eLife

Reviewed Preprint
v1 • March 9, 2026
Not revised

For correspondence:
ogino-m@g.ecc.u-tokyo.ac.jp
c-oizumi@g.ecc.u-tokyo.ac.jp

Competing interests: No competing interests declared

Funding: See page 23

Reviewing editor: Peter Latham, University College London, United Kingdom

Designing optimal perturbation inputs for system identification in neuroscience

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

[1]Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan • [2]School of Data Science, Yokohama City University, Yokohama, Japan

eLife Assessment

The authors establish **solid** theoretical principles for designing brain perturbations under the assumption that brain activity evolves under a linear model. By prioritizing low-variance components, resonant frequencies, and hub nodes, this framework provides an **important** foundation for optimizing information gain, neural state classification, and the control of neural dynamics. However, the lack of investigation of model mismatch makes the study **incomplete**.

https://doi.org/10.7554/eLife.110030.1.sa1

Abstract

Investigating the dynamics of neural networks, which are governed by connectivity between neurons, is a fundamental challenge in neuroscience. Because passive (spontaneous) activity provides only limited information for estimating connectivity, perturbation-based approaches are widely applied in neuroscience, as they can evoke underlying hidden dynamics. However, the characteristics of such perturbations have typically been designed based on empirical or biological intuition. To enable more accurate estimation of connectivity, we propose a data-driven and theoretically grounded framework for optimally designing perturbation inputs, based on formulating the neural model as a control system. The core theoretical insight underlying our approach is that neural signals observed in the passive state lack sufficient latent information, which leads to failures in the system identification. Perturbations reveal these hidden dynamics and lead to improved estimation. Guided by these insights, we derive a theoretical basis for optimizing perturbation inputs that minimize estimation errors in neural system identification. Building upon this, we further explore the relationship of this theory with stimulation patterns commonly used in neuroscience, such as frequency, impulse, and step inputs. We demonstrate the effectiveness of this framework for neuroscience through simulations grounded in experimental paradigms such as neural state classification and optimal control of neural states. Our theoretical analysis, together with multiple simulations, consistently shows that perturbations designed according to our framework achieve substantially more accurate system identification compared to the conventional, intuition-based inputs. This study provides a theoretical foundation for designing perturbation inputs to achieve accurate estimation of neural dynamics. This, in turn, enables reliable discrimination of neural states such as levels of consciousness and pathological conditions, and facilitates precise control of their transitions toward recovery from abnormal states.

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. In particular, a commonly studied aspect is functional connectivity, which captures the temporal dependencies between neural activities [8–11]. The functional connectivity is estimated through recording techniques such as neural spike recording techniques [12,13], electrocorticography (ECoG) [14,15], functional magnetic resonance imaging (fMRI) [16, 17] and electroencephalography (EEG) [18, 19]. While various methods have been proposed and used to estimate these connections (see for example [2, 20] for a comprehensive review), one typical method is based on dynamic modeling, among which is a simple but widely used linear autoregressive model (Fig. 1a) [21–25]. Connectivity is statistically estimated from the time-series data of neural activity (Fig. 1b) by fitting the model parameters (Fig. 1c). A neural connectivity matrix helps visualize the connections between different brain regions, and provides a detailed map of functional 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 [8, 26–30].

A fundamental problem with passive observation is that it fails to reveal hidden dynamics, which leads to an invalid estimate of the corresponding parts of the model. This limitation stems from the attenuation of latent dynamical modes, such as transient or damped components. When neural activity is modeled as a linear dynamical system, as shown in Fig. 1a, the true connectivity matrix governs the generation of time-series signals (Fig. 1b). Passive observation records these spontaneous neural activities within an arbitrarily predefined temporal window and attempts to estimate the connectivity matrix using parameter estimation techniques. However, the resulting matrix $\boldsymbol{A}_{\text{passive}}$ deviates significantly from the true model (Fig. 1c), because some modes quickly decay and vanish from the observable data (see also Fig. 2 for an intuitive explanation). Consequently, relying solely on passive recordings leads to misinterpretations of the neural system—such as overlooking fast transient dynamics or underestimating connectivity strength—which, in turn, may result in inaccurate conclusions in both basic neuroscience and clinical contexts.

To overcome these limitations, the application of perturbations has emerged as a powerful approach for improving the estimation of connectivity in neuroscience [29, 31–33]. Its effectiveness lies in its ability to actively manipulate the neural system, rather than merely observing it, and to thereby uncover dynamic and causal interactions that are otherwise hidden [32]. Recent studies using stimulation techniques such as optogenetics [34, 35] and transcranial magnetic stimulation (TMS) [29, 36, 37] have demonstrated that these interventions can significantly improve the detection of causal relationships within the neural system. Such approaches have been applied to the study of cognitive processes and states of consciousness [29, 31]. These studies underscore the value of perturbation inputs in driving state transitions in the neural system, thereby enabling more accurate and reliable connectivity estimation.

However, the parameters of perturbation protocols, such as the location, intensity, and shape of stimulation, have often been determined based on anatomical and physiological insights and on empirically established techniques [29,38–41]. While these approaches have provided practical utility, they may not optimally exploit the underlying neural dynamics or maximize the informativeness of the perturbation. To overcome these limitations and enhance the reliability of connectivity inference, there is a pressing need to design data-driven and theoretically grounded perturbations.

In this paper, we propose a framework for designing the optimal perturbation input through control theory in neuroscience. We interpret neural dynamics as a control system [42–47], and treat external perturbations as control inputs to design properties of neural stimulation (Fig. 1d). If the optimal perturbation input can be systematically designed, it becomes possible to steer the neural system toward states that are maximally informative (Fig. 1e), thereby enhancing the accuracy of the inferred connectivity (Fig. 1f). To our knowledge, this framework has not been investigated in the field of neuroscience. We first describe how to formulate neural dynamics as a control system and how to estimate the model parameters from observed data. Building upon this formulation, we derive a theoretical basis that enables us to design the optimal

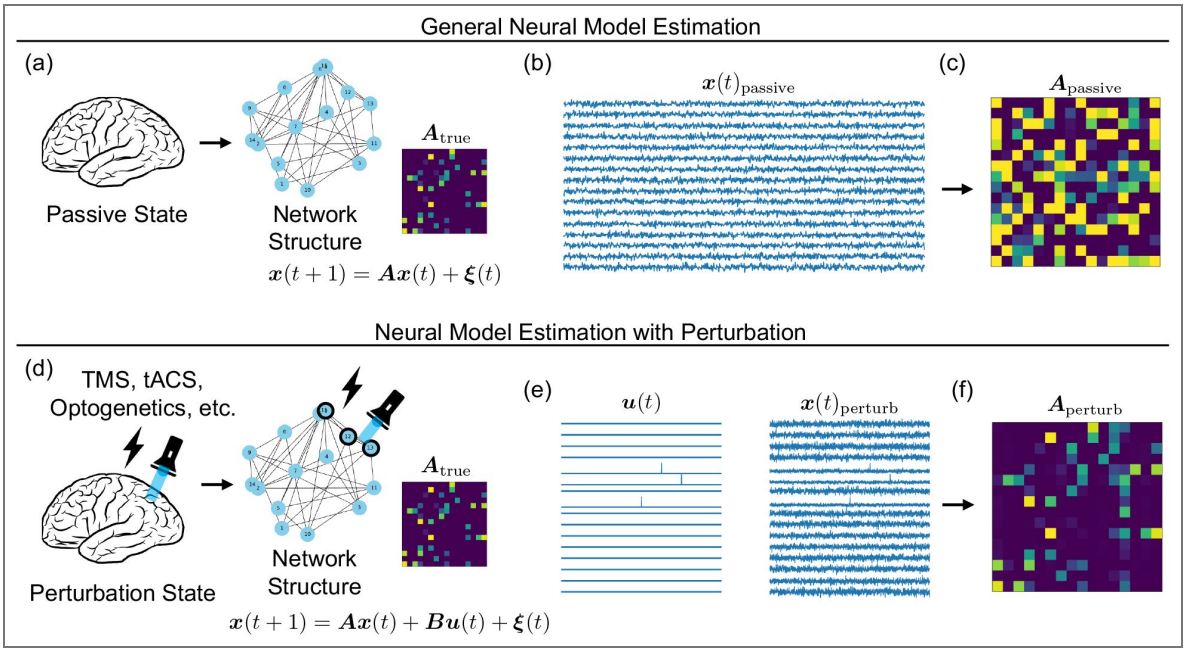


Fig. 1. Passive and perturbation states and the resulting neural dynamics models.

(a) Passive state of the neural network and the corresponding ground-truth model. (b) Recorded neural activities without perturbation. (c) Estimated model obtained without perturbation. (d) Perturbation state of the neural network and corresponding ground-truth model. (e) Recorded neural activity with perturbation. (f) Estimated model obtained with perturbation.

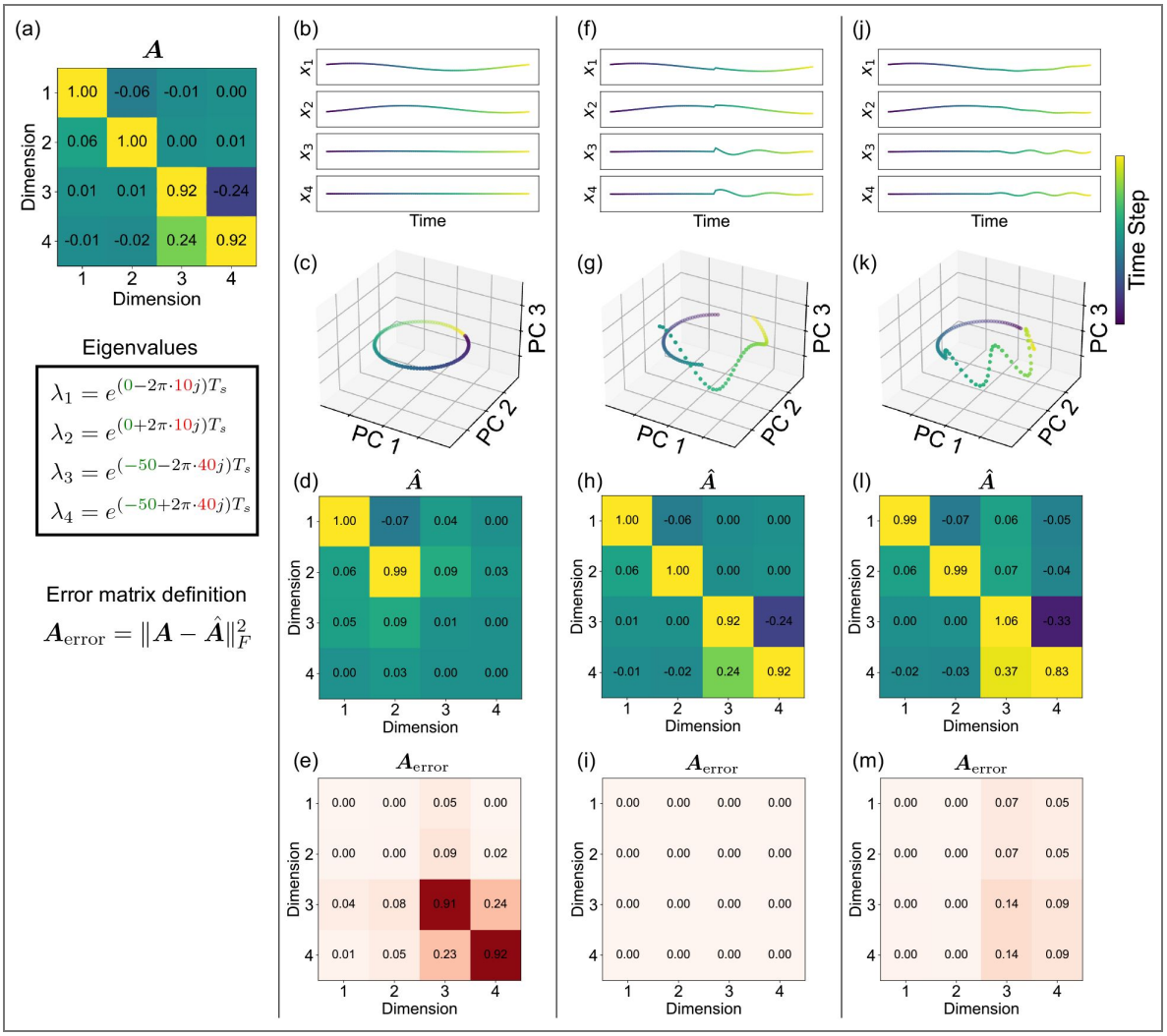


Fig. 2. Visualization of continuous-time system dynamics by perturbations.

(a) System matrix ***A*** representing the dynamics and its eigenvalues. (b) Temporal changes in the state variables of the passive state without perturbation. (c) Time-varying trajectories of each PCA component. The color gradient indicates temporal progression. (d) Estimated connectivity matrix in the passive condition. (e) Error matrix for the passive-state estimation. (f-i) Temporal changes and estimation results in response to an impulse input. (j-m) Temporal changes and estimation results in response to a sinusoidal input.

perturbation inputs for the neural system identification. We demonstrate the validity and utility of this theoretical basis by exploring its implications for optimizing parameters of common neurostimulation techniques (such as TMS, tDCS, and tACS) and by applying it to practical examples, including neural state classification [31, 36, 47, 48] and control of neural states [42, 44, 45, 49]. In these demonstrations, we define concrete problems and apply the theory to validate its practical utility.

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.

Background

To evaluate how external perturbations enhance the accuracy of system identification, this section reviews the estimation formulations for linear dynamical systems, comparing the cases with and without perturbation inputs.

System Identification with External Perturbation

This section reviews a well-established framework for system identification of linear dynamical systems with external perturbations. Researchers have modeled brain dynamics as a discrete-time linear dynamics with external inputs and stochastic system noise [42, 43, 50–52]. We assume that the brain state evolves according to the following linear dynamics:

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

where $\boldsymbol{x}(t) \in \mathbb{R}^n$ denotes the neural state vector at time t, such as the activity of neurons, populations, or brain regions, and n represents its dimension. The connectivity matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ characterizes intrinsic interactions reflecting functional connectivity. The input matrix $\boldsymbol{B} \in \mathbb{R}^{n\times m}$ specifies how external inputs $\boldsymbol{u}(t) \in \mathbb{R}^m$—including sensory stimuli or interventions such as TMS—affect the system, where m represents the number of perturbation channels. The input $\boldsymbol{u}(t)$ is freely designed and known to the experimenter. Finally, $\boldsymbol{\xi}(t) \in \mathbb{R}^n$ represents stochastic fluctuations (e.g., synaptic noise or unobserved inputs), modeled as Gaussian noise with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\Xi}} \in \mathbb{R}^{n\times n}$. We remark that neural activities measured at the macroscopic level, such as functional magnetic resonance imaging (fMRI) or intracranial electroencephalography (iEEG) have been experimentally and theoretically validated to follow approximately linear dynamical systems in Refs. [53, 54].

Here, we assume the input matrix $\boldsymbol{B}$ is known, and we are only estimating the connectivity matrix $\boldsymbol{A}$ from the time series data of $\boldsymbol{x}$ and $\boldsymbol{u}$. We construct the data matrices as

$$\boldsymbol{X} = [\boldsymbol{x}(0), \boldsymbol{x}(1), \ldots, \boldsymbol{x}(T-1)], \quad \boldsymbol{Y} = [\boldsymbol{x}(1), \ldots, \boldsymbol{x}(T)], \quad \boldsymbol{U} = [\boldsymbol{u}(0), \ldots, \boldsymbol{u}(T-1)], \tag{2}$$

where T represents the length of the time series data. Using these matrices, the parameter matrix $\boldsymbol{A}$ can be estimated by minimizing the reconstruction error in the sense of ordinary least squares (OLS):

$$\hat{\boldsymbol{A}} = \arg\min_{\boldsymbol{A}} \|\boldsymbol{Y} - (\boldsymbol{A}\boldsymbol{X} + \boldsymbol{B}\boldsymbol{U})\|_F^2 \tag{3}$$

$$= (\boldsymbol{Y} - \boldsymbol{B}\boldsymbol{U})\boldsymbol{X}^\dagger, \tag{4}$$

where $\|\cdot\|_F$ represents the Frobenius norm and the symbol † denotes the pseudoinverse of a matrix.

System Identification without External Perturbation

To compare the quality of system identification with and without perturbations, we also consider the passive case without external input. This corresponds to setting $\boldsymbol{u}(t) = \mathbf{0}$ in the formulation in the perturbed case, yielding the following dynamics:

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

Here, $\boldsymbol{A}$ and $\boldsymbol{\xi}(t)$ are identical to those defined in Eq. 1. Because this system evolves differently from the perturbed case, we denote its state trajectory as $\boldsymbol{x}_{\text{passive}}(t)$ to explicitly distinguish it from the dynamics under perturbation.

Similarly, the parameter matrix $\boldsymbol{A}$ in the passive condition can be estimated by applying the ordinary least squares (OLS) method:

$$\hat{\boldsymbol{A}}_{\text{passive}} = \boldsymbol{Y}_{\text{passive}}\boldsymbol{X}^{\dagger}_{\text{passive}}, \tag{6}$$

where

$$\boldsymbol{X}_{\text{passive}} = [\boldsymbol{x}_{\text{passive}}(0), \boldsymbol{x}_{\text{passive}}(1), \ldots, \boldsymbol{x}_{\text{passive}}(T-1)], \quad \boldsymbol{Y}_{\text{passive}} = [\boldsymbol{x}_{\text{passive}}(1), \ldots, \boldsymbol{x}_{\text{passive}}(T)]. \tag{7}$$

The subscript “passive” is used again to explicitly distinguish variables associated with the unperturbed dynamics from those obtained under external perturbations.

Results

We present a theoretical framework for determining the optimal perturbation inputs to enhance neural system identification. We begin by presenting an intuitive example in which the failure of passive estimation is demonstrated, and the benefit of perturbation inputs in re-exciting weakly observable dynamics becomes evident. Following this, we provide a guiding theoretical principle that estimation error of $\boldsymbol{A}$ is reduced by excitation of the state $\boldsymbol{x}$ by perturbation input $\boldsymbol{u}$. We then validate this theoretical foundation by applying it to canonical perturbation signals used in neuroscience—sinusoidal inputs and impulse as typically observed in tACS, tDCS, and TMS—and derive analytical relationships between input parameters and estimation errors. Furthermore, we demonstrate the applicability of the proposed framework to practical problems in neuroscience, such as neural state classification and optimal control for neural state transitions. Finally, we outline a practical framework for designing optimal perturbation inputs to estimate neural dynamics.

Typical Failure of Passive-State Model Estimation

We demonstrate why passive state observation, a common experimental condition in system neuroscience, fails to estimate the system model of neural dynamics. Although passive observation is widely employed due to its convenience and non-invasiveness, this practice has fundamental limitations: it inevitably overlooks causal and dynamical information that vanishes under spontaneous conditions but can be recovered through external perturbations, leading to degraded connectivity estimates. Such limitations lead to misinterpretations of the underlying neural functions when neural dynamics are modeled as a control system.

We estimate the connectivity matrix from simulated time series data generated by the true model. The true model, characterized by a connectivity matrix $\boldsymbol{A}$, is illustrated in Fig. 2a. Neural data are generated based on the connectivity matrix, yielding the time series shown in Fig. 2b. These data are not influenced by external input (passive state), and thus the oscillations of x_3 and x_4 gradually attenuate over time. This attenuation is further confirmed by principal component analysis (PCA), which reveals that the first and second principal components capture the 10Hz dynamics, while the third principal component, associated with the 40Hz mode, contributes negligibly to the total variance (Fig. 2c). Consequently, the matrix estimated by OLS deviates substantially from the true connectivity matrix (Fig. 2d), particularly showing errors in the lower-right components, as highlighted in Fig. 2e. This failure arises because the true matrix $\boldsymbol{A}$ contains two distinct dynamical modes: one is a continuous oscillatory mode at 10Hz, and the

other is a damped mode at 40Hz. The damped mode becomes unobservable in the time series once its contribution has attenuated and vanished. Passive recordings capture only superficial aspects of the connectivity and fail to estimate the true underlying neural dynamics.

The limitation of the passive state can be resolved by applying perturbation inputs. Figures 2f–2i show the application of an impulse input $\boldsymbol{u}$ to the system. This perturbation input primarily 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 estimate than in the passive case. Similarly, an improvement is observed when different types of input—a sinusoidal input for example—are used instead of the impulse input, as shown in Figs. 2j–2m. These results illustrate that appropriate perturbation inputs can effectively re-excite weak dynamics, thereby enhancing system identification accuracy. This can be theoretically supported by the concept of persistent excitation [55–58]. According to this theory, a perturbation input is said to be persistently exciting if it causes the system's state to sufficiently explore the state space over time. These insights underscore the importance of incorporating controlled stimulation in experimental design, particularly when accurate system identification is desired.

Intuitive Perturbation Design Informed by Covariance Eigenvalues

In this section, we show how perturbations reduce the estimation error of the system matrix $\boldsymbol{A}$ by exciting the state dynamics, thereby providing a theoretical basis for designing effective perturbation inputs. When the data length T is sufficiently large ($T \to \infty$), the estimation error of $\boldsymbol{A}$ asymptotically converges to the following expression [59, 60]:

$$\mathbb{E}\left[\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_F^2\right] \approx \frac{1}{T}\,\mathrm{tr}(\boldsymbol{\Sigma}_\Xi)\,\mathrm{tr}(\boldsymbol{\Sigma}_X^{-1}). \tag{8}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{X}} = \frac{1}{T}(\boldsymbol{X} - \bar{\boldsymbol{X}})(\boldsymbol{X} - \bar{\boldsymbol{X}})^\top$ with $\bar{\boldsymbol{X}} = \frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{x}(t)$ represents the covariance matrix of the state vector $\boldsymbol{x}$, and tr denotes the trace of a matrix. The detailed derivation of this equation is provided in the Supplementary Material A.1. This equation indicates that the estimation error of $\boldsymbol{A}$ can be reduced either by increasing the measurement duration T or by minimizing the trace of the inverse covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1}$, which depends on the perturbation input $\boldsymbol{u}(t)$. Intuitively, this relationship resembles a signal-to-noise ratio, where increasing the covariance of the state dynamics relative to the noise covariance leads to improved estimation accuracy.

The asymptotic expression in Eq. 8 reveals that the estimation error of $\boldsymbol{A}$ depends inversely on the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$. Expressing this relationship explicitly in terms of the eigenvalues yields

$$\mathbb{E}\left[\|\hat{\boldsymbol{A}} - \boldsymbol{A}\|_F^2\right] \approx \frac{1}{T}\,\mathrm{tr}(\boldsymbol{\Sigma}_\Xi)\sum_{i=1}^{n}\frac{1}{\mu_i}. \tag{9}$$

where μ_i denotes the i-th eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$. This relationship indicates that small eigenvalues dominate the estimation error through their inverse contributions. Therefore, the perturbation input should be designed not only to increase the overall variance of the state $\boldsymbol{x}$, but to enlarge the eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ more uniformly across all directions of the state space. In other words, exciting all dynamical modes of the system–rather than amplifying a limited subset–is essential for improving estimation accuracy.

This theoretical insight can be illustrated intuitively in Fig. 3. The figure visualizes ellipsoids constructed from the eigenvalues μ_i and their reciprocals $1/\mu_i$ of the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$. The covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ are calculated from the time series of the state vector $\boldsymbol{x}$ obtained under passive and perturbation conditions (Fig. 3a). The length of each axis of the ellipsoid corresponds to the variance of the state along that direction, which is associated with the eigenvalue μ_i. Its reciprocal counterpart $1/\mu_i$ represents the contribution to the estimation error (Fig. 3b). As shown in Fig. 3c, when the neural dynamics is dominated by a single eigenvalue μ_1, the other eigenvalues μ_2 and μ_3 remain small, resulting in an elongated reciprocal ellipsoid (Fig. 3d) and a large estimation error. When perturbation inputs that excite the directions associated with μ_2 and μ_3 are applied, the ellipsoid expands and becomes closer to a sphere (Fig.

3e). As a result, the estimation error decreases in all directions, as illustrated in Fig. 3f. This uniform enlargement of the eigenvalues provides a clear guideline: perturbations should be designed so that the variance becomes large in all directions, rather than being confined to specific modes.

Theoretical Formulation of Sinusoidal Inputs for an Overall Increase in Eigenvalues

Based on the guideline presented in the previous section, we derive analytical expressions to investigate which frequency of perturbation input will efficiently excite the dynamical modes of a system. In neuroscience, such a perturbation is analogous to tACS. Since neural dynamics possess modal frequencies, there should exist perturbation input frequencies that resonate with these modes. Such resonance leads to an amplification of $\boldsymbol{x}(t)$, which generally results in larger eigenvalues μ_i of the covariance matrix $\boldsymbol{\Sigma_X}$, reflecting increased variability along the corresponding modes. To clarify this relationship, we derive an analytical expression for $\boldsymbol{x}(t)$ to investigate how the input frequency affects its amplitude. The solution $\boldsymbol{x}(t)$ to the model in Eq.1 can be obtained by iterating the state transition matrix. Specifically,

$$\boldsymbol{x}(t) = \boldsymbol{A}^t \boldsymbol{x}(0) + \sum_{k=0}^{t-1} \boldsymbol{A}^{t-1-k} \boldsymbol{B}\boldsymbol{u}(k) + \sum_{k=0}^{t-1} \boldsymbol{A}^{t-1-k} \boldsymbol{\xi}(k) \tag{10}$$

We assume a cosine input $\boldsymbol{u}(k) = \sum_{l=1}^{L} \cos(\omega_l k)\boldsymbol{u}_0$, where ω_l are the angular frequencies. By substituting this into the second term, we can obtain the following equation by performing diagonalization of $\boldsymbol{A}$ and expressing the eigenvalues in polar form.

$$\boldsymbol{x}(t) = \boldsymbol{x}_{\text{passive}}(t) + \boldsymbol{x}_{\text{diff}}(t), \tag{11}$$

$$\boldsymbol{x}_{\text{passive}}(t) = \boldsymbol{A}^t \boldsymbol{x}(0) + \sum_{k=0}^{t-1} \boldsymbol{A}^{t-1-k} \boldsymbol{\xi}(k) \tag{12}$$

$$\boldsymbol{x}_{\text{diff}}(t) = \sum_{l=1}^{L} \sum_{d=1}^{n} \text{Re}\left[\frac{r_d^{t+1} e^{i\theta_d(t-1)} - r_d^t e^{i(\theta_d t - \omega_l)} - r_d e^{i(\omega_l t - \theta_d)} + e^{i\omega_l(t-1)}}{r_d^2 - 2r_d \cos(\theta_d - \omega_l) + 1} \boldsymbol{v}_d \boldsymbol{w}_d^\top \boldsymbol{B}\, \boldsymbol{u}_0 \right] \tag{13}$$

where each eigenvalue of $\boldsymbol{A}$ is written in polar form as $\lambda_d = r_d e^{i\theta_d}$, with $r_d = |\lambda_d|$ denoting the magnitude and $\theta_d = \arg(\lambda_d)$ denoting the argument. $\boldsymbol{x}_{\text{passive}}$ represents the state $\boldsymbol{x}$ that would be realized if no perturbation were applied, i.e., the state that would evolve according to Eq. 5. The vector $\boldsymbol{v}_d$ denotes the right eigenvector of $\boldsymbol{A}$ associated with λ_d, and $\boldsymbol{w}_d^\top$ denotes the corresponding left eigenvector. The detailed derivation is provided in the Supplementary Material A.2.1.

We focus on $\boldsymbol{x}_{\text{diff}}(t)$ simply because it is the component directly driven by the control input. When the input frequencies ω_l coincide with the mode's angular component θ_d, a resonance-like amplification occurs in $\boldsymbol{x}_{\text{diff}}(t)$—that is, the amplitude of $\boldsymbol{x}(t)$ increases markedly. Increasing the magnitude of $\boldsymbol{x}(t)$ through resonance naturally leads to an increase in the overall variance in $\boldsymbol{\Sigma_X}$. This enhancement directly contributes to increasing all eigenvalues μ_i of the covariance matrix rather than leaving some unexcited. Therefore, the analysis of Eq. 13 provides the necessary guidelines to target and amplify specific dynamical modes, suggesting that oscillatory inputs, such as tACS, should be designed with frequencies that correspond to the true dynamical mode frequencies.

Theoretical Formulation of Impulse and Step Inputs

The eigenvalues of the covariance matrix can also be increased by manipulating the intensity of the perturbation input. In neuroscience, external perturbations such as TMS and tDCS can be modeled as impulse-like or step-like inputs to neural systems. The covariance matrices of $\boldsymbol{x}(t)$ for impulse inputs can be written as:

$$\boldsymbol{x}_{\text{passive}}(t) = \boldsymbol{A}^t \boldsymbol{x}(0) + \sum_{k=0}^{t-1} \boldsymbol{A}^{t-1-k} \boldsymbol{\xi}(k), \quad \boldsymbol{x}_{\text{diff}}(t) = \alpha\, \boldsymbol{A}^{t-1} \boldsymbol{B}\boldsymbol{u}_0. \tag{14}$$

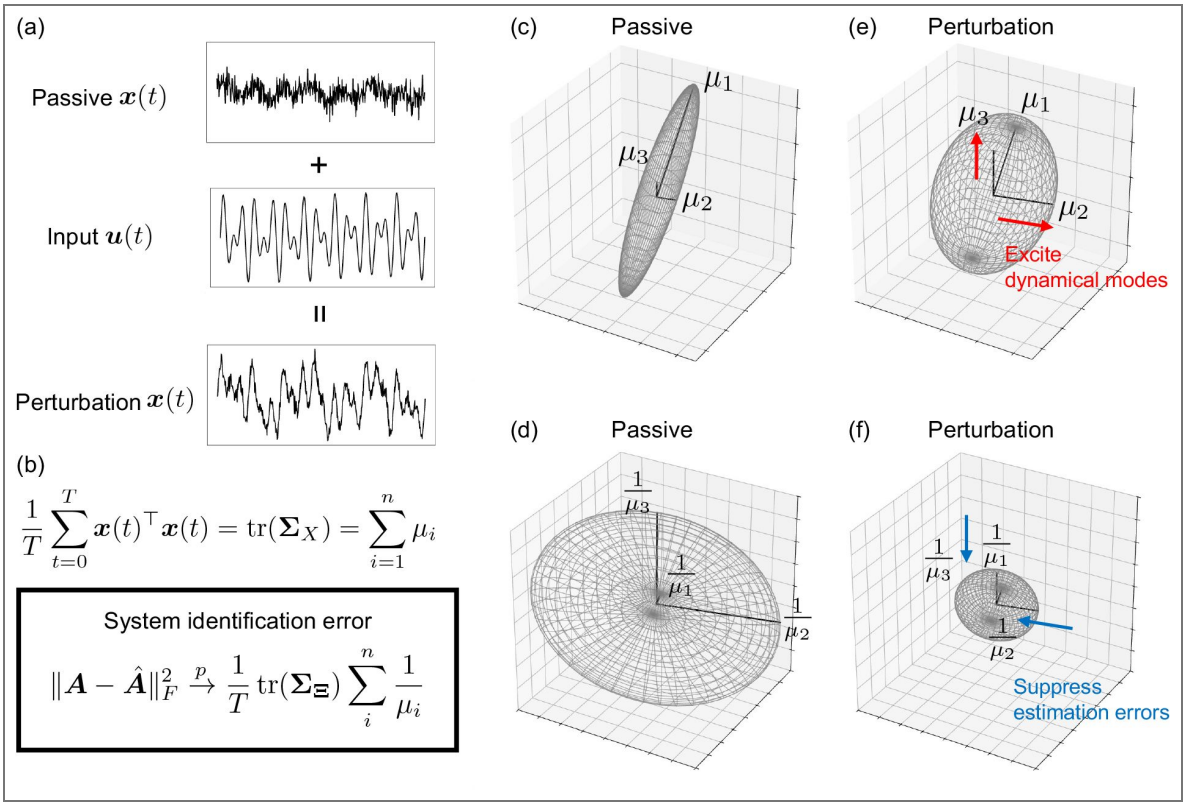


Fig. 3. Effect of perturbation on the eigenvalues of the state covariance matrix.

Ellipsoidal plots show the eigenvalues and reciprocal eigenvalues of the covariance matrix of the state trajectories ***X***, comparing passive dynamics and the case with perturbation. (a) Examples of a passive signal, a perturbation input, and a perturbed signal. (b) Analytical relationship linking the eigenvalues of the state covariance matrix to system identification error. (c) Passive: The covariance matrix is dominated by a single eigenvalue direction. (d) Passive (reciprocal): A single dominant eigenvalue produces a widely spread reciprocal-space ellipsoid (e) Perturbation: Additional dynamical modes are excited, increasing the smaller eigenvalues. (f) Perturbation (reciprocal): Perturbations yield a more uniform reciprocal-space ellipsoid, reducing estimation error.

For step input, the state vector is described by

$$\boldsymbol{x}_{\text{passive}}(t) = \boldsymbol{A}^t \boldsymbol{x}(0) + \sum_{k=0}^{t-1} \boldsymbol{A}^{t-1-k} \boldsymbol{\xi}(k), \quad \boldsymbol{x}_{\text{diff}}(t) = \beta \sum_{k=0}^{t-1} \boldsymbol{A}^{t-1-k} \boldsymbol{B} \boldsymbol{u}_0. \tag{15}$$

where α and β are intensity of the impulse and step inputs. The detailed derivation is provided in the Supplementary Material A.2.2. From these equations, it is clear that the contribution of the perturbation to the state covariance increases proportionally to the squares of the input strengths, α^2 and β^2. However, estimation accuracy is not determined by input intensity alone. Eqs.14 and 15 show that the resulting dynamics $\boldsymbol{x}_{\text{diff}}(t)$ are critically dependent on the term $\boldsymbol{B}\boldsymbol{u}_0$. This term represents how the input vector $\boldsymbol{u}_0$ (which defines the spatial pattern of the stimulation, i.e., which nodes are targeted) interacts with the system's input matrix $\boldsymbol{B}$. Therefore, while increasing intensity (α, β) within experimental constraints is beneficial, the spatial pattern $\boldsymbol{u}_0$ is the key design parameter that determines which dynamical modes are excited. An improperly chosen $\boldsymbol{u}_0$ may excite only the modes already dominant in the passive state, failing to enlarge the smaller eigenvalues that are critical for system identification. The problem of how to design the optimal spatial pattern $\boldsymbol{u}_0$ to most effectively excite the weakly observable dynamics will be addressed in a later section.

Demonstration of Sinusoidal Inputs

In this section, we demonstrate how to design the frequency of oscillatory input by analyzing the eigenvalues of the covariance matrix. Theoretical formulations indicate that tuning the frequency of oscillatory perturbation inputs is more complex compared to that of impulse and step inputs. We generate neural dynamics governed by the connectivity matrix $\boldsymbol{A}$, which is characterized by designated dynamical modes and compare its eigenvalues λ_i with eigenvalues μ_i of the covariance matrix of the perturbed state vectors. The demonstration for impulse and step inputs can be found in the Supplementary Material B.2.

The simulation results show that an oscillatory input with the same frequency as a mode of the system matrix minimized the estimation error when the system had a single mode. Figure 4a shows the system matrix and its eigenvalues, which correspond to a single 10 Hz mode. We generated time series data using this system matrix and estimated the matrix. Figure 4b illustrates the changes in the sum of eigenvalues μ_i, the sum of reciprocal eigenvalues, and the estimation error defined by the Frobenius norm. The sum of eigenvalues is maximized by a 10 Hz sinusoidal input, while the sum of reciprocal eigenvalues is minimized. Consistent with these results, the estimation error is minimized by the 10 Hz sinusoidal input. This tendency can be explained more clearly by plotting the eigenvalues as ellipsoids. We visualized the eigenvalues using eigenvalue-scaled ellipsoids, as shown in Figs. 4c and 4d. Both the major and minor axes of the ellipsoids in Fig. 4c are extended by the sinusoidal inputs, especially by the 10 Hz input. In contrast, the major and minor axes of the ellipsoids in Fig. 4d are substantially shortened by the 10 Hz sinusoidal input.

When the system exhibits multiple modes, the input frequencies should be designed to reflect all these modes. We demonstrate this using the system matrix shown in Fig. 5a, along with simulated time series and the corresponding matrix estimation. Because the system matrix contains two distinct pairs of eigenvalues (10 Hz and 20 Hz), we constructed the evaluation input $\boldsymbol{u}$ as a combination of two frequencies. The sum of eigenvalue reciprocals is minimized only when the input contained both 10 Hz and 20 Hz components; inputs with a single frequency yield larger values (Fig. 5b). To further evaluate the effect of input design, we compared the two-frequency input with a flat-spectrum input (Fig. 5c). When the total input energy was normalized, the two-frequency input achieves better identification performance than the flat-spectrum input, which is commonly employed in control and system identification studies. This result highlights the importance of tailoring the input spectrum to the system's intrinsic dynamics rather than relying on uniform excitation. The superior performance of the two-frequency combination can be understood through the eigenvalue-scaled ellipsoids in Figs. 5d–g. Figure 5d illustrates ellipsoids determined by the three dominant eigenvalues. The ellipsoid volume is expanded by

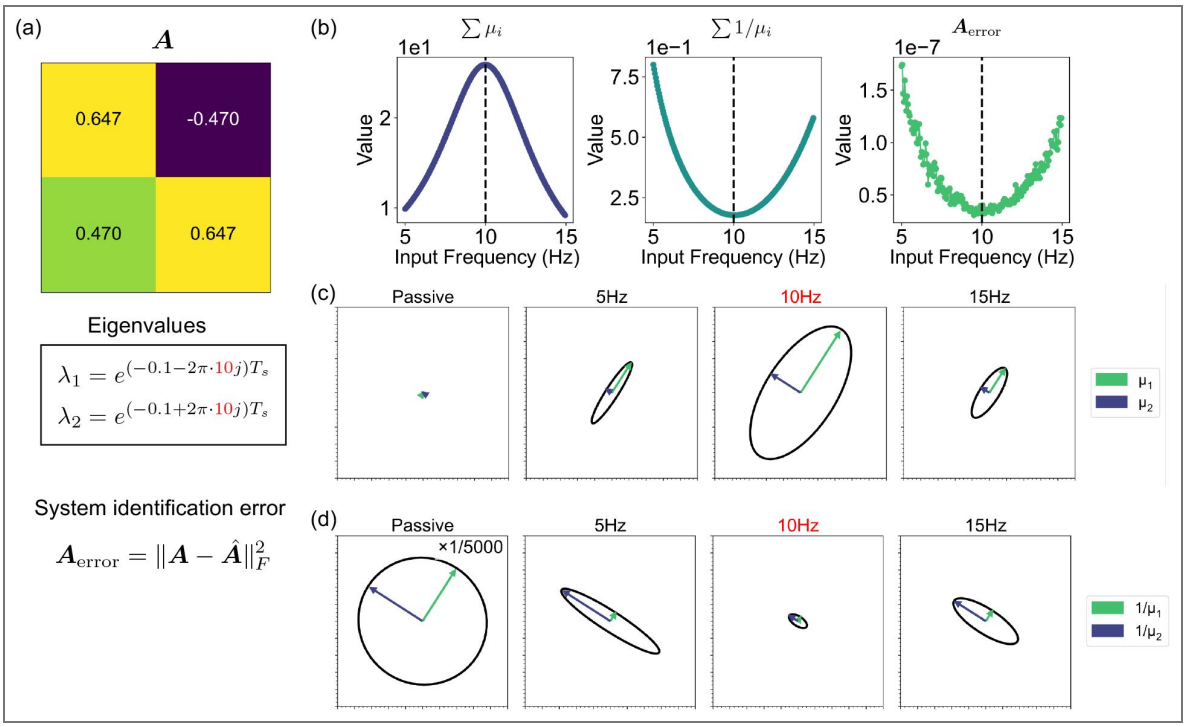


Fig. 4. Error in the system matrix and eigenvalue-scaled ellipsoids.

(a) State transition matrix ***A*** and its eigenvalues λ_i. (b) System frequency responses, showing the sum of eigenvalues (left), the sum of inverse eigenvalues (middle), and the system matrix error (right). (c) Eigenvalue ellipsoids under passive, 5 Hz, 10 Hz, and 15 Hz conditions. (d) Reciprocal-eigenvalue ellipsoids under passive, 5 Hz, 10 Hz, and 15 Hz conditions.

three types of sinusoidal inputs (10 Hz, 20 Hz, and the combined input). However, the third axis (blue line) is not extended under single-frequency inputs, as shown in Fig. 5e, leading to larger estimation errors. Small eigenvalues disproportionately increase the sum of reciprocal eigenvalues, as seen in

Figs. 5f and 5g. In these cases, the reciprocal eigenvalue-scaled ellipsoids are enlarged along the third axis (blue line) when only a single-frequency input is applied. By contrast, the combined input uniformly increases all eigenvalues, including the third, thereby successfully minimizing the reciprocal eigenvalue-scaled ellipsoid volume. These findings indicate that sinusoidal inputs should be designed to match multiple system modes rather than single modes. It should be noted that the true system modes cannot generally be known a priori. They must be identified through iterative experiments and refinement of the estimated system matrix. This practical procedure for optimal perturbation design is described in a later section (see Practical Designing Procedure for Optimal Perturbation Inputs).

Demonstration of Location Tuning

This section illustrates how our theoretical framework enables location-specific tuning of perturbation inputs within neural networks. Placing the perturbation input on a node that has high-weight connections with other nodes effectively minimizes tr $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1})$, and leads to an improvement in system identification, in accordance with Eq. 8.

To examine this relationship in practice, we constructed an 8-node network as shown in Figs. 6a and 6b. ... The network consists of eight nodes (N = 8), structured into three modes (Nodes 1-6) and two additional nodes (Nodes 7-8). Nodes 7 and 8 have only outgoing edges. This structural distinction between Nodes 7 and 8 plays a critical role in the overall connectivity and dynamics of the network. As shown in Fig. 6c, Node 8 has a high weighted out-degree [61], 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. 6d, absolute eigenvalues are differentiated. The eigenvectors shown in Fig. 6e 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 an impulse-shaped perturbation input.

The results, shown in Figs. 6f-g, demonstrate a clear relationship between the input location and estimation accuracy. Both the trace of the inverse covariance matrix tr $\mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{X}}^{-1})$ and the estimation error of $\boldsymbol{A}$ are minimized when the impulse input is applied to Node 8.This finding directly corresponds to the network's structural properties, specifically the weighted out-degree (Fig. 6c). Node 8 possesses the highest weighted out-degree, followed by Node 7, which is the second-best stimulation target.The theoretical explanation is straightforward: nodes with higher out-degrees function as "hub nodes" or "broadcasters". A perturbation applied to such a hub propagates more effectively and widely throughout the entire network. This broad excitation increases the variance of the state $\boldsymbol{x}$ in multiple directions, leading to a more uniform enlargement of the eigenvalues μ_i of the state covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{X}}$ (as discussed in Section). By preventing any μ_i from remaining small, the total estimation error, which depends on $(1/\mu_i)$, is effectively minimized.

This finding is further supported by analyzing the system's dynamical modes ($\boldsymbol{A}$'s eigenvectors and eigenvalues). Fig. 6d shows that the system possesses modes with small absolute eigenvalues. These modes are the most critical for system identification, as their small eigenvalues mean they decay rapidly and become unobservable in the passive state, thus dominating the estimation error. Fig. 6e (Eigenvectors) reveals the spatial structure of these modes. Crucially, the eigenvectors corresponding to these rapidly decaying modes (e.g., evec 7, evec 8) have their largest components concentrated at Nodes 7 and 8. This provides a precise dynamic explanation for our results: applying an input to Node 8 is optimal because it most efficiently targets and re-excites the

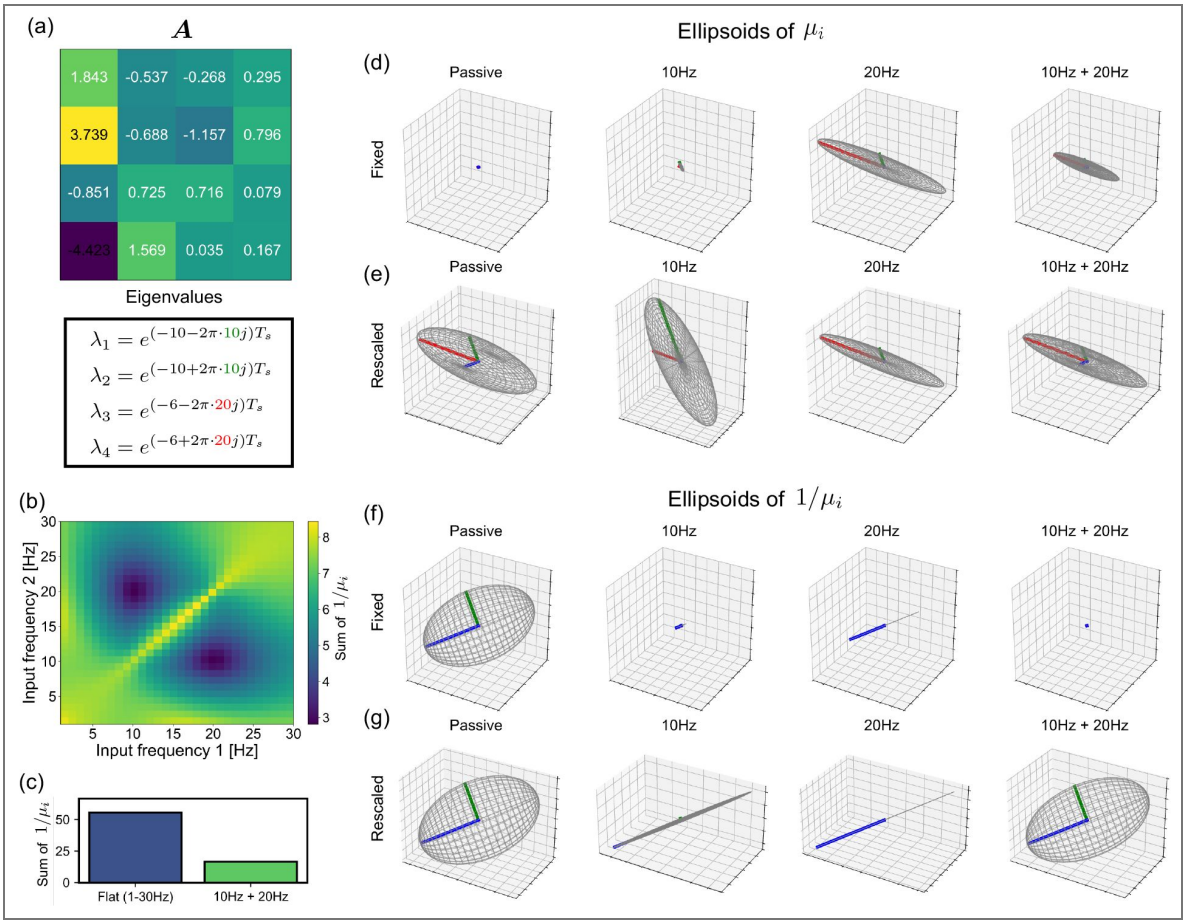


Fig. 5. Simulation of a multi-mode system and eigenvalue-scaled ellipsoids.

(a) System matrix ***A*** and its eigenvalues; the system exhibits two damped oscillatory modes at 10 Hz and 20 Hz. (b) Theoretical estimation error, defined as $\sum(1/\mu_i)$ for two-frequency input combinations; cooler colors indicate smaller estimation errors. (c) Comparison of the reciprocal eigenvalue sum for flat-spectrum versus composite-frequency inputs. (d) Eigenvalue-scaled ellipsoids (μ-scaled) under four input conditions (columns): Passive, 10 Hz, 20 Hz, and 10 Hz + 20 Hz; ellipsoid axes align with the eigenvectors. (e) Relative-scale view of (d). (f) Eigenvalue-scaled ellipsoids (inverse-scaled, $1/\mu_i$) for the same four conditions. (g) Relative-scale view of (f).

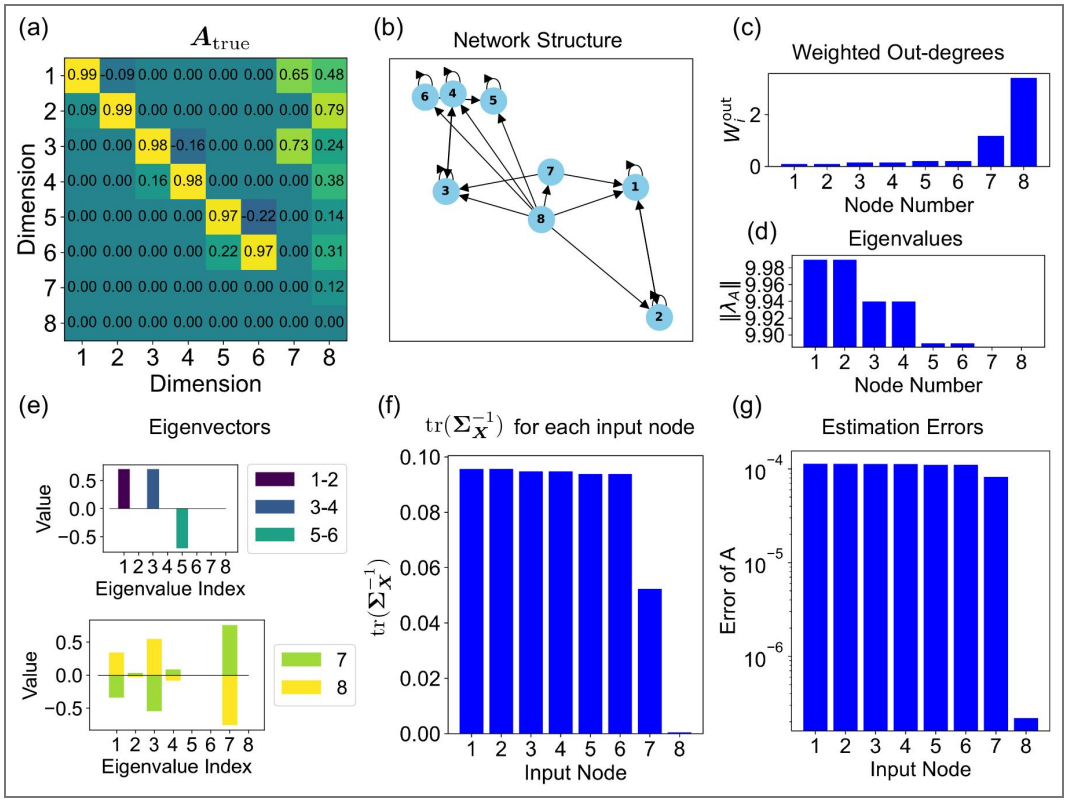


Fig. 6. Perturbation input locations and model estimation errors.

(a) The network was constructed with an 8 × 8 adjacency matrix ***A*** and designed to have three distinct dynamical modes. Nodes 7 and 8 have outgoing edges only, with node 8 having a higher number of connections. (b) Graph representation of the network structure. (c) Weighted out-degree for each node. (d) Absolute eigenvalues of matrix ***A***. (e) Eigenvectors of matrix ***A***. (f) Trace of the inverse covariance matrix of ***X*** when impulse input was applied to each node. (g) Estimation errors of the adjacency matrix ***A*** when the impulse input was applied individually to each node.

system's weakest and most unobservable dynamical modes. This alignment between the input location and the structure of the critical modes ensures that all dynamics are sufficiently excited, minimizing the estimation error.

This simulation demonstrates that, given a tentative connectivity matrix, an effective perturbation input (such as TMS or tDCS) can be designed by targeting the node with the highest weighted out-degree. This structural heuristic succeeds because it effectively identifies the nodes that can most efficiently excite the system's weakest and most unobservable dynamical modes—those that decay rapidly in the passive state. This alignment between the static network structure and the dynamic modes ensures that all aspects of the system's dynamics are sufficiently excited, minimizing the estimation error.

Neural State Classification

We demonstrate the effectiveness of our framework by applying it to a neural state classification problem. Neural state classification is crucial for diagnosing illnesses and assessing brain conditions in many experimental neuroscience settings [31, 36, 47, 48]. Some of these studies have already applied arbitrary perturbation inputs for classification, but not optimal perturbation inputs. By simulating such real-world applications, we demonstrate how well optimal perturbation design can contribute to neuroscience research.

We designed a neural network with clearly distinct states and considered a simulation setting in which these states are classified using signals of a fixed duration. These distinct states consist of five types, each defined by a unique linear dynamical system characterized by differing eigenvalue spectra and connectivity topologies of matrix $\boldsymbol{A}$ (Fig. 7a). These latent states are intended to mimic neural states associated with different cognitive or behavioral conditions. For example, in a typical motor task experiment, such states could correspond to task contexts such as motor execution or imagery involving the left or right hand, or resting state [47, 48]. The neural signal was simulated under five different states and two stimulation conditions: passive observation and external perturbation. Perturbation was applied as impulse-type inputs, such as TMS. The location of these impulse inputs was determined according to weighted outdegree, in order to minimize the error of the model estimation. The resulting time-series data are shown in Fig. 7b. Using this data, we estimated the underlying dynamical system via a control-based identification approach presented in Eq. 4, which corresponds to an estimation of functional connectivity.

The simulation demonstrates that classification performance under perturbation is superior to that under the passive condition, both in visual inspection and in quantitative evaluation. As revealed by multidimensional scaling (MDS) analysis, state clusters in the passive condition exhibit substantial overlap, whereas perturbation leads to clear separation among the states (Fig. 7c). Classification with linear discriminant analysis (LDA) and 10-fold cross-validation shows the average accuracy is substantially higher in the perturbation condition (98.20%) compared to the passive condition (67.40%), as shown in Fig. 7d. ROC curve analysis further confirms that perturbation substantially improves discriminability across all states (Fig. 7e). These results can be explained by Eq. 8; the perturbation inputs undoubtedly increases the eigenvalues of the covariance matrix, including even the smallest ones, which in turn leads to a decrease in the estimation error of $\boldsymbol{A}$.

To obtain estimates under the passive condition that are comparable to those derived under perturbation, it is necessary to experimentally observe extensive time-series data. Figure 7f illustrates the MDS representation and classification accuracy for different time-series lengths. The leftmost MDS plot ($T = 5$) is the same as the passive condition plot shown in Fig. 7c), indicating that the estimation performance in the passive condition is inferior to that in the perturbation condition. This low accuracy can be improved by using a longer time window ($T = 100$, 1000, 10000). At $T = 10000$, the classification accuracy exceeds the perturbation condition. However, this longer window corresponds to a 2000-fold increase in the required time window. In practical experiments, such long windows are unrealistic because neural states change rapidly over time.

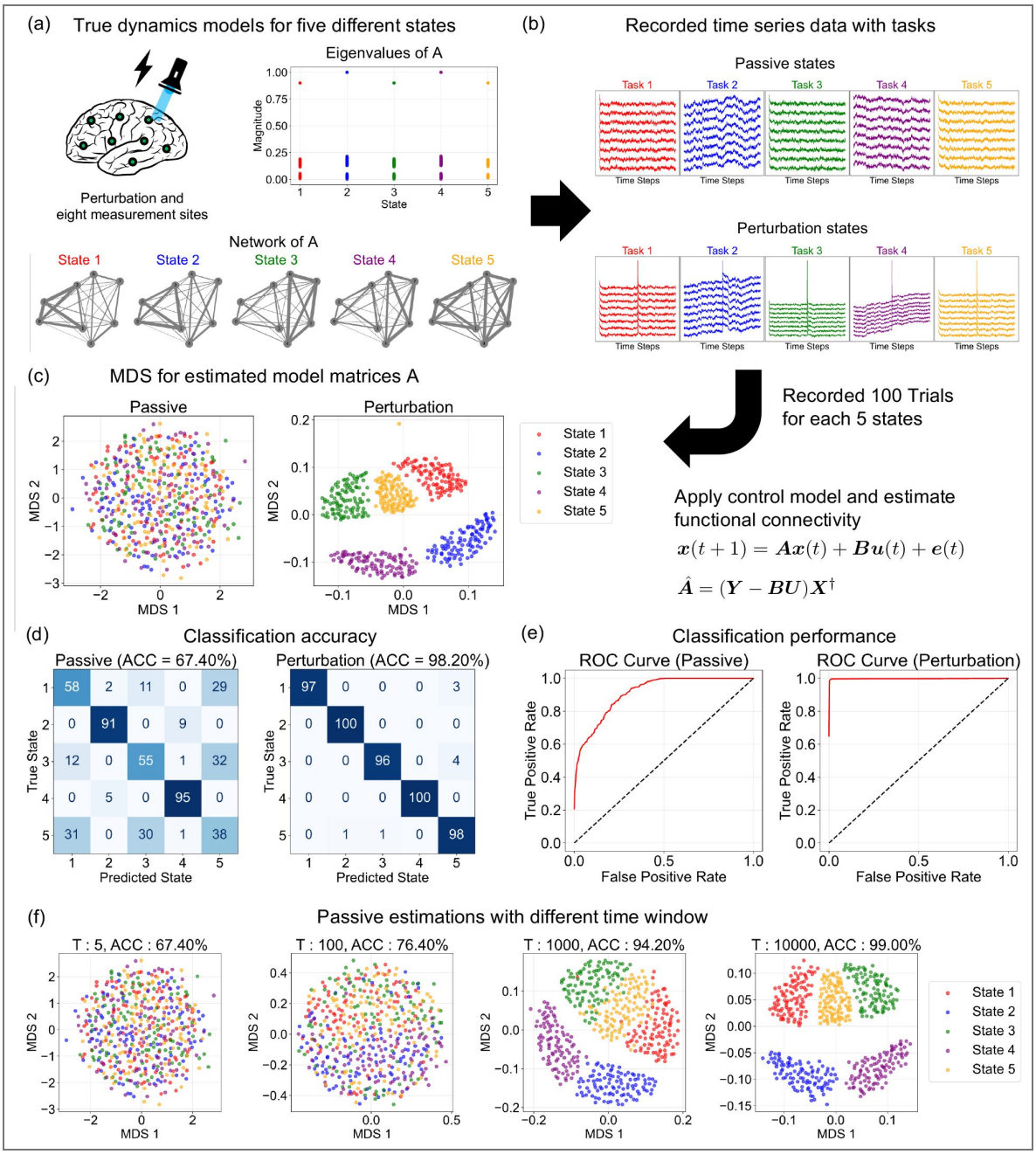


Fig. 7. Neural state classification enhanced by perturbation-aided model estimation.

(a) Ground-truth dynamics models for five distinct latent states, each with unique eigenvalue distributions and network structures of matrix ***A***. (b) Simulated neural activity under five task conditions inducing distinct states, shown separately for the passive (top) and perturbed (bottom) regimes. (c) Estimated model matrices ***A*** visualized using MDS. (d) Confusion matrices showing classification accuracy. (e) ROC curves further confirm the improved classification performance achieved with perturbation. (f) MDS plots for passive condition with increasing time window length, showing clustering and classification accuracy as the time window length *T* increases from 5 to 10,000.

These findings highlight the advantage of perturbation-based neural model estimation, particularly in scenarios demanding fast and accurate neural state classification.

Neural State Transitions via Optimal Control

This section demonstrates the practical utility of our framework by applying it to a neural state control problem [42,49]. Here, we estimate the system parameters from both passive and perturbation states. Then, we determine optimal control inputs which control the neural states to desired targets, and associated control costs. By using perturbation inputs, $\boldsymbol{A}$ is accurately estimated, which in turn allows precise determination of the optimal inputs and the corresponding controlled state transitions (Fig. 8a). Moreover, use of the perturbation approach to derive the model allows the controllability Gramian [42,47], as well as the estimation of control costs for each network area, to be more reliably assessed (Fig. 8b and 8c). 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. 9a 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 a known 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. We generated passive and perturbation states to estimate the parameter matrices. The perturbation condition uses an impulse input with sufficient intensity ($\alpha = 10^{20}$). By using the estimated system model, 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 25 while keeping all other nodes at 0 without any movement.

The simulation results show that perturbation-based estimation enables the precise control of neural states. If the estimation of $\boldsymbol{A}$ is inaccurate, the state transitions under optimal control cannot be executed correctly (Fig. 9b). Nodes 3 and 4 fail to reach the target, and unnecessary activations occur in nodes 1 and 2. Furthermore, unnecessary control inputs appear in nodes 1 and 2. On the other hand, when the model is correctly estimated using a high-strength impulse, the resulting control inputs enable accurate state transitions (Fig. 9c). In this case, the strength of the impulse is correctly concentrated on node 5. As the error in matrix $\boldsymbol{A}$ increases, the estimation errors of the controllability

Gramian $\boldsymbol{W}$ lead to inaccuracies in estimating the control cost (Fig. 9d). Here, we adopt average controllability as an example measure of control cost [42, 46]. The key point here is that, since the controllability Gramian involves multiple products of $\boldsymbol{A}$ (Eq. 22), 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.

Practical Designing Procedure for Optimal Perturbation Inputs

This section presents a practical framework for designing optimal perturbation inputs to estimate neural dynamics. The practical framework progressively refines the input signal through repeated cycles of model identification and perturbation input design. By leveraging this alternating scheme, each iteration utilizes the current model estimate to design a more informative input for the next round of identification. We also focus on perturbations characterized by a broad and uniform frequency distribution, a form commonly adopted in control engineering and

Fig. 8. Effects of system identification on network control theory.

(a) State transitions using models identified from the passive and perturbed conditions. (b) Controllability Gramians with two models. (c) Control costs compputed from the estimated controllability Gramians.

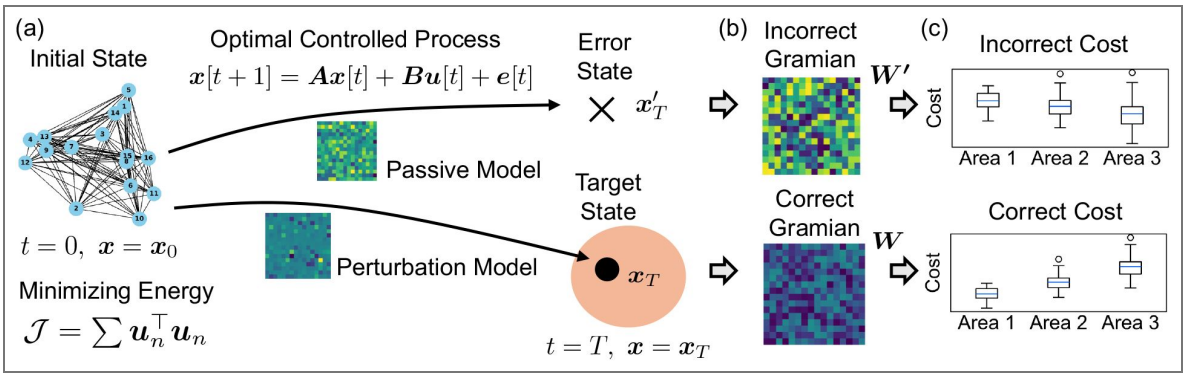
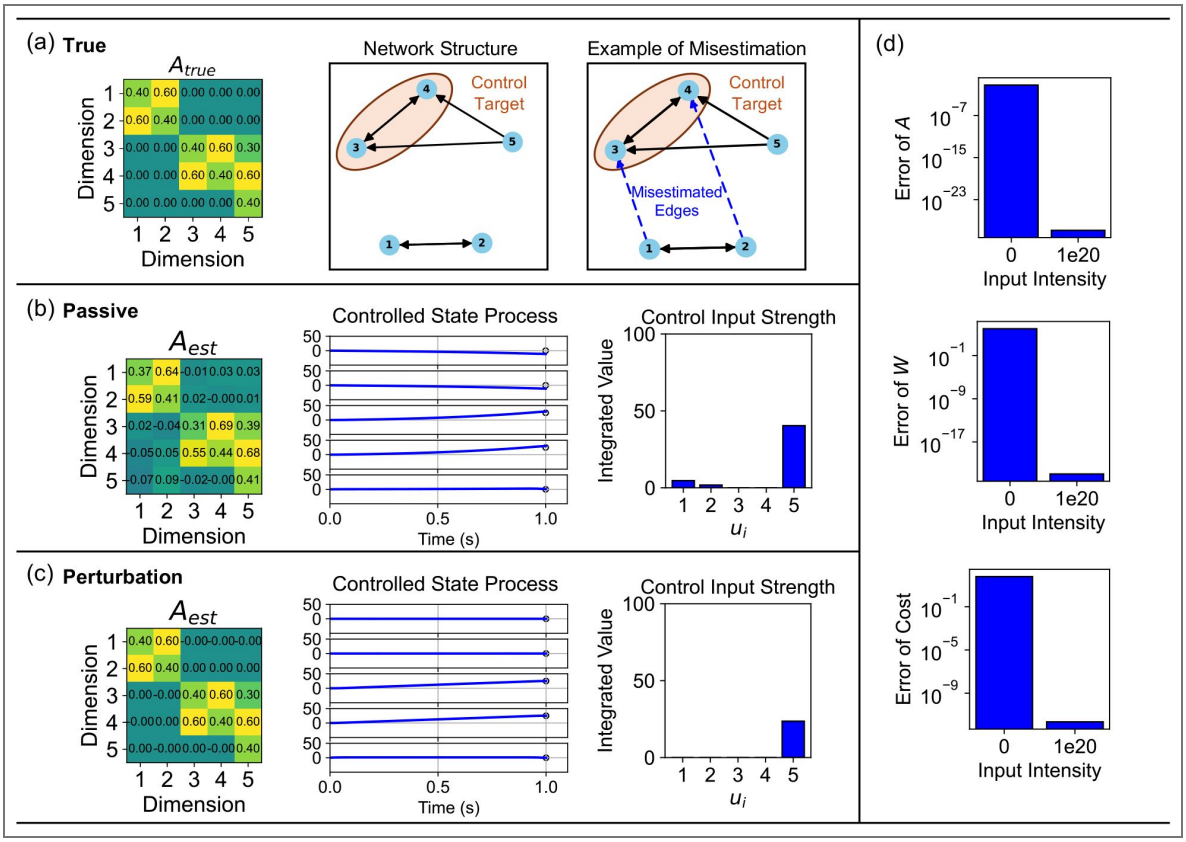


Fig. 9. Optimal control difference between passive and perturbation-based models.

(a) True matrix ***A*** and its network structure. 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 nodes 3 and 4. (b) Matrix estimation failure under passive condition and controlled state transitions. 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 the estimated matrices ***A*** and ***W*** and the control cost (average controllability)



conceptually analogous to transcranial random noise stimulation (tRNS) in neuroscience [62, 63]. As shown in Figs. 5 and B.2, the designed composite-wave input allows for more accurate model estimation than the uniform-frequency input when their total input energies are equal.

Iterative refinement of both the perturbation design and the estimation process progressively improves the accuracy of $\boldsymbol{A}$. The time-series data is collected from 32 points, where the matrix $\boldsymbol{A}$ (Fig. 10a) is designed to have 16 modes. Matrix $\boldsymbol{B}$ was set as the identity matrix. The following procedure outlines a practical approach (Fig. 10c) for this simulation, aiming to precisely estimate the matrix by designing an optimal perturbation input (assuming frequency stimulation, such as tACS).

1. Record spontaneous activity as the passive state and estimate $\boldsymbol{A}_{\text{passive}}$.
2. Based on the estimated matrix, the optimal frequency and target node of the perturbation input are determined through numerical optimization following Eq. 9.
3. The designed perturbation $\boldsymbol{u}_{\text{design1}}$ is applied to the system, and a more precise matrix $\boldsymbol{A}_{\text{design1}}$ is identified.
4. The obtained matrix $\boldsymbol{A}_{\text{design1}}$ is again used to design a new perturbation input $\boldsymbol{u}_{\text{design2}}$ and estimate a more accurate matrix $\boldsymbol{A}_{\text{design2}}$. This iterative process can be continued until convergence criteria, such as changes in $\boldsymbol{A}$, are met.

The iterative design framework dramatically improved estimation accuracy. As shown in Fig. 10b, the estimation process converges with each design iteration: the error from the second iteration (2nd design) is smaller than the first (1st design), which in turn is substantially lower than the initial estimates. This demonstrates the practical power of the framework. Even when starting with a poor model derived from passive data (purple line, Iteration 1), the first designed input (design-1) significantly improves the model. Furthermore, we compared our designed inputs (conceptually analogous to optimized tACS) against a flat-frequency input (a practical approach analogous to tRNS). While the flat-frequency input (teal line, Iteration 1) provided a better starting point than passive data, our iterative design procedure quickly outperformes it. This confirms the central hypothesis of this paper: a model-based, iteratively designed input is superior to both passive observation and non-specific (flat-spectrum) stimulation. Even without a priori knowledge, this practical framework allows for the progressive refinement of system identification, enabling a deeper and more accurate understanding of neural dynamics.

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 [9, 42, 64]. Linear models are particularly effective for providing accurate approximations of nonlinear systems within a specific operating range [65]. 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

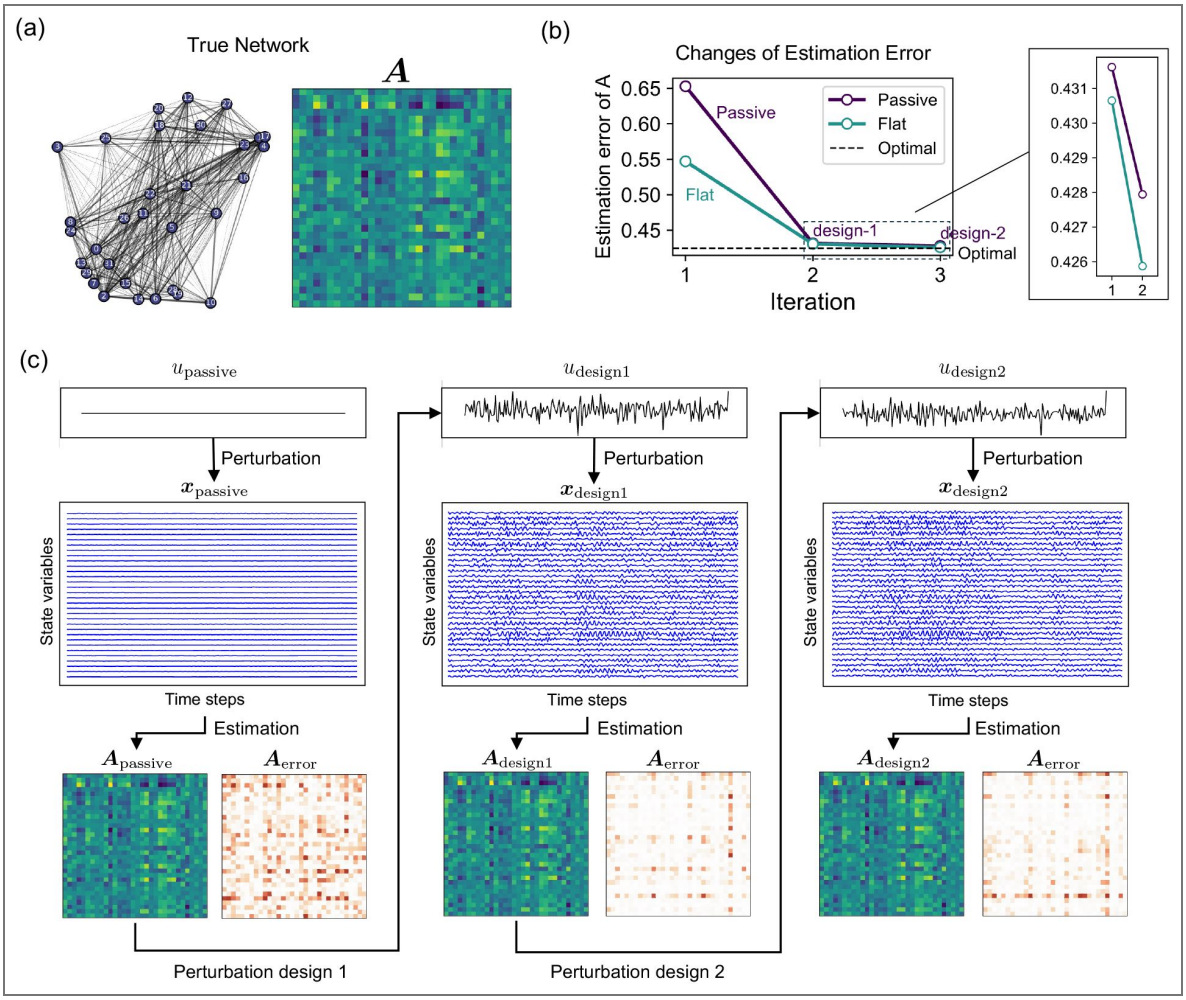


Fig. 10. Simulation results for a 32 channel neural activity model.

(a) True matrix $\boldsymbol{A}$ and its network structure, which was designed to include both oscillatory and attenuating components. (b) Comparison of Estimation errors. The purple line indicates the iterative design process initialized from the passive-state estimate (Iteration 1). The teal line shows the iterative design starting from a flat-spectrum input (Iteration 1). Both processes converge towards the theoretically optimal error (dashed line) as the design is refined (design-1, design-2). (c) Schematic of the iterative procedure. An initial estimated system matrix ($\boldsymbol{A}_{\text{passive}}$) is used to design the first perturbation input ($\boldsymbol{u}_{\text{design1}}$), which yields a better estimate ($\boldsymbol{A}_{\text{design1}}$), and the cycle repeats.

necessary to incorporate advanced frameworks such as the Koopman operator [66,67] and system identification methods leveraging machine learning techniques [68,69]. 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 [70–72]. The framework proposed in this paper can be naturally and meaningfully extended to higher-order VAR models, broadening its applicability to more complex temporal dependencies.

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 [17, 73]. For stimulus intensity, our proposed framework suggested that a stronger stimulus improves system identification; however, an excessively strong stimulus poses safety risks [74]. With respect to sinusoidal 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 [74, 75]. Furthermore, the selection of stimulation sites is often dictated by experimental accessibility or ethical considerations [76, 77]. 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 [47]. 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, 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.

Methods

State Vector Simulation and Model Estimation

To investigate appropriate perturbation inputs for the identification of neural systems, we constructed neural dynamics using matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ to generate the temporal evolution of state variables. From the generated dynamics, we estimated the model matrix $\boldsymbol{A}$, while assuming that the model matrix $\boldsymbol{B}$ was known.

Given an initial condition $\boldsymbol{x}(0)$, the neural dynamics were generated according to the true matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and the governing equation for $\boldsymbol{x}(t)$ (Eq. 1), with a predefined noise time series $\boldsymbol{\xi}(t)$ added at each time step. For the discrete-time simulations, the sampling rate was set to an appropriate value for each simulation. The model matrix $\boldsymbol{A}$ was then estimated from the generated time-series states $\boldsymbol{x}(t)$ and applied perturbation inputs $\boldsymbol{u}(t)$ using the OLS.

Both the generation of neural dynamics and the model estimation were repeated for a predefined number of trials, with variations in initial conditions and noise sequences. The estimation errors of the matrices were quantified using the Frobenius norm. Details of the simulation parameters are provided in the Supplementary Material B.1.

Eigenvector Alignment for eigenvalue-scaled ellipsoids

To examine the relationship between the eigenvalues μ_i of the state covariance matrix $\boldsymbol{\Sigma_X}$ and the input frequencies, we plotted an eigenvalue-scaled ellipsoid, thereby providing an intuitive representation of optimal input design. However, the eigenvectors of $\boldsymbol{X}$ are not fixed; they may vary depending on the applied input, which complicates direct comparison across conditions. To enable consistent comparison of eigenvalues across input conditions, we reordered eigenvalues according to the similarity of their associated eigenvectors to those obtained in the passive condition, which served as the reference basis. For each reference eigenvector $\boldsymbol{t}_i$, we identified the most similar eigenvector $\boldsymbol{u}_j$ from the input condition by maximizing the absolute inner product

$$j^* = \arg\max_j \left| \boldsymbol{t}_i^\top \boldsymbol{u}_j \right|. \tag{16}$$

The eigenvalue $\boldsymbol{\mu_{j^*}}$ corresponding to $\boldsymbol{u_{j^*}}$ was then assigned to the (i)-th position of the reordered list. Each $\boldsymbol{u}_j$ was used only once, ensuring a one-to-one correspondence between reference and input eigenvectors. This alignment resolves the permutation and sign ambiguities inherent in eigendecomposition, and allows eigenvalues to be compared across conditions along a common axis.

Weighted out-degree

To quantify which node influences other nodes, metrics designed for weighted networks were used [10, 78, 79]. The weighted out-degree [61] of node *i*, excluding self-loops, is given by:

$$w_i^{\text{out}} = \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij}, \tag{17}$$

where the element a_{ij} represents the weight of the directed edge from node *i* to node *j*. The summation explicitly excludes self-loops ($j \neq i$). 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.

Optimal Control and Network Controllability

Accurately estimating dynamical models under perturbation not only facilitates precise inference of the neural dynamics but also contributes to the accurate estimation of control inputs for neural state transitions. Recently, the control theory which utilizes the controllability Gramian and control costs is applied for neuroscience, and provides insights into the efficiency and feasibility of inducing specific neural state transitions [42, 49] with some limitations [80, 81]. For clarify, 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 $\boldsymbol{x}_T$:

$$\mathcal{M}(\{\boldsymbol{u}(t)\}_{t=0}^{T-1}) = \sum_{t=0}^{T-1} \boldsymbol{u}(t)^\top \boldsymbol{u}(t), \tag{18}$$

under

$$\boldsymbol{x}(0) = \boldsymbol{x}_0, \quad \boldsymbol{x}(T) = \boldsymbol{x}_T. \tag{19}$$

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

$$\boldsymbol{u}(t)^* = \arg\min_{\boldsymbol{u}} \mathcal{M}(\{\boldsymbol{u}(t)\}_{t=0}^{T-1}). \tag{20}$$

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

$$\boldsymbol{u}(t)^* = \boldsymbol{B}^\top (\boldsymbol{A}^\top)^{T-t-1} \boldsymbol{W}^{-1} (\boldsymbol{x}(T) - \boldsymbol{A}^T \boldsymbol{x}(0)), \tag{21}$$

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

$$\boldsymbol{W} = \sum_{t=0}^{T-1} \boldsymbol{A}^t \boldsymbol{B} \boldsymbol{B}^\top (\boldsymbol{A}^\top)^t. \tag{22}$$

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 [42, 47].

Average Controllability for Control Cost

While the optimal control problem (Eq. 20) calculates the specific input energy for a given state transition, a more general, state-independent metric is often used to characterize the system's overall controllability. This metric, average controllability, is a state-independent metric used to quantify the system's overall ease of control [42, 46]. It is defined as the trace of the controllability Gramian:

$$\text{Average Controllability} = \text{tr}(\boldsymbol{W}) = \sum_{i=1}^{N} \lambda_i(\boldsymbol{W}) \tag{23}$$

A larger trace indicates that the system is, on average, more controllable (i.e., can be moved to various states with less input energy). This metric is therefore used as a measure of control efficiency.

Data availability

The current manuscript is a computational study, so no biological data have been generated. The code for the simulations and analyses presented in this paper is openly accessible at *https://github.com/mikito-ogino/NeuroPerturbID*

Acknowledgements

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.

Additional files

Supplementary Material

Additional information

Funding

Funder	Grant reference number	Author
MEXT | Japan Science and Technology Agency (JST)	https://doi.org/10.52926/jpmjms2012	Mikito Ogino
MEXT | Japan Society for the Promotion of Science (JSPS)	24K20462	Masafumi Oizumi

Author ORCID iDs

Mikito Ogino: https://orcid.org/0000-0003-1089-4469
Daiki Sekizawa: https://orcid.org/0009-0004-0196-2612
Jun Kitazono: https://orcid.org/0000-0001-6701-3947

Masafumi Oizumi: https://orcid.org/0000-0001-8802-2607

References

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

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

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

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

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

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

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

[8] **Stevens M. C.** (2009) The developmental cognitive neuroscience of functional connectivity. *Brain Cogn* **70**:1-12

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

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

[11] **Friston K.,** Moran R., Seth A. K. (2013) Analysing connectivity with granger causality and dynamic causal modelling. *Curr. Opin. Neurobiol* **23**:172-178

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

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

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

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

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

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

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

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

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

[21] **Goebel R.,** Roebroeck A., Kim D.-S., Formisano E. (2003) Investigating directed cortical interactions in timeresolved fMRI data using vector autoregressive modeling and granger causality mapping. *Magn Reson Imaging* **21**:1251-1261

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

[23] **Ting C.,** Seghouane A., Khalid M. U., Salleh S. (2015) Is first-order vector autoregressive model optimal for fMRI data?. *Neural Comput* **27**:1857-1871

[24] **Seth A.,** Barrett A., Barnett L. (2015) Granger causality analysis in neuroscience and neuroimaging. *J. Neurosci* **35**:3293-3297

[25] **Cekic S.,** Grandjean D., Renaud O. (2018) Time, frequency, and time-varying granger-causality measures in neuroscience. *Stat Med* **37**:1910-1931

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

[27] **Bassett D.,** Sporns O. (2017) Network neuroscience. *Nat. Neurosci* **20**:353-364

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

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

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

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

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

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

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

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

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

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

[38] **Massimini M.,** et al. (2007) Triggering sleep slow waves by transcranial magnetic stimulation. *Proc. Natl. Acad. Sci. U. S. A* **104**:8496-8501

[39] **George M. S.** (2019) Whither TMS: A one-trick pony or the beginning of a neuroscientific revolution?. *Am. J. Psychiatry* **176**:904-910

[40] **Bernal-Casas D.,** Lee H. J., Weitz A. J., Lee J. H. (2017) Studying brain circuit function with dynamic causal modeling for optogenetic fMRI. *Neuron* **93**:522-532.e5

[41] **Grimm C.,** et al. (2024) Tonic and burst-like locus coeruleus stimulation distinctly shift network activity across the cortical hierarchy. *Nat. Neurosci* **27**:2167-2177

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

[43] **Deng S.,** Li J., Thomas Yeo B. T., Gu S. (2022) Control theory illustrates the energy efficiency in the dynamic reconfiguration of functional connectivity. *Commun. Biol* **5**:295

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

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

[46] **Moradi Amani A.,** et al. (2024) Controllability of functional and structural brain networks. *Complexity* **2024**

[47] **Shikauchi Y.,** et al. (2025) Quantifying state-dependent control properties of brain dynamics from perturbation re-sponses. *bioRxiv*

[48] **Angulo-Sherman I. N.,** Rodríguez-Ugarte M., Sciacca N., Iáñez E., Azorín J. M. (2017) Effect of tDCS stimulation of motor cortex and cerebellum on EEG classification of motor imagery and sensorimotor band power. *J. Neuroeng. Rehabil* **14**:31

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

[50] **Betzel R. F.,** Gu S., Medaglia J. D., Pasqualetti F., Bassett D. S. (2016) Optimally controlling the human connectome: the role of network topology. *Sci. Rep* **6**:30770

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

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

[53] **Ahmed S.,** Nozari E. (2022) On the linearizing effect of spatial averaging in large-scale populations of homogeneous nonlinear systems.

[54] **Nozari E.,** et al. (2024) Macroscopic resting-state brain dynamics are best described by linear models. *Nat. Biomed. Eng* **8**:68-84

[55] **Green M.,** Moore J. B. (1986) Persistence of excitation in linear systems. *Syst Control Lett* **7**:351-360

[56] **Shimkin N.,** Feuer A. (1987) Persistency of excitation in continuous-time systems. *Syst Control Lett* **9**:225-233

[57] **Jenkins B. M.,** Annaswamy A. M., Lavretsky E., Gibson T. E. (2018) Convergence properties of adaptive systems and the definition of exponential stability. *SIAM J Control Optim* **56**:2463-2484

[58] **Lu X.,** Cannon M. (2023) Robust adaptive model predictive control with persistent excitation conditions. *Automatica* **152**:110959

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

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

[61] **Acemoglu D.** (2012) The network origins of aggregate fluctuations. *Econometrica* **80**:1977-2016

[62] **Paulus W.** (2011) Transcranial electrical stimulation (tES - tDCS; tRNS, tACS) methods. *Neuropsychol Rehabil* **21**:602-617

[63] **Paulus W.,** Nitsche M. A., Antal A. (2016) Application of transcranial electric stimulation (tDCS, tACS, tRNS): From motor-evoked potentials towards modulation of behaviour. *Eur. Psychol* **21**:4-14

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

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

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

[67] **Chow C.,** Dan T., Styner M., Wu G. (2024) Understanding brain dynamics through neural koopman operator with structure-function coupling. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Springer. pp. 509-518

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Peer reviews

Joint Public Review:

Summary:

Inferring so-called "functional connectivity" between neurons or groups of neurons is important both for validating models and for inferring brain state. Under the assumption that brain dynamics is linear, the authors show that the error in estimating functional connectivity depends only on the eigenvalues of the covariance matrix of the observed data, and it is the small eigenvalues -corresponding to directions in which the variance of the brain activity is low - that lead to large estimation errors. Based on this, the authors show that to achieve low estimation error, it's important to excite the resonant frequencies and perturb well-connected hubs. The authors propose a practical iterative approach to estimate the functional connectivity and demonstrate faster convergence to the optimal estimate compared to passive observation.

Strengths:

The main contribution of the study is the derivation of an explicit expression for the error in functional connectivity that depends only on the covariance matrix of the observed data. If valid, this result can have a profound impact on the field. The study also motivates the current shift to closed-loop experiments by demonstrating the effectiveness of active learning in the system using perturbation, in comparison to passive estimation from resting-state activity. Finally, the relative simplicity of the model makes its practical applications straightforward, as the authors illustrate in the context of brain state classification and neural control.

Weaknesses:

The derivation of the main error term misses some important steps, which complicates peer review at this stage. In particular, factorisation of the covariance into noise and the inverse of the observation covariance matrix needs a more thorough justification. The cited sources do not contain the derivation for a noise term with full covariance, which is essential for deriving this error term.

The practical recommendation at the end of the paper also requires clearer guidance on how the design perturbations are constructed, and how many times and for how long the system is stimulated in each iteration of the experiment.

Finally, there is no analysis of model mis-specification. In particular, the true dynamics are unlikely to be linear; the noise is unlikely to be either Gaussian or uncorrelated across time; and the B matrix is unlikely to be known perfectly. We're not suggesting that the authors consider a more complex model, but it's important to know how sensitive their method is to model mismatch. If nothing can be done analytically, then simulations would at least provide some kind of guide.

https://doi.org/10.7554/eLife.110030.1.sa0

Author response:

We thank the editors and reviewers for their careful reading of our manuscript and for their insightful comments. We appreciate the opportunity to clarify several aspects of the derivations and experimental design, and we will revise the manuscript accordingly. Below we provide responses to the major weaknesses raised by the reviewers.

> *The derivation of the main error term misses some important steps, which complicates peer review at this stage. In particular, factorisation of the covariance into noise and the inverse of the observation covariance matrix needs a more thorough justification. The cited sources do not contain the derivation for a noise term with full covariance, which is essential for deriving this error term.*

Thank you for pointing this out. We agree that the derivation of the main error term should be presented more explicitly to facilitate peer review. In the revised manuscript, we will explicitly cite the relevant equation numbers from the references to make each step of the argument easier to follow. We will also revise the text to more clearly discuss the assumption on the noise covariance matrix.

> *The pratical recommendation at the end of the paper also requires clearer guidance on how the design perturbations are constructed, and how many times and for how long the system is stimulated in each iteration of the experiment.*

Thank you for this helpful suggestion. We agree that the practical implementation of the experimental design should be explained more clearly. In the revised manuscript, we will provide a more explicit description of how the input perturbations are constructed in each iteration. To more clearly explain how many times and for how long the system is stimulated, we will clarify the stopping criterion used in the iterative procedure and the time length of the external inputs. As shown in Eq. (8), the estimation error scales approximately as 1/T, so longer measurements improve accuracy. For clearer guidance, we will add additional explanations on the relation between the stimulation time and estimation accuracy, as well as on the role of iterative input design.

> *Finally, there is no analysis of model mis-specification. In particular, the true dynamics are unlikely to be linear; the noise is unlikely to be either Gaussian or uncorrelated across time; and the B matrix is unlikely to be known perfectly. We're not suggesting that the authors consider a more complex model, but it's important to know how sensitive their method is to model mismatch. If nothing can be done analytically, then simulations would at least provide some kind of guide.*

We thank the reviewer for raising this important point. We agree that it is important to understand how sensitive the proposed method is to model mismatch. While our current theoretical analysis assumes linear dynamics with Gaussian noise for analytical tractability,

real systems may deviate from these assumptions in several ways, including nonlinear dynamics, temporally correlated noise, or imperfect knowledge of the input matrix B. To address this concern, we will add simulation experiments to examine the robustness of our method under several types of model misspecification. These simulations will provide practical guidance on how deviations from the assumed model affect estimation performance. We will include these results and discuss their implications in the revised manuscript.

https://doi.org/10.7554/eLife.110030.1.sa2