsymbol{\Phi}}$ can be rewritten as the sum of the true matrix $\boldsymbol{\Phi}$ and the remaining terms:

$$\hat{\boldsymbol{\Phi}} = \boldsymbol{Y}\boldsymbol{Z}^{\dagger} = \boldsymbol{Y}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1} = (\boldsymbol{\Phi}\boldsymbol{Z} + \boldsymbol{\Xi})\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1} = \boldsymbol{\Phi} + \boldsymbol{\Xi}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1} \tag{10}$$

where a matrix with a $\dagger$ symbol represents a pseudoinverse matrix. The expression for the pseudoinverse $\boldsymbol{Z}^\dagger = \boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1}$ is derived under the assumption that $\boldsymbol{Z}$ has full row rank, ensuring that $\boldsymbol{Z}\boldsymbol{Z}^\top$ is invertible. The error in the coefficient matrix can be written as follows:

$$\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi} = \boldsymbol{\Xi}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1} = \left[\frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{\xi}(t)\boldsymbol{z}(t)^\top\right]\left[\frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{z}(t)\boldsymbol{z}(t)^\top\right]^{-1}. \tag{11}$$

Representing the noise vector $\boldsymbol{\xi}(t)$ in component form as $\boldsymbol{\xi}(t) = [\xi_1(t), \ldots, \xi_n(t)]^\top$, we define $\boldsymbol{\eta}(t)$ as the following $n(n+m) \times 1$ vector:

$$\boldsymbol{\eta}(t) = \begin{bmatrix} \xi_1(t)\boldsymbol{z}(t) \\ \vdots \\ \xi_n(t)\boldsymbol{z}(t) \end{bmatrix}. \tag{12}$$

Calculating the covariance matrix of this $\boldsymbol{\eta}(t)$,

$$\mathbb{E}(\boldsymbol{\eta}(t)\boldsymbol{\eta}(t)^\top) = \boldsymbol{\Omega} \otimes \boldsymbol{Q} \tag{13}$$

where $\otimes$ denotes the Kronecker product, and we define the following:

$$\boldsymbol{Q}_T = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{z}(t)\boldsymbol{z}(t)^\top \xrightarrow{p} \mathbf{Q} \quad \text{where} \quad \mathbf{Q} = \mathbb{E}(\boldsymbol{z}(t)\boldsymbol{z}(t)^\top). \tag{14}$$

$$\boldsymbol{\Omega}_T = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\xi}(t)\boldsymbol{\xi}(t)^\top \xrightarrow{p} \boldsymbol{\Omega} \quad \text{where} \quad \boldsymbol{\Omega} = \mathbb{E}(\boldsymbol{\xi}(t)\boldsymbol{\xi}(t)^\top). \tag{15}$$

Here, $\xrightarrow{p}$ indicates convergence in probability. As a result, the distribution of $\boldsymbol{\xi}(t)$ can be asymptotically approximated as follows:

$$\frac{1}{\sqrt{T}}\sum_{t}^{T}\boldsymbol{\eta}_t \xrightarrow{L} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega} \otimes \boldsymbol{Q}) \tag{16}$$

where $\xrightarrow{L}$ denotes convergence in distribution.

Rewriting Eq. (11) using $\boldsymbol{\eta}$ and $\boldsymbol{Q}_T$, we obtain the following equation:

$$\sqrt{T}(\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi}) = \left(\left[\frac{1}{\sqrt{T}} \sum_t^T \xi_1(t)\boldsymbol{z}(t)^\top, \cdots, \frac{1}{\sqrt{T}} \sum_t^T \xi_n(t)\boldsymbol{z}(t)^\top \right] \begin{bmatrix} \boldsymbol{Q}_T^{-1} & 0 & \cdots & 0 \\ 0 & \boldsymbol{Q}_T^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{Q}_T^{-1} \end{bmatrix} \right)^\top \tag{17}$$

$$= \begin{bmatrix} \boldsymbol{Q}_T^{-1} & 0 & \cdots & 0 \\ 0 & \boldsymbol{Q}_T^{-1} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{Q}_T^{-1} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{T}} \sum_t^T \xi_1(t)\boldsymbol{z}(t) \\ \vdots \\ \frac{1}{\sqrt{T}} \sum_t^T \xi_n(t)\boldsymbol{z}(t) \end{bmatrix} \tag{18}$$

$$= (\boldsymbol{I} \otimes \boldsymbol{Q}_T^{-1}) \frac{1}{\sqrt{T}} \sum_t^T \boldsymbol{\eta}(t) \tag{19}$$

By considering Eq. (14) and (16), the covariance of the distribution in Eq. (19) can be asymptotically approximated as follows:

$$(\mathbf{I} \otimes \mathbf{Q}^{-1})(\boldsymbol{\Omega} \otimes \mathbf{Q})(\mathbf{I} \otimes \mathbf{Q}^{-1}) = (\mathbf{I} \otimes \mathbf{Q}^{-1})\boldsymbol{\Omega}(\mathbf{I} \otimes \mathbf{Q}^{-1}) = \boldsymbol{\Omega} \otimes \mathbf{Q}^{-1}. \tag{20}$$

Therefore, the estimation error of $\boldsymbol{\Phi}$ asymptotically follows the normal distribution given by:

$$\hat{\boldsymbol{\Phi}} - \boldsymbol{\Phi} \xrightarrow{p} \mathcal{N}(\mathbf{0}, \frac{1}{T}(\boldsymbol{\Omega} \otimes \boldsymbol{Q}^{-1})). \tag{21}$$

2.3 Error Minimization Criteria for LTI System

To reduce the elements of the covariance matrix covariance matrix of Eq. (21), it is necessary to minimize the elements of $\boldsymbol{\Omega} \otimes \boldsymbol{Q}^{-1}$ or maximize the number of data points T, which in turn improves the accuracy of estimating the matrix $\hat{\boldsymbol{\Phi}}$. The asymptotic distribution of the variance can also be explained using the Cramér-Rao bound with the Fisher information matrix [45–47], which provides the lower bound of the variance of unbiased estimators.

The Cramér-Rao bound can be expressed as follows [37, 48]:

$$\mathrm{Cov}(\hat{\boldsymbol{\Phi}}) \succeq \mathcal{I}(\boldsymbol{\Phi})^{-1}, \tag{22}$$

where $\mathcal{I}(\boldsymbol{\Phi})^{-1}$ is the inverse of the Fisher information matrix. The notation $\succeq$ indicates a matrix inequality, which implies that the matrix $\mathrm{Cov}(\hat{\boldsymbol{\Phi}}) - \mathcal{I}(\boldsymbol{\Phi})^{-1}$ is positive semidefinite. By carefully designing the perturbation inputs or experimental conditions, it is possible to enhance the Fisher information matrix $\mathcal{I}(\boldsymbol{\Phi})$ in a way that improves the accuracy of parameter estimation. Various strategies to design such conditions have been proposed in the design of experiment literature [36].

Here, we summarize some representative approaches:

1. A-optimality: Minimizes the sum of the variances of parameter estimates, corresponding to minimizing the trace of $\mathcal{I}(\boldsymbol{\Phi})^{-1}$:

$$\text{Objective: } \min \mathrm{tr}(\mathcal{I}(\boldsymbol{\Phi})^{-1}) \tag{23}$$

2. D-optimality: Maximizes the determinant of $\mathcal{I}(\boldsymbol{\Phi})$, which minimizes the volume of the confidence ellipsoid:

$$\text{Objective: } \max \det(\mathcal{I}(\boldsymbol{\Phi})) \tag{24}$$

3. E-optimality: Maximizes the smallest eigenvalue of $\mathcal{I}(\boldsymbol{\Phi})$, ensuring robustness by reducing the worst-case variance:

$$\text{Objective: } \max \lambda_{\min}(\mathcal{I}(\boldsymbol{\Phi})) \tag{25}$$

These criteria provide guidelines for designing perturbation inputs that lead to accurate parameter estimation of the LTI system.

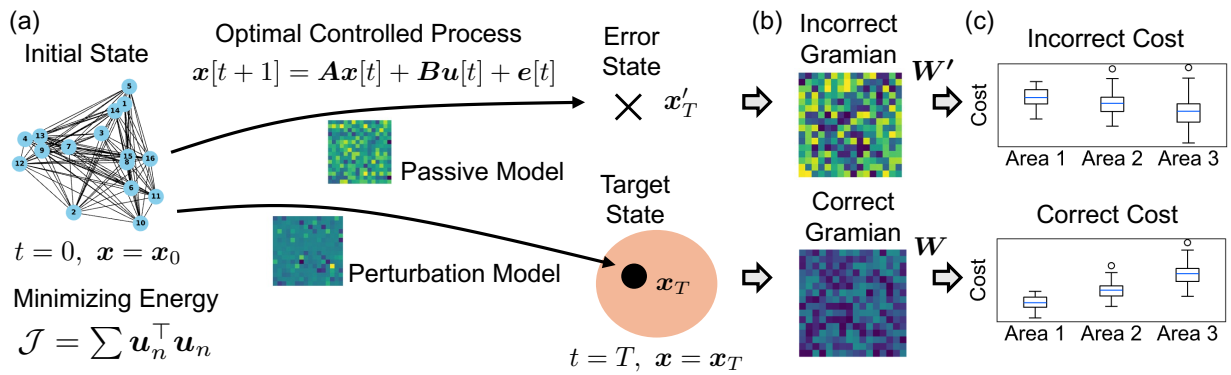

Fig 2: Effects of system identification on network control theory. (a) State transitions with the identified models by passive and perturbation states. (b) Controllability Gramians with two models. (c) Control costs calculated by the estimated Gramians.

2.4 Optimal Control and Network Controllability

While previous explanations have focused on the estimation of dynamical models in neural systems, it is also worth briefly considering how these estimated parameters influence optimal control strategies in neuroscience. Accurately estimating matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ under perturbation not only facilitates precise inference of the neural dynamics but also contributes to the accurate estimation of optimal control inputs for neural state transitions and the precise evaluation of controllability. As preparation for a demonstration of benefit in the optimal control theory (see Section 5), we review here the basic control theory framework in neuroscience. The approach utilizes the controllability Gramian and control costs, and provides insights into the efficiency and feasibility of inducing specific neural state transitions [38, 39] with some limitations [49, 50]. Hereafter, we refer to the input for system identification as the *perturbation input* and the input for state control in the optimal control theory as the *control input*.

The optimal control input is derived under the condition of assuming the linear system is controllable. We consider the sequence of control inputs that minimizes the following cost function (input energy minimization) while driving the state from the initial state $\boldsymbol{x}_0$ to the terminal state $\boldsymbol{x}_T$:

$$\mathcal{M}(\{\boldsymbol{u}(t)\}_{t=0}^{T-1}) = \sum_{t=0}^{T-1} \boldsymbol{u}(t)^\top \boldsymbol{u}(t) \tag{26}$$

under

$$\boldsymbol{x}(0) = \boldsymbol{x}_0, \quad \boldsymbol{x}(T) = \boldsymbol{x}_T. \tag{27}$$

Here, the initial and terminal state conditions represent constraints on the optimization problem. This problem is expressed as follows:

$$\boldsymbol{u}(t)^* = \arg\min_{\boldsymbol{u}} \mathcal{M}(\{\boldsymbol{u}(t)\}_{t=0}^{T-1}). \tag{28}$$

The constrained optimization problem can be solved using Lagrange multipliers. The optimal control input is given by

$$\boldsymbol{u}(t)^* = \boldsymbol{B}^\top (\boldsymbol{A}^\top)^{T-t-1} \boldsymbol{W}^{-1} (\boldsymbol{x}(T) - \boldsymbol{A}^T \boldsymbol{x}(0)) \tag{29}$$

where $\boldsymbol{W}$ is the controllability Gramian, defined as:

$$\boldsymbol{W} = \sum_{t=0}^{T-1} \boldsymbol{A}^t \boldsymbol{B}\boldsymbol{B}^\top (\boldsymbol{A}^\top)^t \tag{30}$$

This is a simplified form of controllability Gramian in the linear stochastic system as follows,

$$\boldsymbol{M} = \sum_{t=0}^{\infty} \boldsymbol{A}^t \boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^\top \boldsymbol{B}^\top (\boldsymbol{A}^\top)^t \tag{31}$$

where the noise terms are assumed to be temporally independent ($\boldsymbol{C} = \boldsymbol{I}$) [41].

We can express the optimal input energy in terms of the controllability Gramian and $\boldsymbol{x}_d = \boldsymbol{x}(T) - \boldsymbol{A}^T \boldsymbol{x}(0)$ as follows:

$$\mathcal{M}(\{\boldsymbol{u}(t)\}_{t=0}^{T-1}) = \boldsymbol{x}_d^\top \boldsymbol{W}^{-1} \boldsymbol{x}_d \tag{32}$$

Here, $\boldsymbol{x}_d$ represents the difference between the desired terminal state $\boldsymbol{x}(T)$ and the state predicted by the system dynamics when starting from $\boldsymbol{x}(0)$ without control input.

The controllability Gramian plays a pivotal role in determining the optimal control and control input cost. It has been used to identify the functional roles of individual brain regions [38, 51]. Eq. (30) utilizes the dynamical model matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Therefore, using perturbations, $\boldsymbol{A}$ and $\boldsymbol{B}$ are accurately estimated, which in turn allows precise determination of the optimal inputs and the corresponding controlled state transitions (Fig 2 (a)). Moreover, use of the perturbation approach to derive the model allows the controllability Gramian obtained from Eq. (30), as well as the estimation of control costs for each network area, to be more reliably assessed (Fig 2 (b)(c)). To accurately control neural states and understand neural functions, precise estimation of $\boldsymbol{A}$ and $\boldsymbol{B}$ is essential to the application of network control theory in neuroscience.

3 Design of Perturbation Input for System Identification

In this section, we present an unexplored approach in neuroscience in which we derive the theoretical framework for determining the optimal perturbation inputs to enhance the estimation of model parameters in the LTI system. We begin by providing an intuitive demonstration of our approach that perturbation inputs can re-excite weakly observable dynamics and improve the model estimation. Following this, we derive the objective function to design the optimal perturbation input for minimizing the estimation errors of $\boldsymbol{A}$ and $\boldsymbol{B}$. Next, we apply this theory to impulse, step, and frequency inputs, as typically assumed in neuroscience applications such as TMS, tDCS, and tACS, and derive the relationships between the perturbation input parameters and estimation errors. Finally, we perform simulations to demonstrate the optimal frequency and location of perturbation inputs relative to our theory. These explorations demonstrate how perturbations can be effectively designed to enhance the accuracy of parameter estimates in complex neural network models.

3.1 Intuitive Demonstration of Dynamics Re-excitation and Accurate Parameter Estimation

Using a simple simulation, we demonstrate how perturbation inputs can intuitively enhance system identification accuracy. Weakly observable dynamics in a system can be re-excited by applying appropriate perturbations, thereby enriching the available information and improving the accuracy of parameter estimation. Here, we consider an LTI system represented by the connectivity matrix $\boldsymbol{A}$, as shown in Fig 3 (a). For simplicity, the input matrix $\boldsymbol{B}$ is assumed to be an identity matrix. We assume that the system operates under minimal noise conditions. The eigenvalues of matrix $\boldsymbol{A}$ correspond to modes at 10 Hz and 40 Hz. Over time, the dynamics associated with the 40 Hz mode decay, leading to its near disappearance from the observed time-series data. This phenomenon is depicted in Fig 3 (b), where one cycle of a 10 Hz frequency is marked in the plot, and the oscillations in the states x_3 and x_4 almost completely disappear. When principal component analysis (PCA) is applied to this time-series data, the first and second principal components reflect the 10 Hz dynamics, whereas the information along the third principal component related to the 40 Hz approaches zero. Estimating matrix $\boldsymbol{A}$ from this time-series data reveals significant errors in the lower-right components of the matrix.

This difficulty can be resolved by applying perturbation inputs. Fig 3 (c) shows the application of an impulse input $\boldsymbol{u}$ to the system. This perturbation input primally affects the nodes corresponding to x_3 and x_4, reintroducing variations in the third principal component direction. Estimating matrix $\boldsymbol{A}$ from this perturbed time-series data yields a more accurate representation than with the first case. Similarly, improvement is observed when different types of input—a frequency input for example—is used instead of the impulse input, as shown in Fig 3 (d). These results illustrate that appropriate perturbation inputs can effectively re-excite weak dynamics, thereby enhancing parameter estimation accuracy.

3.2 An Objective Function for Optimal System Identification

In this section, we derive an objective function for optimizing perturbation inputs to identify an LTI system with minimized errors. We show that the errors of parameters A and B, expressed in terms of the Frobenius norm, are proportional to the ratio of the trace of the covariance matrix of the noise $\boldsymbol{\Xi}$ and the covariance matrix of matrix $\boldsymbol{Z}$.

First, from Eq. (9), the true $\boldsymbol{\Phi}$ is as follows:

$$\boldsymbol{\Phi} = (\boldsymbol{Y} - \boldsymbol{\Xi})\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1}. \tag{33}$$

The estimate of $\boldsymbol{\Phi}$ is obtained by the OLS as follows:

$$\hat{\boldsymbol{\Phi}} = \boldsymbol{Y}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1}. \tag{34}$$

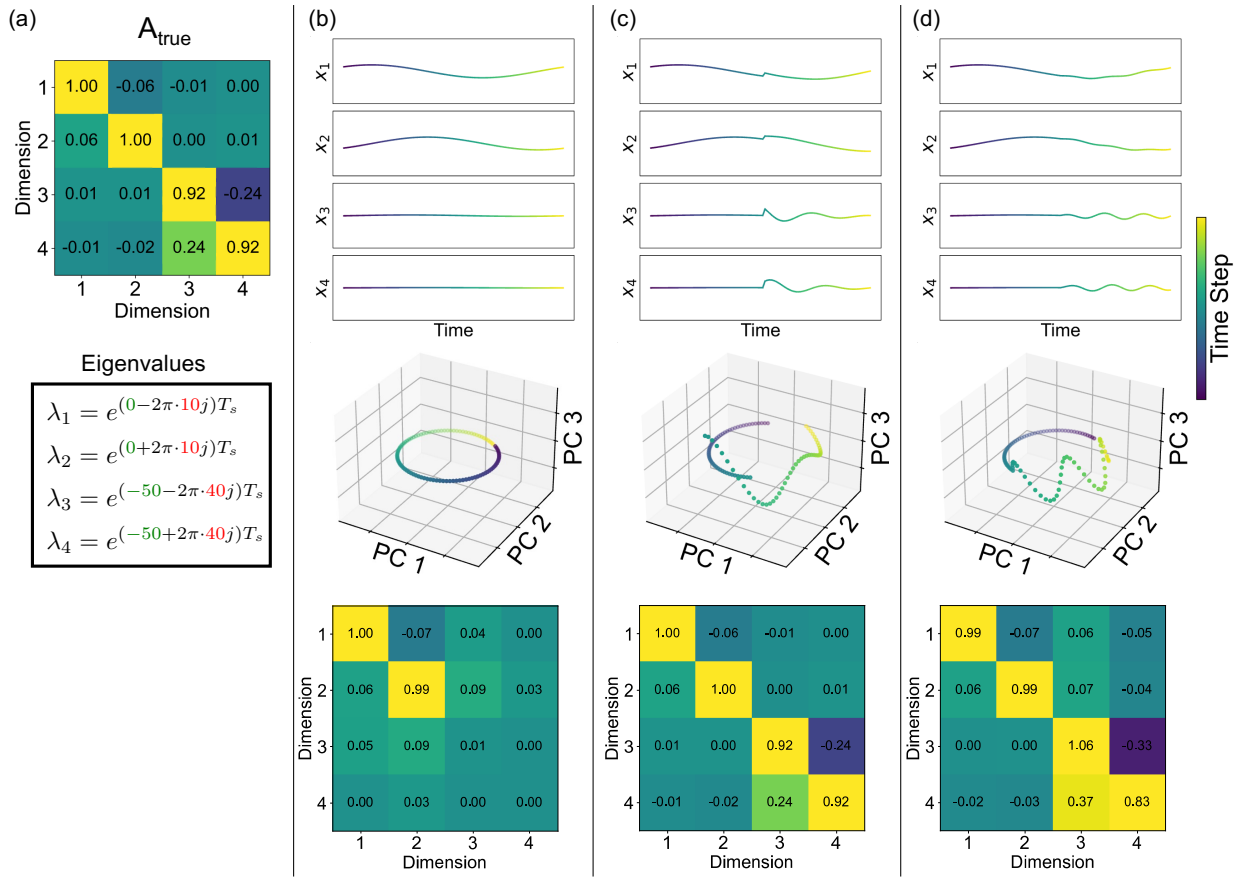

Fig 3: Visualization of continuous-time system dynamics by perturbations. (a) Matrix $\boldsymbol{A}$ represents the dynamics and its eigenvalues. (b) Temporal changes in the state variables of the passive state without perturbation and variations in each PCA component. The color gradient represents temporal change. (c) Temporal changes in the state variables and variations in each PCA component induced by an impulse input. (d) Temporal changes in the state variables and variations in each PCA component induced by a frequency input.

From these, the difference between the true $\boldsymbol{\Phi}$ and the estimator $\hat{\boldsymbol{\Phi}}$, expressed in terms of the Frobenius norm, is given by

$$\|\boldsymbol{\Phi} - \hat{\boldsymbol{\Phi}}\|_F^2 = \|\boldsymbol{\Xi}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top)^{-1}\|_F^2 = \mathrm{tr}\left[\boldsymbol{\Xi}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top)^{-2}\boldsymbol{Z}\boldsymbol{\Xi}^\top\right]. \tag{35}$$

By performing singular value decomposition (SVD) of $\boldsymbol{Z}$, we can rewrite Eq. (35). The SVD of $\boldsymbol{Z}$ is given by

$$\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top, \tag{36}$$

where $\boldsymbol{U}$, $\boldsymbol{S}$ are $(n+m)\times(n+m)$ matrices, and $\boldsymbol{V}$ is a $T\times(n+m)$ matrix. Here, $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{U}^\top\boldsymbol{U} = I$, $\boldsymbol{V}\boldsymbol{V}^\top = I$, and $\boldsymbol{S}$ is a diagonal matrix. Using this, we can rewrite Eq. (35) as follows:

$$\begin{aligned}
\|\boldsymbol{\Phi} - \hat{\boldsymbol{\Phi}}\|_F^2 &= \mathrm{tr}\left[\boldsymbol{\Xi}\boldsymbol{Z}^\top(\boldsymbol{Z}\boldsymbol{Z}^\top)^{-2}\boldsymbol{Z}\boldsymbol{\Xi}^\top\right] && (37)\\
&= \mathrm{tr}\left[\boldsymbol{\Xi}(\boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^\top)(\boldsymbol{U}\boldsymbol{S}^{-4}\boldsymbol{U}^\top)(\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top)\boldsymbol{\Xi}^\top\right] && (38)\\
&= \mathrm{tr}\left[\boldsymbol{\Xi}\boldsymbol{V}\boldsymbol{S}^{-2}\boldsymbol{V}^\top\boldsymbol{\Xi}^\top\right] && (39)\\
&= \sum_{i=1}^{n+m}\frac{1}{s_i^2}\|\boldsymbol{\Xi}\boldsymbol{v}_i\|^2. && (40)
\end{aligned}$$

Here, $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_{n+m}$ are the column vectors of $\boldsymbol{V}$, which are orthogonal and of length 1, and are T-dimensional vectors. Assuming that $\boldsymbol{\Xi}$ is independent at each time step (i.e., $E[\boldsymbol{\Xi}_{i,j,k}\boldsymbol{\Xi}_{i',j',k'}] = 0$ for $k \neq k'$), Eq. (40) can be asymptotically evaluated as follows:

$$\|\boldsymbol{\Phi} - \hat{\boldsymbol{\Phi}}\|_F^2 \xrightarrow{p} \sum_{i=1}^{n+m}\frac{1}{s_i^2}\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Xi}}) \tag{41}$$

Here, $\xrightarrow{p}$ indicates convergence in probability and $\boldsymbol{\Sigma}_{\boldsymbol{\Xi}}$ represents the covariance matrix of the noise $\boldsymbol{\Xi}$. The detailed derivation of this equation is provided in the Supplementary Material. From the above, the estimation error of $\boldsymbol{\Phi}$ can be written using the covariance matrix of $\boldsymbol{\Xi}$ and the singular values of data $\boldsymbol{Z}$. The singular values s_i of $\boldsymbol{Z}$ are related to the eigenvalues λ_i of the covariance matrix of $\boldsymbol{Z}$ as $\lambda_i = s_i^2/T$. Thus, the above equation can be re-written as

$$\|\boldsymbol{\Phi} - \hat{\boldsymbol{\Phi}}\|_F^2 \xrightarrow{p} \sum_i \frac{\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Xi}})}{T\lambda_i} = \frac{1}{T}\,\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Xi}})\,\mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}\right) \tag{42}$$

Since the system noise covariance $\boldsymbol{\Sigma}_{\boldsymbol{\Xi}}$ and the measurement time T are independent of the perturbation input, the objective function for optimizing the perturbation input to minimize the estimation error Eq. (42) is defined as

$$\mathcal{J}(\boldsymbol{u}) = \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1}). \tag{43}$$

By designing $\boldsymbol{u}(t)$ to minimize the objective function, the estimation accuracy of $\boldsymbol{\Phi}$ is improved, leading to more accurate estimation of the system matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

We interpret this result from the perspectives of asymptotic variance theory and the design of experiments, as explained in the Background section. When the time horizon is sufficiently long, the left-hand side of the inequality in Eq. (22) converges to the right-hand side [37, 52, 53]. In other words, the covariance matrix described in Eq. (21) converges to the inverse of the Fisher information matrix. In the design of experiments theory, the A-optimality criterion aims to minimize the trace of the inverse of the Fisher information matrix. This coincides with the last element of Eq. (42).

3.3 Clarifying the Relationship Between the Objective Function and Perturbation Input

In this section, we reformulate the objective function $\mathcal{J}(\boldsymbol{u})$ derived in the previous section to clarify its relationship with the perturbation input $\boldsymbol{u}$. This reformulation leads to an intuitive interpretation of the relationship between parameter estimation and perturbation input.

To simplify the analysis, we assume that $\boldsymbol{B}$ is known. Then, the estimation error Eq. (42) is expressed as

$$\|\boldsymbol{A} - \hat{\boldsymbol{A}}\|_F^2 \xrightarrow{p} \frac{1}{T}\,\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Xi}})\,\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}). \tag{44}$$

The derivation is given in the Supplementary Material. As in the derivation of Eq. (43), since the system noise covariance $\boldsymbol{\Sigma}_{\boldsymbol{\Xi}}$ is independent of the input, the objective function for optimizing the input to minimize the estimation error is defined as

$$\mathcal{J}(\boldsymbol{u}) = \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}). \tag{45}$$

We have derived Eq. (45) to characterize the estimation error of $\boldsymbol{A}$, but further analytical investigation of its dependence on the input is challenging. To address this, rather than minimizing $\mathcal{J}(\boldsymbol{u}) = \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1})$, we consider maximizing $\mathcal{J}'(\boldsymbol{u}) = \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{X}})$ as a proxy. While this substitution is not strictly valid, it is less problematic because $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ is a positive semidefinite matrix.

The trace of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ can be expressed as the sum of the squared norms of the state vector $\boldsymbol{x}(t)$ over all time steps t:

$$\frac{1}{T}\,\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{X}}) = \mathbb{E}[\boldsymbol{x}(t)^\top \boldsymbol{x}(t)], \tag{46}$$

where $\boldsymbol{x}(t)$ is the state of the system at time t and can be estimated by Eq. (1) as follows:

$$\boldsymbol{x}(t) = \boldsymbol{A}^t \boldsymbol{x}(0) + \sum_{k=0}^{t-1} \boldsymbol{A}^{t-1-k}\boldsymbol{B}\boldsymbol{u}(k) + \sum_{k=0}^{t-1} \boldsymbol{A}^{t-1-k}\boldsymbol{\xi}(k). \tag{47}$$

Substituting Eq. (47) into Eq. (46), we obtain:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{x}(t)^\top \boldsymbol{x}(t)] = {} & \mathbb{E}\left[\boldsymbol{x}(0)^\top (\boldsymbol{A}^t)^\top \boldsymbol{A}^t \boldsymbol{x}(0)\right] + 2\mathbb{E}\left[\boldsymbol{x}(0)^\top (\boldsymbol{A}^t)^\top \sum_{k=0}^{t-1} \boldsymbol{A}^{t-1-k}\boldsymbol{B}\boldsymbol{u}(k)\right] \\
& + 2\mathbb{E}\left[\boldsymbol{x}(0)^\top (\boldsymbol{A}^t)^\top \sum_{k=0}^{t-1} \boldsymbol{A}^{t-1-k}\boldsymbol{\xi}(k)\right] + \mathbb{E}\left[\sum_{k=0}^{t-1} \boldsymbol{u}(k)^\top \boldsymbol{B}^\top (\boldsymbol{A}^{t-1-k})^\top \boldsymbol{A}^{t-1-k}\boldsymbol{B}\boldsymbol{u}(k)\right] \\
& + 2\mathbb{E}\left[\sum_{k=0}^{t-1} \boldsymbol{\xi}(j)^\top (\boldsymbol{A}^{t-1-j})^\top \sum_{j=0}^{t-1} \boldsymbol{A}^{t-1-k}\boldsymbol{B}\boldsymbol{u}(k)\right] + \mathbb{E}\left[\sum_{k=0}^{t-1} \boldsymbol{\xi}(k)^\top (\boldsymbol{A}^{t-1-k})^\top \boldsymbol{A}^{t-1-k}\boldsymbol{\xi}(k)\right].
\end{aligned} \tag{48}$$

Assuming that only the perturbation input $\boldsymbol{u}$ can be manipulated, the proxy objective function $\mathcal{J}'(\boldsymbol{u})$, which is maximized to approximately minimize the estimation error, is expressed as follows:

$$\mathcal{J}'(\boldsymbol{u}) = 2\mathbb{E}\left[\left(\sum_{j=0}^{t-1}\boldsymbol{\xi}(j)^\top(\boldsymbol{A}^{t-1-j})^\top + \boldsymbol{x}(0)^\top(\boldsymbol{A}^t)^\top\right)\sum_{k=0}^{t-1}\boldsymbol{A}^{t-1-k}\boldsymbol{B}\boldsymbol{u}(k)\right] + \mathbb{E}\left[\sum_{k=0}^{t-1}\|\boldsymbol{A}^{t-1-k}\boldsymbol{B}\boldsymbol{u}(k)\|^2\right]. \tag{49}$$

Therefore, the reformulated function provides a clearer understanding of how the perturbation input $\boldsymbol{u}$ influences the parameter estimation. Accordingly, it provides that the relationship with true parameters, particularly with $\boldsymbol{A}$, becomes crucial to designing the optimal perturbation input. In the next section, we further elaborate on the perturbation inputs commonly used in neuroscience by using this objective function.

3.4 Designing Optimal Neural Stimulation

In this section, we further express the perturbation input $\boldsymbol{u}$ in forms that correspond to commonly used stimulation methods in neuroscience. We focus on impulse, step, and frequency inputs, which correspond to well-known neural stimulation methods such as TMS, tDCS, and tACS. Among these, the frequency input exhibits relatively more complexity, and thus optimal frequency is derived first. Following this, the impulse and step inputs, which adhere to simpler principles, are then addressed.

3.4.1 Frequency Input

When the perturbation input is sinusoidal, $\boldsymbol{u}(t)$ is given by:

$$\boldsymbol{u}(t) = a\sin(\omega t)\boldsymbol{u}_0. \tag{50}$$

where a is the amplitude and ω is the angular frequency. The term $\boldsymbol{u}_0$ represents the direction of the input in the state space or system input space. Using the Euler formula, the sine function can be expressed in terms of complex exponentials:

$$\sin(\omega k) = \frac{1}{2i}\left(e^{i\omega k} - e^{-i\omega k}\right). \tag{51}$$

Thus, the perturbation input becomes:

$$\boldsymbol{u}(k) = \frac{a}{2i}\left(e^{i\omega k} - e^{-i\omega k}\right)\boldsymbol{u}_0. \tag{52}$$

Next, we express system matrix $\boldsymbol{A}$ in terms of its eigenvalue decomposition:

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^{-1}, \tag{53}$$

where $\boldsymbol{V}$ is the matrix of eigenvectors of $\boldsymbol{A}$, and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Using this decomposition, powers of $\boldsymbol{A}^k$ can be written as:

$$\boldsymbol{A}^k = \boldsymbol{V}\boldsymbol{\Lambda}^k\boldsymbol{V}^{-1}, \tag{54}$$

We focus specifically on the term involving the perturbation input $\boldsymbol{u}$, which is the component we can directly manipulate. By substituting $\boldsymbol{A}^k$ into the part of $\mathcal{J}'(\boldsymbol{u})$, this can be expressed as follows:

$$\begin{aligned}\mathcal{J}'(\boldsymbol{u})_{\text{part}} &= \boldsymbol{A}^{t-1-k}\boldsymbol{B}\boldsymbol{u}(k)\\ &= \frac{a}{2i}\sum_{k=0}^{t-1}\boldsymbol{V}\boldsymbol{\Lambda}^{t-1-k}\boldsymbol{V}^{-1}\boldsymbol{B}\left(e^{i\omega k} - e^{-i\omega k}\right)\boldsymbol{u}_0\end{aligned} \tag{55}$$

The diagonal matrix $\boldsymbol{\Lambda}^{t-1-k}$ is expressed as follows:

$$\boldsymbol{\Lambda}^{t-1-k} = \text{diag}\left(\lambda_1^{t-1-k}, \lambda_2^{t-1-k}, \ldots, \lambda_n^{t-1-k}\right), \tag{56}$$

where each eigenvalue λ_j, corresponding to the eigenvalues of the discrete-time system matrix $\boldsymbol{A}$, is represented in polar form as:

$$\lambda_j = |\lambda_j|e^{i\theta_j}. \tag{57}$$

Here, $|\lambda_j|$ denotes the magnitude of the eigenvalue, and θ_j represents the phase (angular frequency) of the eigenvalue. The phase θ_j corresponds to the imaginary part of the eigenvalue of the continuous-time system matrix $\boldsymbol{A}$.

By substituting these expressions into the terms in Eq. (55), the following expression can be derived.

$$\mathcal{J}'(\boldsymbol{u})_{\text{part}} = \frac{a}{2i}\sum_{k=0}^{t-1} \boldsymbol{V}\text{diag}\left(|\lambda_j|^{t-1-k}e^{i[(t-1-k)\theta_j+\omega k]} - |\lambda_j|^{t-1-k}e^{i[(t-1-k)\theta_j-\omega k]}\right)\boldsymbol{V}^{-1}\boldsymbol{B}. \tag{58}$$

This equation is proportional to $e^{i[(t-1-k)\boldsymbol{\theta}_j+\omega k]}$, with which the cumulative effect becomes evident when the frequency of the periodic input ω matches the angular frequency $\boldsymbol{\theta}_j$. The term becomes $e^{i(t-1)\boldsymbol{\theta}_j}$, which is independent of k. The result can also be derived from the continuous-time equations (see the Supplementary Material). When the dynamics have a single mode or a dominant mode, the optimal perturbation input aligns directly with the frequency θ_1. Thus, the optimal perturbation input to maximize the objective function is:

$$\boldsymbol{u}^* = a\sin(\theta_1 t)\boldsymbol{u}_0. \tag{59}$$

When the dynamics have multiple modes, the optimal perturbation input needs to account for the contributions from all significant modes, potentially requiring a combination of frequencies and spatial patterns. When the frequencies of these modes are close to each other, the optimal frequency falls between them. Conversely, when the frequencies are sufficiently distinct, the optimal frequency aligns with each mode individually, maximizing the contribution of each specific mode. Which case applies depends on the characteristics of the dynamics. Therefore, when multiple modes are present, it becomes necessary to numerically determine the optimal frequencies to account for all significant modes.

3.4.2 Impulse and Step Input

Impulse and step inputs follow simple rules. Let α represent the scaling factor for impulse input. Then, input $\boldsymbol{u}(t)$ is defined as:

$$\boldsymbol{u}(t) = \begin{cases} \alpha\boldsymbol{u}_0 & \text{if } t = 0, \\ 0 & \text{if } t > 0. \end{cases} \tag{60}$$

Substituting Eq. (60) into Eq. (49), we obtain

$$\mathcal{J}'(\boldsymbol{u}) = 2\alpha\mathbb{E}\left[\left(\sum_{j=0}^{t-1}\boldsymbol{\xi}(j)^\top(\boldsymbol{A}^{t-1-j})^\top + \boldsymbol{x}(0)^\top(\boldsymbol{A}^t)^\top\right)\boldsymbol{A}^{t-1}\boldsymbol{B}\boldsymbol{u}_0\right] + \alpha^2\mathbb{E}\left[\|\boldsymbol{A}^{t-1}\boldsymbol{B}\boldsymbol{u}_0\|^2\right]. \tag{61}$$

For a scaled step input, $\boldsymbol{u}(t)$ is defined as

$$\boldsymbol{u}(t) = \beta\boldsymbol{u}_0, \quad \forall t \geq 0. \tag{62}$$

where β is input strength. We then get

$$\mathcal{J}'(\boldsymbol{u}) = 2\beta\mathbb{E}\left[\left(\sum_{j=0}^{t-1}\boldsymbol{\xi}(j)^\top(\boldsymbol{A}^{t-1-j})^\top + \boldsymbol{x}(0)^\top(\boldsymbol{A}^t)^\top\right)\sum_{k=0}^{t-1}\boldsymbol{A}^{t-1-k}\boldsymbol{B}\boldsymbol{u}_0\right] + \beta^2\mathbb{E}\left[\sum_{k=0}^{t-1}\|\boldsymbol{A}^{t-1-k}\boldsymbol{B}\boldsymbol{u}_0\|^2\right]. \tag{63}$$

In both the impulse and step input cases, the objective function $\mathcal{J}'(\boldsymbol{u})$ increases as the input strength (α or β) increases because the contribution of the first-order term in α and β becomes negligible relative to the quadratic term α^2 and β^2 when these coefficients are sufficiently large. These findings indicate that designing impulse and step inputs with greater intensity leads to more accurate system identification.

3.5 Estimation of Optimal Perturbation Input

This section illustrates the relationship between the system's intrinsic parameters and the optimal perturbation input parameters in two representative scenarios through simulation. The first scenario examines the optimal frequency of a frequency input. We exemplify the system's intrinsic frequencies as playing a crucial role in determining the optimal perturbation input frequency. The second focuses on determining the optimal location for applying an impulse input under the condition that it can only be applied to a single target node. We demonstrate that the optimal location (node) is the one exerting the greatest influence on other nodes, enabling the most accurate system identification.

3.5.1 Optimal Frequency Inputs for System Identification

In this section, we exemplify the relationship between the system's intrinsic frequencies and the optimal input frequency. When the system has a single mode, the optimal input frequency corresponds to the imaginary part of the eigenvalues of matrix $\boldsymbol{A}$. When the system has multiple modes, the optimal frequency lies between the system's intrinsic frequencies or is biased toward one of them. In the simulations, we numerically solve the exact objective function

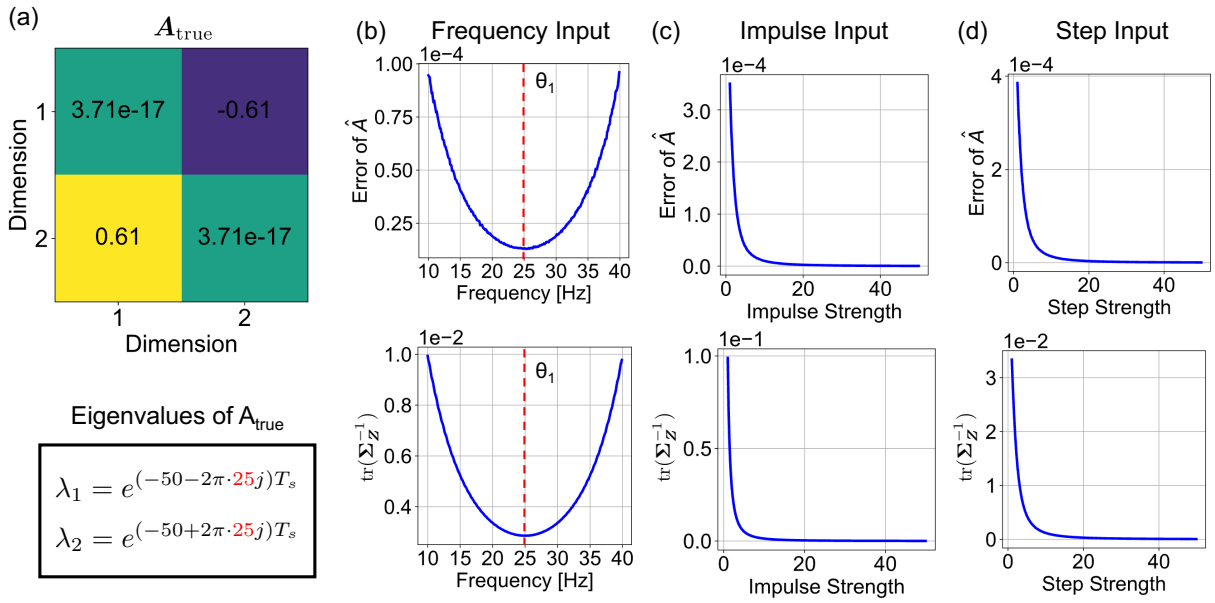

Fig 4: Estimation errors in relation to perturbation input frequency. (a) True matrix $\boldsymbol{A}$ and its eigenvalues. (b) Estimation error of $\boldsymbol{A}$ from frequency input and the trace of $\boldsymbol{Z}$'s inverse covariance matrix. (c) Impulse input strength and estimation errors. (d) Step input strength and estimation errors.

shown in Eq. (42) for the optimal design of perturbation input, albeit that we derived an intuitive understanding of what frequency would be optimal in Eq. (59).

We generated signals using Eq. (1). The matrix $\boldsymbol{A}$ had eigenvalues as shown in Fig 4 (a), and the matrix $\boldsymbol{B}$ was set as an identity matrix. The sampling frequency was set to $f_s = 100$ Hz, and the time step was $T_s = 1/f_s$. The noise vector $\boldsymbol{\xi}$ was generated from a multivariate normal distribution with zero mean and a covariance matrix scaled by 0.1^2 times the identity matrix. The simulation ran for $T = 1$ second and was executed for 250,000 trials. The perturbation input $\boldsymbol{u}$ was randomly applied to one node per trial and 50 trials were concatenated to calculate the parameters. This prevents an increase in the condition number and enables simulation of the optimal frequency in accordance with the theoretical framework. The results including the issues are provided in the Supplementary Material. Parameter estimation was performed over 5000 trials, and the average error was reported as the result. The true matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, along with the perturbation inputs, were used to generate the signal $\boldsymbol{x}$, and then $\hat{\boldsymbol{A}}$ was estimated from the signal using Eq. (4). The amplitude of the frequency input was set to 1, and the frequency varied from 10 to 40 in steps of 0.1. The estimation error of the model parameters was evaluated using the Frobenius norm, and the results were averaged over all trials. Additionally, we also applied impulse and step inputs. The input strength varied from 1 to 50 in steps of 0.1.

As shown in Fig 4 (b), the simulation results indicate that a frequency input matching the system's eigenfrequencies (25 Hz) minimizes the error of $\boldsymbol{A}$. This trend was proportional to $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1})$, as derived by our proposed theory. Intuitively, this occurs because increasing the information in $\boldsymbol{Z}$ also enhances the information in $\boldsymbol{X}$, which directly impacts the estimation of $\boldsymbol{A}$. For details on $\boldsymbol{B}$ and $\boldsymbol{\Phi}$, refer to the Supplementary Material. This simulation demonstrates that the frequency inherent to the dynamics minimizes parameter errors most effectively.

For the impulse and step inputs, the estimation errors decrease as a function of input strength (Fig4 (d)(e)). These trends are proportional to $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1})$. Although these results may appear straightforward, the attenuation of parameter errors is proportional to the square of the stimulation intensity, as shown in Eq. (61) and (63). This indicates that even slight adjustments in stimulation intensity can lead to a substantial reduction in parameter errors.

To investigate the relationship in networks with more nodes, we extended our analysis beyond the two-dimensional case to a four-dimensional network. Simulations were conducted using a network with frequency modes set at 10 Hz and 40 Hz, while maintaining all parameters consistent with the two-dimensional case, except for the matrix $\boldsymbol{A}$ and the sampling frequency ($f_s = 100$ Hz). In the four-dimensional model, the eigenvalue spectrum of $\boldsymbol{A}$, which included two conjugate pairs of eigenvalues, is shown in Fig 5 (a).

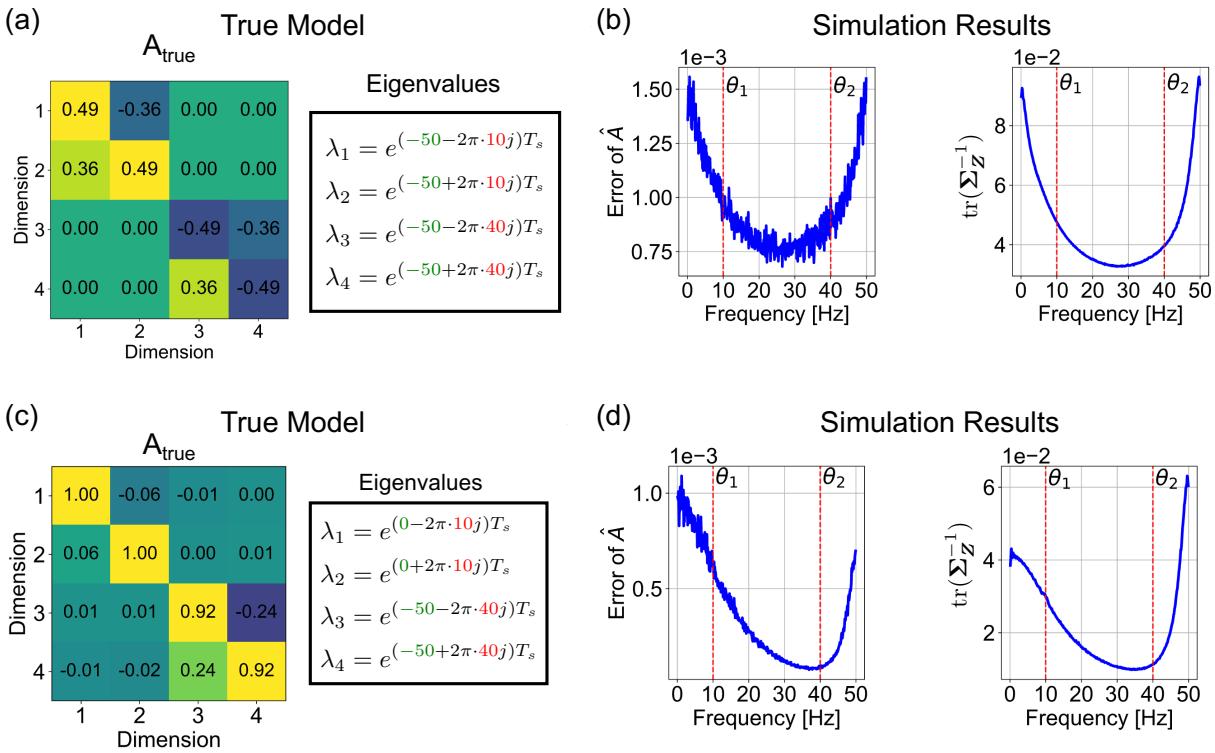

Fig 5: Estimation errors for a network with two modes. (a) True matrix $\boldsymbol{A}$ and its eigenvalues. Eigenvalues corresponding to 10 Hz and 40 Hz oscillations. Both modes are attenuated. (b) Relationship between the frequency of perturbation inputs and the estimation errors of $\boldsymbol{A}$, and $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1})$ when both modes are equally attenuated. (c) True matrix $\boldsymbol{A}$ and its eigenvalues. Eigenvalues corresponding to 10 Hz and 40 Hz oscillations. Only one mode is attenuated. (d) Relationship between the frequency of perturbation inputs and the estimation errors of $\boldsymbol{A}$, and $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1})$ when only one mode is attenuated.

The simulation results revealed that the estimation error of $\boldsymbol{A}$ is minimized by a unique input frequency between the resonant frequencies (Fig 5 (b)). Unlike the two-dimensional case, the frequency at which the estimation error is minimized does not coincide with the mode frequencies. The optimal frequency lies between 10 Hz and 40 Hz but does not correspond to their exact midpoint. The results of a similar simulation conducted for the dynamics shown in Fig 3 are shown in Fig 5 (c)(d). As depicted in Fig 5 (c), the matrix $\boldsymbol{A}$ has the same eigenvalues as in Fig 3, where it is represented at a sampling frequency of 1000 Hz for better comparability. Unlike the scenario in Fig 5 (a)(b), one of the modes retains oscillations even in the absence of perturbation. In this case, as illustrated in Fig 5 (d), the optimal input frequency is biased toward the frequency at which the mode exhibits strong damping (40 Hz). This phenomenon can also be explained using Eq. (58): the larger the real part of the eigenvalues in the continuous-time system, the greater the effect of resonance on increasing the objective function, which ultimately reduces parameter estimation errors. Thus, when designing frequency inputs, it is necessary to consider not only oscillations but also damping.

These simulations demonstrated that the optimal perturbation frequency corresponds to the intrinsic dynamics represented by matrix $\boldsymbol{A}$. We discuss the optimal input location further in the next section.

3.5.2 Optimal Location of Impulse Inputs

This section demonstrates the relationship between the optimal location of perturbation input and the network connectivity inherent in the system dynamics. Placing the perturbation input on a node that has high-weight connections with other nodes effectively minimizes $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1})$, and leads to an improvement in system identification, in accordance with Eq. (42).

To examine this relationship in practice, we constructed a network as shown in Fig 6 (a)(b). In this network, matrix $\boldsymbol{A}$ represents the adjacency matrix of the network, which coincides with system matrix $\boldsymbol{A}$ of the underlying LTI system. Matrix $\boldsymbol{B}$ was set to an identity matrix. The network consists of six nodes, which are numbered from 1 to 6, each

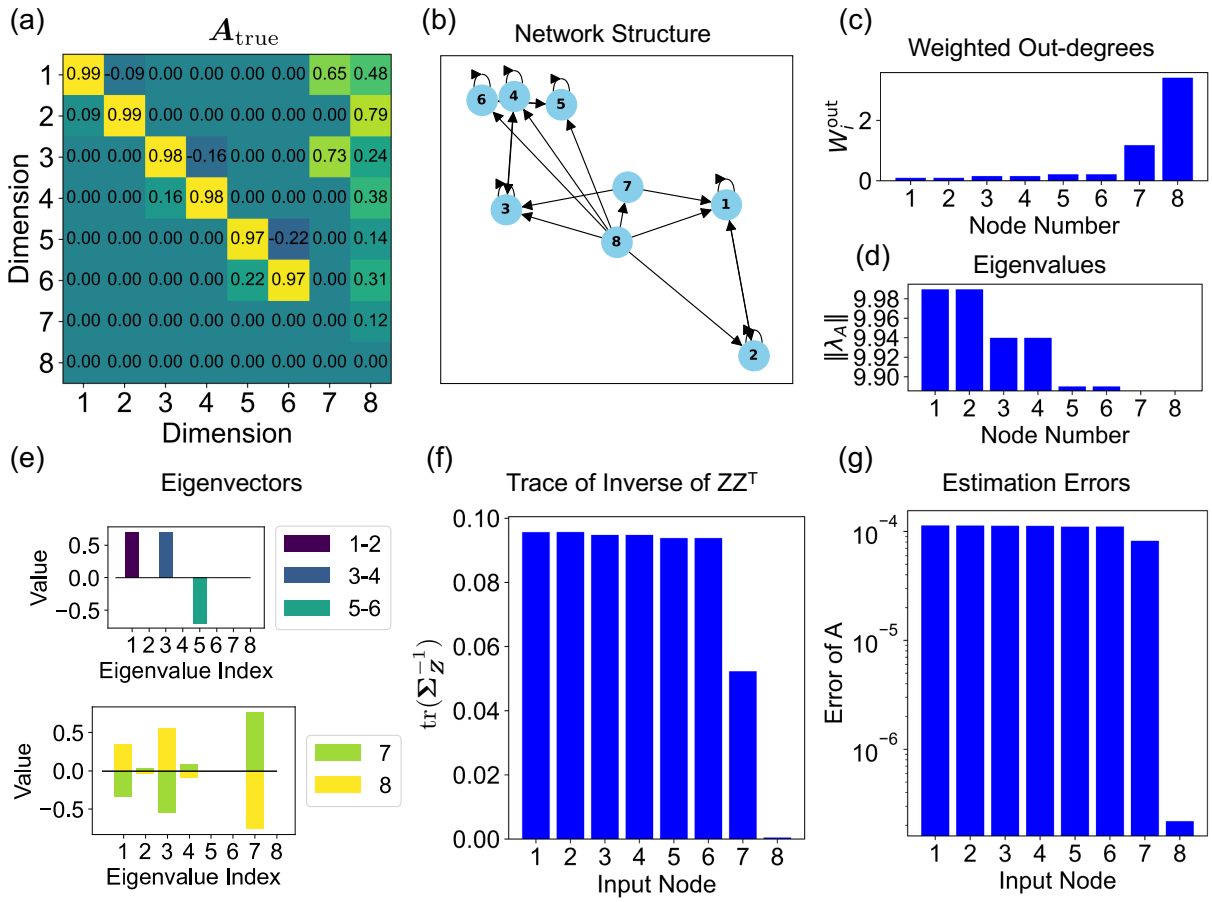

Fig 6: Perturbation input locations and model estimation errors. (a) The network was constructed with an 8×8 adjacency matrix $\boldsymbol{A}$ and designed to have three different modes. Nodes 7 and 8 have outgoing edges only, with node 8 having a higher number of connections. (b) Graph of the network structure. (c) Weighted-out degrees for each node. (d) Absolute eigenvalues of matrix $\boldsymbol{A}$. (e) Eigenvectors of matrix $\boldsymbol{A}$. (f) Trace of the inverse of the covariance matrix of $\boldsymbol{Z}$ when impulse input was applied to each node. (g) Estimation errors of the adjacency matrix $\boldsymbol{A}$ when the impulse input was applied individually to each node.

fulfilling specific roles in the network dynamics, and is structured with three modes. Nodes 7 and 8 have outgoing edges only, with no incoming connections from other nodes. This structural distinction between Nodes 7 and 8 plays a critical role in the overall connectivity and dynamics of the network. To quantify which node influences other nodes, metrics designed for weighted networks were used [54–56]. The weighted out-degree [57] of node i, excluding self-loops, is given by:

$$w_i^{\text{out}} = \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}, \tag{64}$$

where the element a_{ij} represents the weight of the directed edge from node i to node j, and self-loops are excluded by setting $a_{ii} = 0$ for all nodes i. This metric represents the total weighted influence or outgoing connections from node i to all other nodes in the network. It accounts for the sum of the weights of all directed edges originating from node i, excluding any self-loops. As shown in Fig 6 (c), Node 8 has a high weighted-out degree, followed by Node 7. To clearly differentiate the modes, the real parts of the eigenvalues for each mode were set to -1, -6, and -11, respectively. The frequencies were set to 15, 25, and 35 Hz. As shown in Fig 6 (d), differences in absolute eigenvalues can be created. The eigenvectors shown in Fig 6 (e) indicate that eigenvectors 1 through 6 correspond to the directions of each mode, while eigenvectors 7 and 8 are associated with directions related to those nodes that do not have modes but hold interactions with other nodes.

The simulation was conducted with a single node receiving a perturbation input. The input was implemented as an impulse signal with a strength of 10, and the perturbed signal was sampled at a frequency of 1000 Hz. Measurement noise for each node was generated from a Gaussian distribution with a mean of 0 and a standard deviation of 0.01. Data acquisition was performed over a duration of $T = 1$ second, and 300 trials were executed to ensure statistical robustness.

The results, shown in Fig 6 (f)(g), indicate that $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{Z}}^{-1})$ and the estimation error of adjacency matrix $\boldsymbol{A}$ decrease as the out-weighted degree w_i^{out} of the selected input node increases and the absolute eigenvalue of the selected input node decreases. This trend can be attributed to the structural properties of the network: nodes with higher out-weighted degrees possess more outgoing edges with significant weights, allowing perturbation inputs to propagate more effectively throughout the network. This propagation induces larger signal variances across the entire system, which is critical for accurate parameter estimation. The difference in the absolute values of eigenvalues also influences the estimation error. Specifically, smaller absolute eigenvalues lead to faster attenuation, resulting in smaller oscillation amplitudes in the dynamics before the introduction of perturbations. Consequently, when an impulse is applied to a part of the network corresponding to these smaller eigenvalues, the estimation error tends to be reduced.

This simulation demonstrated that applying a perturbation input to the node with the highest weighted out-degree or to one associated with low absolute eigenvalues, as calculated from the intrinsic dynamical model, constitutes an optimal perturbation input design. However, the intrinsic dynamical model cannot be directly observed in actual experiments. In the next section, we outline a method to address this issue.

4 Practical Identification Procedure of Neuroscience Systems

In this section, we outline an approach for determining the optimal perturbation and conducting system identification. The previous sections demonstrated the relationship between the latent neural model and the perturbation input. However, the latent model cannot be known from the outset, and indeed, if it were known, there would be no need for identification. To overcome this challenge, we outline an iterative approach alternating between system identification and input design which gradually minimizes the error of the parameter estimation. We demonstrate in simulations that this iterative approach incrementally optimizes the perturbation input and improves system identification.

We conducted simulations assuming time-series data collected from 32 points over 30 seconds. Sampling frequency was set to $f_s = 128$ Hz, with a time step of $T_s = 1/f_s$, and 10 trials were performed. Noise for each node was assumed to have a mean of 0 and a standard deviation of 10^{-20} for each node. When generating signal $\boldsymbol{x}(t)$, the true matrices $\mathbf{A}$ and $\mathbf{B}$, along with the perturbation inputs, were used. The perturbation input strength was set to 10^5. Matrix $\boldsymbol{A}$ was designed to have 9 modes (2, 3, 4, 11, 16, 18, 26, 33, and 38 Hz) (Fig 7 (a)). Matrix $\boldsymbol{B}$ was set as the identity matrix.

To estimate the matrix from the time-series data, we designed an optimal perturbation input (assuming frequency stimulation, such as tACS) using the following approach:

1. Matrix $\hat{\boldsymbol{A}}_{\text{passive}}$ is initially estimated in a passive state without perturbation input to provide a baseline estimation.
2. An initial perturbation $\boldsymbol{u}_0$, which has an arbitrary frequency and location, is introduced, and more accurate $\hat{\boldsymbol{A}}_0$ is estimated by the perturbation.
3. Using the estimated matrix, the optimal frequency and target node for the perturbation input are determined. The target node is identified based on the out-weighted degree of the nodes. The optimal frequency is identified by numerical calculation.
4. The designed perturbation $\boldsymbol{u}_1$ is applied to the system, and a more precise matrix $\hat{\boldsymbol{A}}_1$ can be identified.
5. The obtained matrix $\hat{\boldsymbol{A}}_1$ is again used to design a new perturbation input $\boldsymbol{u}_2$ and estimate a more accurate matrix $\hat{\boldsymbol{A}}_2$. This iterative process can be continued until the convergence criteria, such as changes in $\boldsymbol{A}$, are satisfied.

The estimation errors for matrix $\mathbf{A}$, measured using the Frobenius norm, were gradually decreased at each design step as shown in Fig 7 (b). We show the state variable plots and the actual estimated matrices in Fig 7 (c). First, the connectivity matrix estimated without perturbation was far from the true matrices. Next, the matrix estimated by initial perturbation was relatively close to the true matrices, but was still not sufficient. By designing the frequency and location of the perturbation input based on the estimated matrix $\boldsymbol{A}_{\text{initial}}$, the estimation error was further minimized. By repeating this procedure, the estimation error of matrix $\boldsymbol{A}$ was gradually decreased, as seen in $\boldsymbol{A}_{\text{design1}}$ and $\boldsymbol{A}_{\text{design2}}$. In this simulation, we demonstrated a framework to iteratively design the perturbation input and improve the accuracy of system identification using simulations. Use of this approach in the design of external stimuli in neuroscience experiments enhances design efficiency and promotes a deeper understanding of neural dynamics.

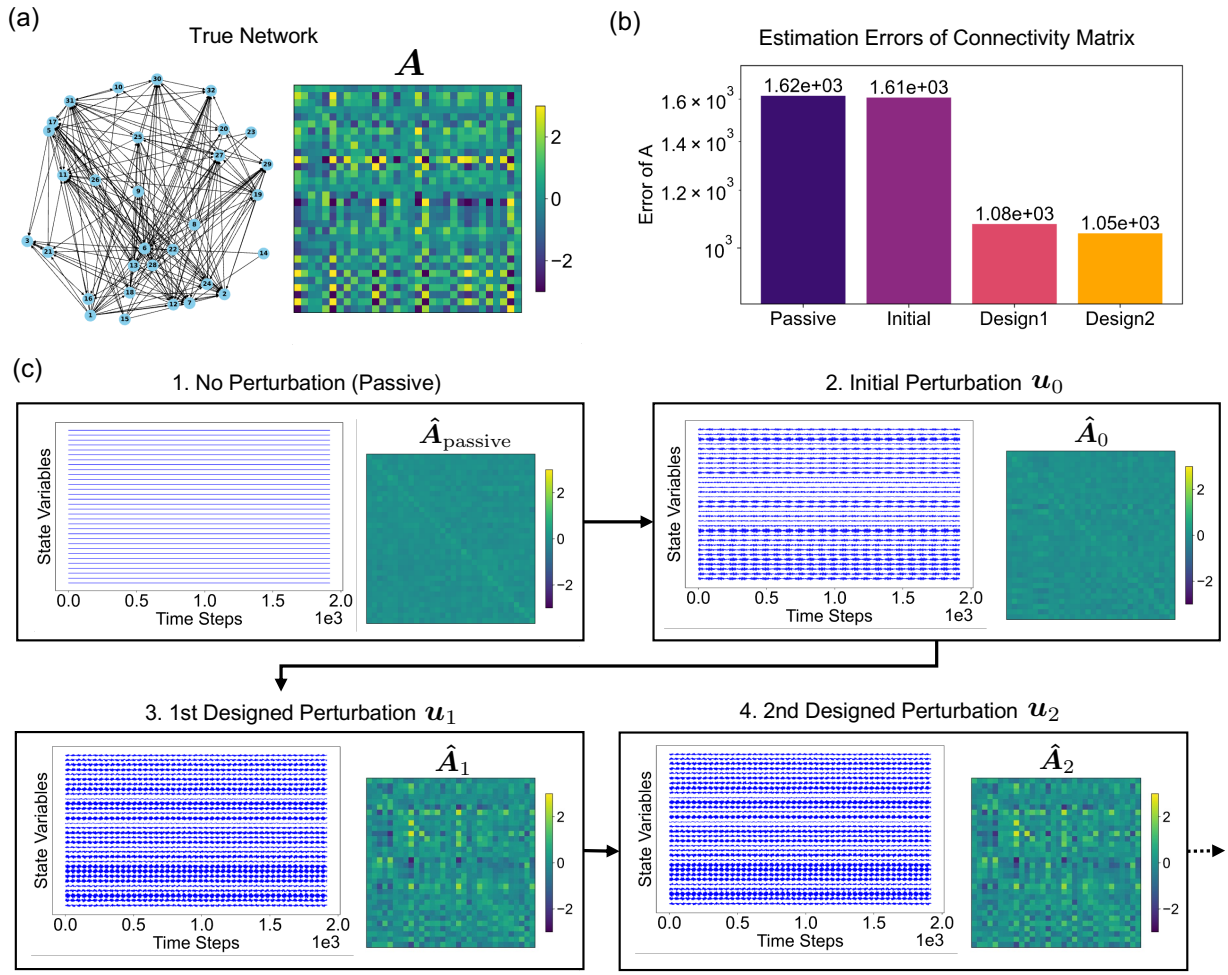

Fig 7: Simulation design results assuming neural activity data of 32 channels. (a) True matrix $\boldsymbol{A}$ and its network structure, which was designed to include both oscillatory and attenuating components. Only strong connections between nodes are shown in the structure (absolute weight > 1.5). (b) Comparison of the estimation errors of matrix $\boldsymbol{A}$ under four different conditions (passive, initial, 1st design, and 2nd design). (c) State variables and estimated matrix $\boldsymbol{A}$ following the iterative approach.

5 Accurate System Identification Enables Accurate Estimate of Optimal Control

Thus far, our discussion has centered on parameter identification of neural systems. In this final section, we explore a different yet related perspective: how these estimated parameters affect the optimal control theory in neuroscience [38,39]. Here, we estimate the system parameters by both passive and perturbation states. Then, we determine optimal control inputs which control the neural states to desired targets, and associated control costs. This simulation underscores the importance of accurate system identification for achieving optimal control and reliable estimation of neural functions. Throughout this section, we use the term *perturbation input* to refer to the input used for system identification and *control input* to refer to the input used for state transitions.

We designed a network as shown in Fig 8 (a) to clearly show the differences in system identification between passive and perturbations states. The network consists of two disconnected groups. Nodes 1 and 2, as well as nodes 3, 4, and 5, are each connected separately, with no connections between these groups, as shown in matrix $\boldsymbol{A}$. We assumed an input matrix $\boldsymbol{B}$ in which nodes 3 and 4 cannot be directly controlled. In this scenario, the optimal control strategy to manipulate nodes 3 and 4 is to apply inputs to node 5. However, if the system identification is inaccurate, there is the possibility that control will be attempted through nodes 1 and 2, which are disconnected from nodes 3 and 4. These matrices were estimated from the time series signals generated by simulation. Sampling frequency was set to $f_s = 1000$ Hz, and the time step was $T_s = 1/f_s$. The signals were generated for 1 second ($T = 1$) and the estimation

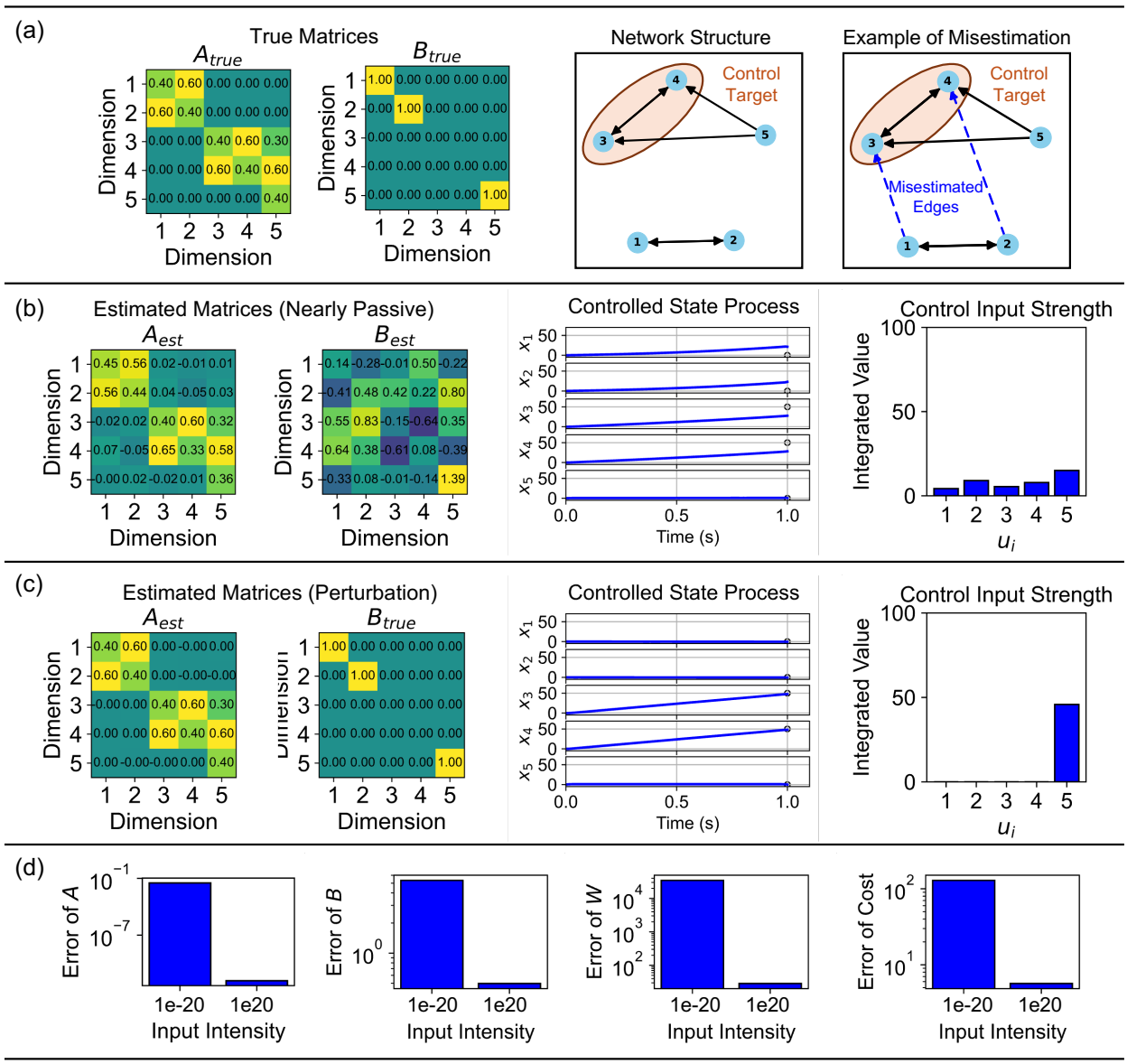

Fig 8: Optimal control difference between a passive-like model and a perturbation-based model. (a) True matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and network structure of the connectivity matrix. Nodes 1, 2, and 5 are controllable nodes. Nodes 3 and 4 are targets controlled by the determined control input. When the connectivity matrix is misestimated, the structure has edges between nodes 1 and 2, and 3 and 4. (b) Failed matrix estimations and controlled state transitions with near-zero impulse inputs ($\alpha = 10^{-20}$). The control objectives are plotted as white circles. (c) Successful matrix estimations and controlled state transitions with sufficiently strong impulse inputs ($\alpha = 10^{20}$). (d) Errors in matrices $\boldsymbol{A}$, $\boldsymbol{B}$, and $\boldsymbol{W}$ and control cost (average controllability)

was repeated over 300 trials. The noise for each node was generated following a normal distribution with a mean of 0 and a standard deviation of 0.01. We generated passive and perturbation states to estimate the parameter matrices. The passive condition uses a minimal intensity impulse input ($\alpha = 10^{-20}$) that was as close to passive as possible, since B estimation was impossible if the intensity was completely zero, and the other was an impulse input with sufficient intensity ($\alpha = 10^{20}$). By using these estimated matrices, we determined optimal control inputs, then evaluated the state transitions and associated costs. The controlled transition test was run with $T = 1$. The control objective was to set nodes 3 and 4 to 50 while keeping all other nodes at 0 without any movement.

If the estimation of $\boldsymbol{A}$ and $\boldsymbol{B}$ is inaccurate, the state transitions under optimal control cannot be executed correctly (Fig 8 (b)). Furthermore, the intensity of the control inputs tends to be distributed approximately equally across all

nodes. On the other hand, when $\boldsymbol{A}$ and $\boldsymbol{B}$ are estimated using a high-strength impulse, the resulting control inputs enable accurate state transitions (Fig 8 (c)). In this case, the strength of the impulse is correctly concentrated on node 5. As the errors in matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ increase, the estimation errors of the controllability Gramian lead to inaccuracies in estimating the control cost (Fig 8 (d)). Here, we adopted average controllability as an example measure of control cost (see also the Supplementary Material). The key point here is that, since the controllability Gramian involves multiple products of $\boldsymbol{A}$ (Eq. (30)), even small errors of $\boldsymbol{A}$ can result in significant discrepancies. Therefore, when estimating optimal control input for neural transitions or control costs, it is more appropriate to identify those model parameters with at least arbitrary perturbations, and ideally with designed perturbations.

6 Discussion

This study addressed the challenge of designing optimal perturbations for effectively identifying neural system dynamics. We introduced a framework for estimating the optimal perturbation input for identification of neural systems. Our findings demonstrated that incorporating perturbation inputs, including TMS, tDCS, and tACS, significantly improves identification accuracy. Specifically, alignment of their parameters with the intrinsic frequencies of matrix $\boldsymbol{A}$, high input intensity, and targeted inputs directed at nodes with high weighted out-degrees enhances system identification. Furthermore, we outlined an approach for designing the optimal perturbation input by iterating system identifications with perturbations, and demonstrated that the approach progressively decreases parameter estimation error. Beyond the parameter identification of neural systems, our findings provide insights into how these estimated parameters influence optimal control theory in neuroscience. These results emphasize the potential of perturbation input design to advance understanding of neural dynamics.

The assumption of linearity and first-order autoregressiveness in neural dynamics warrants discussion. Biological signals often exhibit complex and nonlinear interactions that cannot be fully captured by linear models. Nevertheless, prior studies have demonstrated that nonlinear behaviors can frequently be approximated using linear models [32, 38, 58]. Linear models are particularly effective for providing accurate approximations of nonlinear systems within a specific operating range [59]. Therefore, establishing theoretical foundations based on linear assumptions can contribute to the development of theories for nonlinear systems or be effectively utilized in their advancement. To extend these theories into the nonlinear domain, it would be necessary to incorporate advanced frameworks such as the Koopman operator [60, 61] and system identification methods leveraging machine learning techniques [62, 63]. On the other hand, discussions regarding higher-order VAR models are relatively straightforward due to the simplicity of the estimation method. Previous studies involving actual functional data have employed VAR models of an order greater than one [64–66]. Nonetheless, these studies utilized the same parameter estimation approach introduced in this study, namely the least squares method. Therefore, the framework proposed in this paper can be directly extended to higher-order VAR models.

Safety and experimental constraints play a pivotal role in the practical implementation of this framework, influencing both the scope and design of potential applications. In biological systems, stimulation parameters, such as amplitude, frequency, and location, must adhere to stringent safety thresholds to avoid adverse effects, such as tissue damage or unintended physiological responses [13, 67]. For stimulus intensity, our proposed framework suggested that a stronger stimulus improves system identification; however, an excessively strong stimulus poses safety risks [68]. With respect to frequency input, stimuli are typically applied within ranges corresponding to neural activity. The optimal input frequencies derived from the proposed theory are expected to align with these ranges, suggesting that the proposed framework can be implemented without major complications [68, 69]. Furthermore, the selection of stimulation sites is often dictated by experimental accessibility or ethical considerations [70, 71]. Based on these constraints, it is necessary to determine the most appropriate stimulation parameters to achieve optimal system identification.

With regard to the future direction of this proposed approach, we consider that it should be subject to validation experiments and stimulus experiments to elucidate neural dynamics. Validating a theoretical framework through experimental design is an essential next step in bridging the gap between theory and practice. Recent research involving some of the present authors has demonstrated that TMS can facilitate the discrimination of neural states [51]. Future experiments can build on this finding by incorporating passive, arbitrary, and designed optimal stimuli to estimate the connectivity matrix, and then comparing the ability of these matrices to distinguish different neural states. If these proposed evaluations demonstrate the practical effectiveness of our theory, it could then be applied to actual experiments. As described in Section 4, a preliminary connectivity matrix is obtained through preliminary experiments to design optimal perturbations. These perturbations would then be applied in the main experiments, improving the estimation of connectivity matrices and neural dynamics. Experiments using our approach will enable the more precise and comprehensive analysis of brain and neural function, and in turn facilitate valuable new insights into human cognition and behavior.

7 Acknowledgments

We are grateful to Yumi Shikauchi, Shunsuke Kamiya, and Daiki Kiyooka for their insightful feedback and valuable discussions, which significantly contributed to the development of this research. This work was supported by JST Moonshot R&D Grant Number JPMJMS2012 and JSPS KAKENHI Grant Number 24K20462.

References

[1] Waites, A. B., Briellmann, R. S., Saling, M. M., Abbott, D. F. & Jackson, G. D. Functional connectivity networks are disrupted in left temporal lobe epilepsy. *Ann. Neurol.* **59**, 335–343 (2006).

[2] Friston, K. J. Functional and effective connectivity: a review. *Brain Connect.* **1**, 13–36 (2011).

[3] Boly, M. *et al.* Brain connectivity in disorders of consciousness. *Brain Connect.* **2**, 1–10 (2012).

[4] Sporns, O. Contributions and challenges for network models in cognitive neuroscience. *Nat. Neurosci.* **17**, 652–660 (2014).

[5] Petersen, S. E. & Sporns, O. Brain networks and cognitive architectures. *Neuron* **88**, 207–219 (2015).

[6] Sporns, O. *Networks of the brain*. The MIT Press (MIT Press, London, England, 2016).

[7] Seguin, C. *et al.* Communication dynamics in the human connectome shape the cortex-wide propagation of direct electrical stimulation. *Neuron* **111**, 1391–1401.e5 (2023).

[8] Siegle, J. *et al.* Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592**, 86–92 (2021).

[9] Yan, Y. & Murphy, T. H. Decoding state-dependent cortical-cerebellar cellular functional connectivity in the mouse brain. *Cell Rep.* **43**, 114348 (2024).

[10] Park, A. H. *et al.* Optogenetic mapping of functional connectivity in freely moving mice via insertable wrapping electrode array beneath the skull. *ACS Nano* **10**, 2791–2802 (2016).

[11] Kucyi, A. *et al.* Intracranial electrophysiology reveals reproducible intrinsic functional connectivity within human brain networks. *J. Neurosci.* **38**, 4230–4242 (2018).

[12] Greicius, M. D. *et al.* Resting-state functional connectivity in major depression: abnormally increased contributions from subgenual cingulate cortex and thalamus. *Biol. Psychiatry* **62**, 429–437 (2007).

[13] Rocchi, F. *et al.* Increased fMRI connectivity upon chemogenetic inhibition of the mouse prefrontal cortex. *Nat. Commun.* **13**, 1056 (2022).

[14] Nentwich, M. *et al.* Functional connectivity of EEG is subject-specific, associated with phenotype, and different from fMRI. *Neuroimage* **218**, 117001 (2020).

[15] Bogéa Ribeiro, L. & da Silva Filho, M. Systematic review on EEG analysis to diagnose and treat autism by evaluating functional connectivity and spectral power. *Neuropsychiatr. Dis. Treat.* **19**, 415–424 (2023).

[16] Reid, A. T. *et al.* Advancing functional connectivity research from association to causation. *Nat. Neurosci.* **22**, 1751–1760 (2019).

[17] Bassett, D. & Sporns, O. Network neuroscience. *Nat. Neurosci.* **20**, 353–364 (2017).

[18] Siddiqi, S. H., Kording, K. P., Parvizi, J. & Fox, M. D. Causal mapping of human brain function. *Nature reviews neuroscience* **23**, 361–375 (2022).

[19] Stevens, M. C. The developmental cognitive neuroscience of functional connectivity. *Brain Cogn.* **70**, 1–12 (2009).

[20] Bergmann, T. O. & Hartwigsen, G. Inferring causality from noninvasive brain stimulation in cognitive neuroscience. *J. Cogn. Neurosci.* **33**, 195–225 (2021).

[21] Ruiz, S., Buyukturkoglu, K., Rana, M., Birbaumer, N. & Sitaram, R. Real-time fMRI brain computer interfaces: self-regulation of single brain regions to networks. *Biol. Psychol.* **95**, 4–20 (2014).

[22] Lee, M., Yoon, J.-G. & Lee, S.-W. Predicting motor imagery performance from resting-state EEG using dynamic causal modeling. *Front. Hum. Neurosci.* **14**, 321 (2020).

[23] Stepaniants, G., Brunton, B. W. & Kutz, J. N. Inferring causal networks of dynamical systems through transient dynamics and perturbation. *Phys. Rev. E.* **102**, 042309 (2020).

[24] Steinberg, E. E. & Janak, P. H. Establishing causality for dopamine in neural function and behavior with optogenetics. *Brain Res.* **1511**, 46–64 (2013).

[25] Lepperød, M. E., Stöber, T., Hafting, T., Fyhn, M. & Kording, K. P. Inferring causal connectivity from pairwise recordings and optogenetics. *PLoS Comput. Biol.* **19**, e1011574 (2023).

[26] Hallett, M. *et al.* Contribution of transcranial magnetic stimulation to assessment of brain connectivity and networks. *Clin. Neurophysiol.* **128**, 2125–2139 (2017).

[27] Bergmann, T. O. *et al.* Concurrent TMS-fMRI for causal network perturbation and proof of target engagement. *Neuroimage* **237**, 118093 (2021).

[28] Tik, M. *et al.* Acute TMS/fMRI response explains offline TMS network effects - an interleaved TMS-fMRI study. *Neuroimage* **267**, 119833 (2023).

[29] Tarnowski, W. Transient amplification in balanced neural networks. *arXiv [cond-mat.dis-nn]* (2020).

[30] Casali, A. G. *et al.* A theoretically based index of consciousness independent of sensory processing and behavior. *Sci. Transl. Med.* **5** (2013).

[31] Rotter, S. & Diesmann, M. Exact digital simulation of time-invariant linear systems with applications to neuronal modeling. *Biol. Cybern.* **81**, 381–402 (1999).

[32] Fernández Galán, R. On how network architecture determines the dominant patterns of spontaneous neural activity. *PLoS One* **3**, e2148 (2008).

[33] Zhao, Y. & Cortes, J. Gramian-based reachability metrics for bilinear networks. *IEEE Trans. Control Netw. Syst.* **4**, 620–631 (2017).

[34] Hamilton, J. D. *Time Series Analysis* (Princeton University Press, Princeton, 1994).

[35] Lutkepohl, H. *Introduction to multiple time series analysis* (Springer, Berlin, Germany, 1991).

[36] Montgomery, D. C. & Montgomery. *Design and Analysis of Experiments, 8th Edition* (John Wiley & Sons, 2012), 8 edn.

[37] Ucinski, D. *Optimal measurement methods for distributed parameter system identification* (CRC Press, London, England, 2019).

[38] Gu, S. *et al.* Controllability of structural brain networks. *Nat. Commun.* **6**, 8414 (2015).

[39] Karrer, T. M. *et al.* A practical guide to methodological considerations in the controllability of structural brain networks. *J. Neural Eng.* **17**, 026031 (2020).

[40] Kawakita, G., Kamiya, S., Sasai, S., Kitazono, J. & Oizumi, M. Quantifying brain state transition cost via schrödinger bridge. *Netw. Neurosci.* **6**, 118–134 (2022).

[41] Kamiya, S., Kawakita, G., Sasai, S., Kitazono, J. & Oizumi, M. Optimal control costs of brain state transitions in linear stochastic systems. *J. Neurosci.* **43**, 270–281 (2023).

[42] Kim, J. Z. *et al.* Role of graph architecture in controlling dynamical networks with applications to neural systems. *Nat. Phys.* **14**, 91–98 (2018).

[43] Braun, U. *et al.* Brain network dynamics during working memory are modulated by dopamine and diminished in schizophrenia. *Nat. Commun.* **12**, 3478 (2021).

[44] Ding, M., Chen, Y. & Bressler, S. L. *Handbook of time series analysis: Recent theoretical developments and applications* (Wiley-VCH Verlag, Weinheim, Germany, 2006), 1 edn.

[45] Rafajłowicz, E. Optimum choice of moving sensor trajectories for distributed-parameter system identification. *Int. J. Control* **43**, 1441–1451 (1986).

[46] Subasi, A. Selection of optimal AR spectral estimation method for EEG signals using cramer-rao bound. *Comput. Biol. Med.* **37**, 183–194 (2007).

[47] Rao, C. R. Information and the accuracy attainable in the estimation of statistical parameters. In *Springer Series in Statistics*, Springer series in statistics, 235–247 (Springer New York, New York, NY, 1992).

[48] Cavanaugh, J. E. & Shumway, R. H. On computing the expected fisher information matrix for state-space model parameters. *Stat. Probab. Lett.* **26**, 347–355 (1996).

[49] Tu, C. *et al.* Warnings and caveats in brain controllability. *Neuroimage* **176**, 83–91 (2018).

[50] Suweis, S. *et al.* Brain controllability: Not a slam dunk yet. *Neuroimage* **200**, 552–555 (2019).

[51] Shikauchi, Y. *et al.* Quantifying state-dependent control properties of brain dynamics from perturbation responses. *bioRxivorg* (2025).

[52] Fedorov, V. V. & Hackl, P. *Model-oriented design of experiments*. Lecture notes in statistics (Springer, New York, NY, 1997), 1997 edn.

[53] Walter, E. & Pronzato, L. *Identification of parametric models*. Communications and Control Engineering (Springer, London, England, 2010).

[54] Rubinov, M. & Sporns, O. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage* **52**, 1059–1069 (2010).

[55] Zuo, X.-N. *et al.* Network centrality in the human functional connectome. *Cereb. Cortex* **22**, 1862–1875 (2012).

[56] Eijlers, A. J. C. *et al.* Increased default-mode network centrality in cognitively impaired multiple sclerosis patients. *Neurology* **88**, 952–960 (2017).

[57] Acemoglu, D. THE NETWORK ORIGINS OF AGGREGATE FLUCTUATIONS. *Econometrica* **80**, 1977–2016 (2012).

[58] Honey, C. J. *et al.* Predicting human resting-state functional connectivity from structural connectivity. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 2035–2040 (2009).

[59] Khalil, H. K. *Nonlinear Systems* (Pearson Education, Delhi, India, 2017).

[60] Koopman, B. O. Hamiltonian systems and transformation in hilbert space. *Proc. Natl. Acad. Sci. U. S. A.* **17**, 315–318 (1931).

[61] Chow, C., Dan, T., Styner, M. & Wu, G. Understanding brain dynamics through neural koopman operator with structure-function coupling: 27th international conference, marrakesh, morocco, october 6–10, 2024, proceedings, part II. In Linguraru, M. G. *et al.* (eds.) *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, vol. 15002 of *Lecture Notes in Computer Science*, 509–518 (Springer Nature Switzerland, Cham, 2024).

[62] Suk, H.-I., Wee, C.-Y., Lee, S.-W. & Shen, D. State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *Neuroimage* **129**, 292–307 (2016).

[63] Glaser, J. I., Benjamin, A. S., Farhoodi, R. & Kording, K. P. The roles of supervised machine learning in systems neuroscience. *Prog. Neurobiol.* **175**, 126–137 (2019).

[64] Tseng, S. Y., Chen, R. C., Chong, F. C. & Kuo, T. S. Evaluation of parametric methods in EEG signal analysis. *Med. Eng. Phys.* **17**, 71–78 (1995).

[65] Chang, J.-Y. *et al.* Multivariate autoregressive models with exogenous inputs for intracerebral responses to direct electrical stimulation of the human brain. *Front. Hum. Neurosci.* **6**, 317 (2012).

[66] Shakeel, A., Onojima, T., Tanaka, T. & Kitajo, K. Real-time implementation of EEG oscillatory phase-informed visual stimulation using a least mean square-based AR model. *J. Pers. Med.* **11**, 38 (2021).

[67] Bikson, M. *et al.* Safety of transcranial direct current stimulation: Evidence based update 2016. *Brain Stimul.* **9**, 641–661 (2016).

[68] Antal, A. *et al.* Low intensity transcranial electric stimulation: Safety, ethical, legal regulatory and application guidelines. *Clin. Neurophysiol.* **128**, 1774–1809 (2017).

[69] Mager, T. *et al.* High frequency neural spiking and auditory signaling by ultrafast red-shifted optogenetics. *Nat. Commun.* **9**, 1750 (2018).

[70] Rossi, S., Hallett, M., Rossini, P. M., Pascual-Leone, A. & Safety of TMS Consensus Group. Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clin. Neurophysiol.* **120**, 2008–2039 (2009).

[71] Heinrichs, J.-H. The promises and perils of non-invasive brain stimulation. *Int. J. Law Psychiatry* **35**, 121–129 (2012).