nature methods

Article

https://doi.org/10.1038/s41592-025-02654-x

Mapping effective connectivity by virtually perturbing a surrogate brain

Received: 13 March 2024

Accepted: 3 March 2025

Published online: 22 April 2025

Check for updates

Zixiang Luo[1,6], Kaining Peng[1], Zhichao Liang[1], Shengyuan Cai[1], Chenyu Xu[2], Dan Li[1], Yu Hu[3,4], Changsong Zhou[5] & Quanying Liu[1]

Effective connectivity (EC), which reflects the causal interactions between brain regions, is fundamental to understanding information processing in the brain; however, traditional methods for obtaining EC, which rely on neural responses to stimulation, are often invasive or limited in spatial coverage, making them unsuitable for whole-brain EC mapping in humans. Here, to address this gap, we introduce Neural Perturbational Inference (NPI), a data-driven framework for mapping whole-brain EC. NPI employs an artificial neural network trained to model large-scale neural dynamics, serving as a computational surrogate of the brain. By systematically perturbing all regions in the surrogate brain and analyzing the resulting responses in other regions, NPI maps the directionality, strength and excitatory/inhibitory properties of brain-wide EC. Validation of NPI on generative models with known ground-truth EC demonstrates its superiority over existing methods such as Granger causality and dynamic causal modeling. When applied to resting-state functional magnetic resonance imaging data across diverse datasets, NPI reveals consistent, structurally supported EC patterns. Furthermore, comparisons with cortico-cortical evoked potential data show a strong resemblance between NPI-inferred EC and real stimulation propagation patterns. By transitioning from correlational to causal understandings of brain functionality, NPI marks a stride in decoding the brain's functional architecture and facilitating both neuroscience studies and clinical applications.

The brain functions as a complex network of interconnected regions that collaboratively process external stimuli to generate behavior[1,2]. Understanding the flow of information between these regions is crucial for unraveling brain functions[3]. While structural connectivity (SC) maps the physical wiring of the brain and functional connectivity (FC) captures statistical dependencies among neural activities, neither fully characterizes the directional flow of information[4,5]. Effective connectivity (EC), delineating the causal interactions between brain regions, is thus essential for understanding information flow and critical in selecting target nodes for neuromodulation in brain disorder treatments[6,7].

EC is traditionally derived through neurostimulation experiments, such as optogenetics[8,9] or deep-brain stimulation (DBS)[10]. These methods involve perturbing specific brain area and monitoring the resultant neural responses in other areas, thereby providing direct evidence of causality; however, such 'perturb and record' procedures are usually invasive and do not scale well for whole-brain analysis. Computational

[1]Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China. [2]Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, USA. [3]Department of Mathematics, The Hong Kong University of Science and Technology, Hong Kong SAR, China. [4]Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China. [5]Department of Physics, Centre for Nonlinear Studies, Hong Kong Baptist University, Hong Kong SAR, China. [6]Present address: Division of Life Science, The Hong Kong University of Science and Technology, Hong Kong SAR, China. e-mail: liuqy@sustech.edu.cn

approaches offer noninvasive alternatives but often suffer from inaccuracies, especially when applied at a whole-brain scale. Model-based methods, such as dynamic causal modeling (DCM), heavily rely on underlying model assumptions and are prone to biases from model mismatches[11]. On the other hand, model-free methods such as Granger causality (GC) are adept at discerning the directionality of EC but struggle to accurately measure its strength or differentiate between excitatory and inhibitory influences[12]. Moreover, the interpretation of EC varies across computational frameworks, leading to ambiguity in the interpretation of EC inferred from computational and experimental approaches.

The advent of big data in neuroscience, propelled by advanced imaging and electrophysiological techniques, has facilitated the use of artificial neural networks (ANNs) to analyze complex neural data[13,14]. For example, recurrent neural network models have been employed to learn temporal dynamics of brain signals and infer EC directly from the learned weight matrices[15,16]. While these models can capture brain dynamics, there is no guarantee that the learned weights reflect the underlying EC, particularly when the model's assumptions do not align with the brain's underlying dynamics or when applied to large-scale networks[17]. Perturbation analysis in ANNs offers a promising alternative for investigating causality by modulating input variables and observing resulting output changes[18,19]. This approach mirrors stimulation-evoked potentials used in neuroscience to infer EC[20,21]. Building on this conceptual parallel, our study integrates perturbation-based analyses into a data-driven framework to uncover brain-wide causality.

Here, we introduce Neural Perturbational Inference (NPI), a noninvasive framework for mapping whole-brain EC. NPI employs an ANN trained to model brain dynamics as a computational surrogate of the brain. Once the ANN is trained to capture large-scale neural interactions, systematic perturbations of the model yield a comprehensive map of causal relationships among brain regions. This framework identifies the directionality, strength, and excitatory/inhibitory properties of whole-brain causal interactions. The effectiveness of NPI is validated using a variety of generative models with known ground-truth EC. Furthermore, NPI demonstrates a substantial match with cortico-cortical evoked potentials, confirming its accuracy in reflecting real causal interactions in the brain.

Results

Neural Perturbational Inference

NPI is a noninvasive framework for inferring EC from neural signals (Fig. 1a–d). Conceptually, NPI mimics experimentally perturbing the real brain through neurostimulation. It uses an ANN as a computational surrogate to replace the real brain, which enables efficient whole-brain perturbation and observation. While brain imaging and electrophysiological recordings provide access to the collective activity of multiple brain regions, the intricate ways these regions interact to process information remain unclear (Fig. 1a). NPI aims to infer EC, directed causal connections, among regions in the entire brain.

First, an ANN is trained to predict the brain state at the next time step based on the preceding three time steps by minimizing the one-step-ahead prediction error (Fig. 1b). To validate the ANN's ability to capture interaction relationships between brain regions, the predicted outputs were recursively fed into the model to generate synthetic neural signals (Fig. 1e). Using human blood-oxygen-level-dependent (BOLD) data, FC calculated from the synthetic signals (model FC) was compared to FC derived from empirical BOLD signals (empirical FC), averaged across 800 participants from the Human Connectome Project (HCP) dataset. The model FC and empirical FC showed a strong correlation ($r = 0.97$), indicating that the ANN effectively captures the complex inter-regional relationships crucial for EC inference (Fig. 1f). This suggests that the trained ANN can serve as a reliable surrogate brain for virtual perturbations. In this study, the ANN was implemented as a multilayer perceptron (MLP; Extended Data Fig. 1), but the NPI framework is flexible and supports various predictive models as long as they are capable of learning brain dynamics and capturing inter-regional relationships (Supplementary Fig. 1 and Supplementary Notes 1 and 2). In addition to the MLP, we tested alternative surrogate models, including convolutional neural network, recurrent neural network (RNN) and vector autoregressive models. These models were evaluated for their performance in signal prediction, FC reproduction and EC inference (Supplementary Table 1), confirming that the NPI framework is robust across different ANN architectures.

After training, the ANN is fixed and serves as a surrogate model for the brain. Virtual perturbations are systematically applied to each node of the ANN, with each node representing a specific brain region (Fig. 1c). Perturbations are introduced as impulse increases in the signal at the selected node at time t (Fig. 1g). The ANN then processes both perturbed and baseline inputs to predict subsequent neural activities $\mathbf{x}(t+1)$ for comparison. Differences in the predicted responses of target regions between perturbed and baseline inputs reflect the EC from the source (perturbed) region to the target region. Increased or decreased activity in the target regions indicates excitatory or inhibitory EC, respectively (Fig. 1h). This one-to-all EC mapping is achieved by perturbing a single node, and systematic perturbations across all nodes provide a comprehensive all-to-all EC mapping (Fig. 1d), capturing the directionality, strength and excitatory/inhibitory properties of causal interactions among brain regions. Mathematically, this process can be interpreted as deriving the Jacobian matrix of the trained ANN (Supplementary Note 4 and Extended Data Fig. 2), quantifying how a small input to one node influences the subsequent states of other nodes.

The effectiveness of NPI was validated using data generated by ground-truth generative models with established EC (Fig. 1i). When applied to real resting-state functional magnetic resonance imaging (fMRI) datasets, NPI successfully revealed seed-based EC and whole-brain EC, uncovering the distribution of EC both within and across functional brain networks (Fig. 1j).

Validation of NPI on generative models

We validated NPI using synthetic datasets generated by models with established ground-truth EC. Three simulated datasets were utilized: one from RNN models (Fig. 2a–h) and two from generative models of fMRI (Fig. 2i–p). Ground-truth EC was derived using the 'perturb and record' method directly applied to the generative models. The performance of NPI was then assessed by comparing the inferred EC with the ground-truth EC.

First, NPI was applied to an RNN model with a predefined weight matrix serving as the SC, where the matrix entries were drawn from Gaussian distributions centered at zero (Fig. 2a). Neural signals were synthesized by executing the RNN (Fig. 2b). Following the NPI framework, a surrogate ANN was trained to model the RNN-generated signals. To evaluate the ANN's ability to learn the RNN dynamics, its output was recursively fed back into the system to generate ANN-simulated signals (Fig. 2c). FC derived from these ANN-generated signals showed a strong correlation with the FC directly calculated from the RNN-generated signals, confirming the ANN's proficiency in capturing inter-regional relationships in the RNN (Fig. 2d).

To infer EC, perturbations were systematically applied to the trained ANN (Fig. 2e and Supplementary Figs. 2 and 3). The RNN's intrinsic EC, obtained by directly perturbing the ground-truth RNN, served as the ground-truth EC (Fig. 2f). Comparisons of NPI-inferred EC with ground-truth EC revealed a high correlation $r = 0.95$, outperforming GC method (Fig. 2g, Supplementary Fig. 4 and Supplementary Note 5). NPI-inferred EC demonstrated a strong correlation with the SC of the RNN, which provides the anatomical foundation for EC (Supplementary Fig. 4). While EC does not perfectly align with SC due to the nonlinearity of neural dynamics and signal noise, the correlation between EC and SC was substantially stronger than that

https://doi.org/10.1038/s41592-025-02654-x

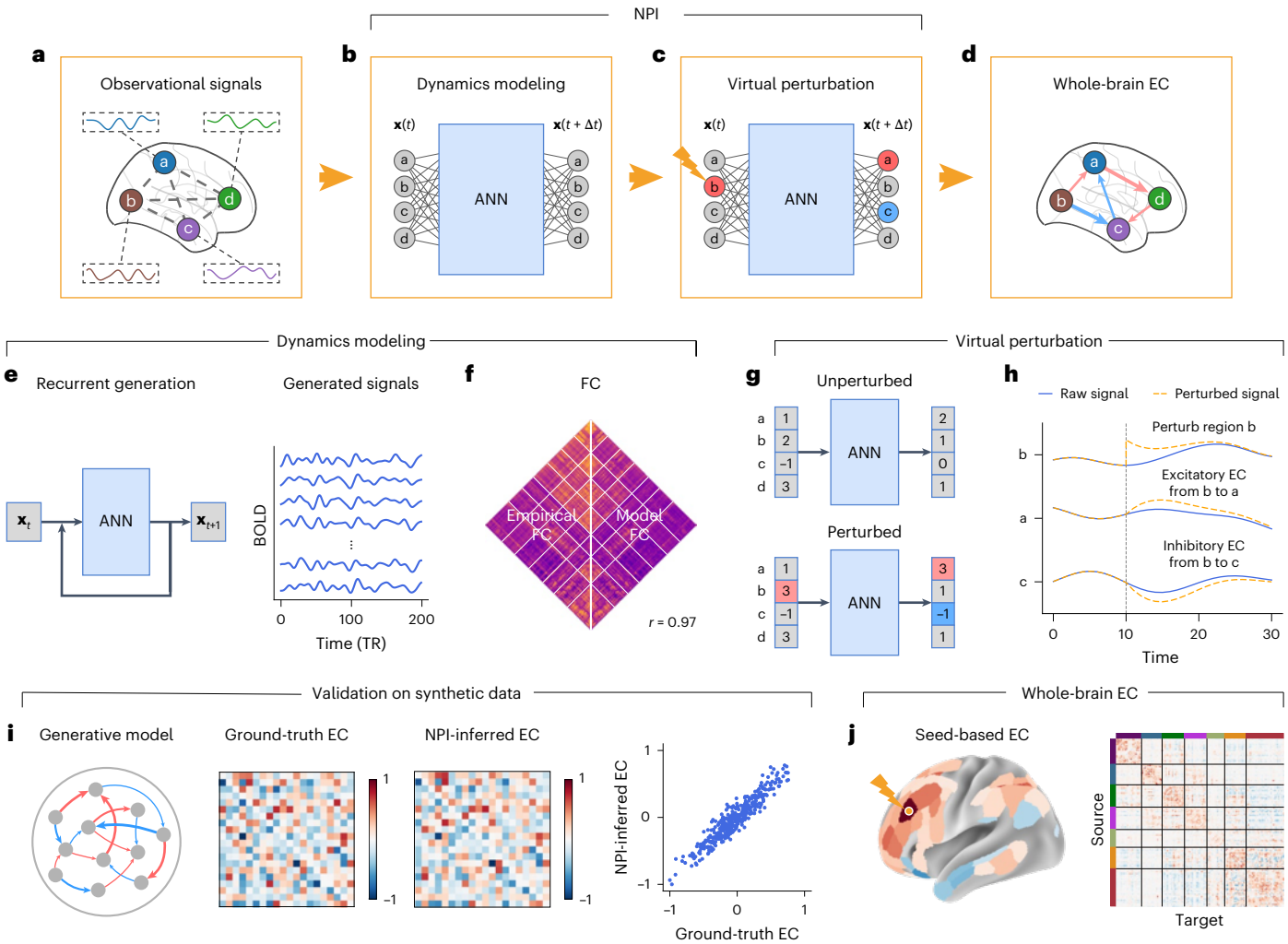


Fig. 1 | Neural Perturbational Inference maps effective connectivity by virtually perturbing a surrogate brain. a, Schematic representation of a brain network with recorded neural signals for each region, from which EC values between regions are inferred. **b**, A surrogate brain, implemented as an ANN, is trained to model brain dynamics. The ANN is optimized to predict subsequent brain states based on previous states. **c**, After training, the ANN is systematically perturbed to infer EC. Perturbing one region yields one-to-all EC values by measuring the perturbation-induced responses in other regions. **d**, All-to-all EC is mapped by perturbing each ANN region systematically. The resulting brain-wide EC map represents causal influences across the brain, capturing directionality, strength and excitatory/inhibitory distinctions. **e**, Recurrently feeding the result of prediction back as input to ANN yields the ANN-generated neural signals. **f**, FC is calculated from generated BOLD signals (model FC) and empirical BOLD signals (empirical FC), both averaged across 800 participants. Model FC and empirical FC exhibit a strong correlation ($r = 0.97$). **g**, Perturbations are applied as increases in neural signals at selected regions. Differences in predicted target region responses, compared to baseline inputs, reflect the EC from the source to the target regions. Perturbation effects are color-coded; red indicates increased neural signals and blue indicates decreased signals. **h**, Perturbing region b increases activity in region a and decreases activity in region c, indicating an excitatory EC from b to a and an inhibitory EC from b to c. **i**, Validation of NPI using generative models with known ground-truth EC. **j**, Application of NPI to resting-state fMRI data yields whole-brain EC maps.

between FC and SC (Supplementary Table 2). This distinction highlights FC's limitations, such as a lack of directionality and susceptibility to spurious connectivity[22].

We further assessed the robustness of NPI against variations in perturbation intensity, system noise levels in the ground-truth RNN, data lengths and RNN sizes (Fig. 2h). The results demonstrated that NPI's performance remained stable across different perturbation magnitudes and experienced only a slight decline with increasing noise levels. Additionally, larger datasets were found to be essential for reliable EC inference, particularly in larger networks. These findings underscore the robustness and scalability of the NPI framework.

To validate NPI in a real-world application scenario, we applied it to two synthetic fMRI datasets. The first application utilized a publicly available synthetic dataset containing BOLD dynamics generated from nine distinct SC configurations[23] (Fig. 2i and Extended Data Fig. 3a). This dataset, commonly used for validating EC inference algorithms, features binary SC and simulates neural firing rates, which are converted into BOLD signals through a hemodynamic response function. As ground-truth EC is unavailable for this dataset, we assessed EC inference performance by calculating the area under the receiver operating characteristic curve (AUC) for classifying the presence or absence of each SC connection after binarizing the NPI-inferred EC. NPI achieved an AUC near 1, outperforming both GC and DCM (Fig. 2j). Across all nine SC configurations, NPI consistently demonstrated superior performance compared to GC and DCM (Extended Data Fig. 3), underscoring its precision and reliability in mapping EC across diverse connection topologies and model structures.

Inferring EC from large-scale networks poses challenges for conventional methods like DCM. To evaluate NPI's effectiveness in large-scale EC inference, we applied it to synthetic BOLD data generated

https://doi.org/10.1038/s41592-025-02654-x

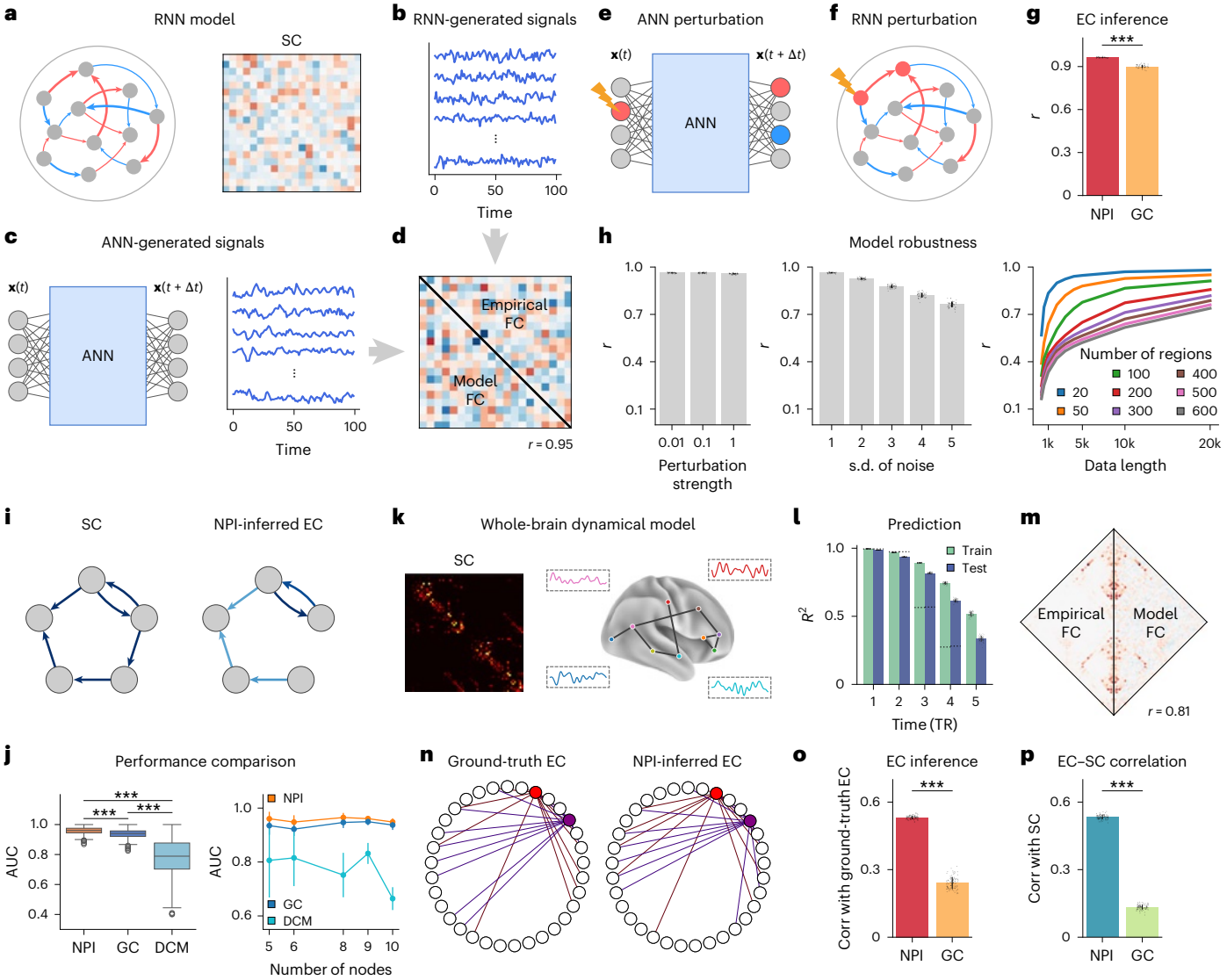


Fig. 2 | Validation of NPI on generative models. a–**h**, Validation using an RNN model. SC (weight matrix) of the RNN model (**a**). RNN-generated signals (**b**). Recursively running the trained ANN yields ANN-generated signals (**c**). Empirical FC and model FC ($r = 0.95$) (**d**). NPI infers EC by perturbing trained ANN (**e**). Ground-truth EC derived by perturbing the ground-truth RNN (**f**). NPI-inferred EC shows a strong correlation with ground-truth EC ($r = 0.95$) and outperforms GC-inferred EC ($P = 1.78 \times 10^{-15}$, two-sided Wilcoxon signed-rank test, $n = 50$) (**g**). Unless otherwise stated, bars and lines indicate mean values, error bars represent s.d. and *** denotes $P < 0.001$. Robustness of NPI under varying conditions: perturbation magnitudes ($n = 50$ for each bar), s.d. of noise ($n = 50$ for each bar) and data lengths (averaged across ten samples for each point) (**h**). **i**–**p**, Validation using synthetic fMRI data. Example of NPI-inferred EC compared to SC (**i**). NPI outperforms GC and DCM in classifying the existence of SC connections ($P < 10^{-60}$ for three comparisons, two-sided Wilcoxon signed-rank test, $n = 540$ for each method) (left, **j**). Unless otherwise stated, boxes indicate the interquartile range (IQR; Q1–Q3), centers represent medians and whiskers extend to the minimum and maximum values. The advantage holds across different node numbers (SC with 5, 6, 8, 9 and 10 nodes: $n = 240, 60, 120, 60$ and 60, respectively) (right, **j**). fMRI signals simulated by a WBM based on real human SC (**k**). The prediction performance of the ANN ($n = 100$ for each bar) (**l**). The dotted line represents the performance of a univariate auto-regression model. The empirical FC and model FC ($r = 0.81$) (**m**). NPI-inferred EC closely resembles ground-truth EC. The strongest 40% output EC from two median-performing nodes are illustrated (**n**). NPI outperforms GC in both capturing EC (**o**, $P = 3.90 \times 10^{-18}$) and reflecting SC (**p**, $P = 3.90 \times 10^{-18}$) (**o**,**p**). Two-sided Wilcoxon signed-rank test, $n = 100$ for both panels. Corr, correlation.

from a whole-brain model (WBM) comprising 66 nodes. The SC matrix for the WBM was derived from neuroanatomical data obtained via diffusion spectrum imaging (DSI) and BOLD time series were simulated using a neurodynamic model (Fig. 2k). Although multistep prediction accuracy declined slightly (Fig. 2l and Supplementary Fig. 7), the FC of ANN-generated signals remained strongly correlated with the FC of WBM-simulated signals (Fig. 2m), confirming the ANN's ability to effectively capture inter-regional relationships. Visualizing the strongest 40% of output EC from two median-performing nodes revealed that NPI-inferred EC closely aligned with ground-truth EC (Fig. 2n). Moreover, NPI-inferred EC more accurately reflected both the ground-truth EC and SC compared to EC inferred using GC (Fig. 2o,p and Supplementary Figs. 5 and 6). These results establish NPI as a robust and reliable method for estimating EC in large-scale brain networks. On this dataset, we further tested the performance of different surrogate models and found that the MLP outperformed other architectures in both FC reproduction and EC inference (Supplementary Tables 1 and 2). We thus selected the MLP as the surrogate model for EC inference for real data analysis.

Human EBC inferred by NPI

We applied NPI to resting-state fMRI (rs-fMRI) data from 800 participants in the HCP dataset, parcellated into 360 regions using the Multi-Modal Parcellation (MMP) atlas[24,25] (Supplementary Table 5).

An individualized ANN was trained for each participant using their rs-fMRI data (Supplementary Video). Testing different input configurations revealed that using signals from the previous three time steps to predict the next time step outperformed using only the previous time step (Supplementary Fig. 8). Consequently, we adopted the three-step-input MLP model for subsequent analyses. The trained ANN functions as an individualized surrogate model. At the group level, the FC calculated from real BOLD signals (empirical FC) strongly correlates and exhibits similar spatial patterns with the FC derived from ANN-generated BOLD signals (model FC; $r = 0.97$; Fig. 3c,d), suggesting that the ANN captures the complex inter-regional interactions in the real brain.

After training the surrogate models, we systematically perturbed each individualized model to derive whole-brain EC, referred to as the effective brain connectome (EBC). Here we perturbed a node by increasing its activity. To adapt NPI to various neuroimaging modalities, ANN architectures and virtual perturbation protocols need to be tailored (Supplementary Notes 1–3 and Supplementary Figs. 13 and 14). We first obtained the individualized EBC by perturbing the individualized surrogate model and then calculated the group-level EBC by averaging the EBC across 800 participants (Fig. 3a and Extended Data Fig. 4). Positive entries in the EBC represent excitatory EC, whereas negative entries indicate inhibitory EC. Brain regions were categorized into seven functional networks according to Yeo et al.[26] (Supplementary Table 3 and Fig. 3b). Seed-based EC analysis revealed the topographic organization of functional networks, showing structural similarities to networks defined by FC. Compared to FC, EC better captures the inhibitory influence of seed regions on other brain areas (Fig. 3e).

The distribution of EC strengths exhibited a long-tail property, with most connections having near-zero strengths and a small fraction showing large strengths. Fitting these distributions to four hypothesized models (log-normal, normal, exponential and inverse Gaussian) revealed that the log-normal distribution best described both excitatory and inhibitory EC, as determined by the Akaike information criterion (Fig. 3f,g and Supplementary Table 4). This distribution aligns with SC distributions observed in experimental studies using tract-tracing techniques in mice and macaques[27,28]. The log-normal distribution of EC strengths was reproducible under the Automated Anatomical Labeling (AAL) parcellation (Supplementary Fig. 9). Excitatory EC was found to have stronger maximum strengths compared to inhibitory EC. When scaled such that the maximum excitatory strength equals 1, the maximum inhibitory strength was 0.16. The strongest excitatory EC primarily comprised intra-network connections (Fig. 3f and Supplementary Fig. 10). In contrast, the strongest inhibitory EC predominantly involved inter-network connections and were mostly inter-hemisphere (Fig. 3g and Supplementary Fig. 10). Nodes with the largest averaged in–out degrees were dispersed across the cortex and spanned multiple functional networks (Fig. 3f). Node degree, defined as the number of connections a node has, serves as a measure of its centrality or importance within the network. Here, we binarized it by applying an 80% threshold on absolute EC strengths (0.06). Connections with absolute strengths below this threshold were set to 0, while those above were set to 1. In this binarized EBC, excitatory and inhibitory EC were not differentiated. As EC is directed and thus asymmetric, the in-degree of a node differs from its out-degree. Overall, 86% of connections in the binarized EBC were bidirectional, consistent with previous findings on SC[29].

NPI-inferred EC is robust and aligns with structural connectivity

To evaluate the reliability and scalability of EBC inferred from real fMRI data, we applied NPI to Schaefer atlases with increasing numbers of brain regions, ranging from 100 to 1,000 regions[30] (Fig. 4a,b and Extended Data Fig. 5). As the number of regions increased, prediction performance slightly declined, likely due to the increased data requirements for training NPI on larger networks (Fig. 2h); however, FC reproduction performance remained stable across different atlas resolutions (Fig. 4b). Additionally, the patterns of inferred EBC were consistent, and the variability of EC was comparable to that of FC, demonstrating the robustness of NPI (Extended Data Fig. 5b,c).

We further examined inter-participant variability and test–retest reliability across ANN trainings, sessions, and datasets (Fig. 4c–e). Inter-participant variability in both within-network and cross-network EC was comparable (Fig. 4c). Across all EC pairs, 55% of connections were significantly different from zero across 800 participants, reflecting a consistent deviation from the null hypothesis of no connection ($P < 0.05$, one-sample t-test, Bonferroni corrected; Supplementary Fig. 11). To assess whether NPI-mapped EC depends on the variability and initialization of ANN training, we trained two ANNs with different initializations for each HCP participant and compared the inferred ECs. The high consistency (termed 'ANN trainings' in Fig. 4d) confirmed the robustness of NPI to variations in ANN training. To distinguish intrinsic individual variability from potential noise introduced by the method, we conducted cross-session, inter-participant and inter-dataset assessments (termed as 'Sessions', 'Participants' and 'Datasets' in Fig. 4d and Supplementary Fig. 15). In cross-session analyses, we applied NPI separately to the first two and last two sessions of each participant's four-session fMRI data and found that cross-session EC correlations were higher than inter-participant correlations. This indicates that NPI captures stable, participant-specific EBC patterns across sessions. The limbic network exhibited the lowest reliability, likely due to the low signal-to-noise ratio of fMRI in this region[26,31]. In cross-dataset analyses, we applied NPI to the Adolescent Brain Cognitive Development (ABCD) dataset[32] and observed strong alignment between population-averaged EBCs from the HCP and ABCD datasets, confirming NPI's generalizability and applicability across datasets (Fig. 4e).

We also investigated the relationship between EBC and its structural foundation derived from DSI data. Our analysis revealed a strong correlation between EBC and SC, confirming that the brain's anatomical structure strongly influences the pathways of functional neural communication (Fig. 4f). Overall, these results demonstrate that NPI reliably captures general EBC patterns across datasets while effectively characterizing individual brain EBCs.

NPI supports clinical applications

To evaluate NPI's potential for clinical applications, we assessed the consistency between the spatial distribution of NPI-inferred EBC and neurostimulation-induced neural responses. We utilized an open-source cortico-cortical evoked potentials (CCEP) dataset (Fig. 5a) from the Functional Tractography (F-TRACT) project[33], which includes intracortical stimulation and intracerebral stereoencephalographic recordings from epileptic patients (Fig. 5b). This dataset aggregates data from a cohort of 613 patients with stimulation sites distributed across various brain regions, resulting in a comprehensive group-level CCEP connectivity matrix of the human brain. This matrix maps neural signal propagation across the cortex, providing a direct measurement of neural connectivity that is ideal for validating NPI-inferred EBC.

Comparisons between the EBC and the CCEP-derived connectivity matrix (Fig. 5c) revealed a strong correlation between NPI-inferred EC and CCEP (whole-brain, $r = 0.33$), notably higher than the correlation between FC and CCEP (whole-brain, $r = 0.20$; Fig. 5d). These findings demonstrate that NPI-inferred EBC, derived from rs-fMRI data, accurately reflects neurostimulation propagation pathways and the thus underlying causal relationships between brain regions.

To further illustrate the potential of EBC in guiding neurostimulation, we analyzed both output and input EC patterns in the CCEP and NPI-inferred EBC matrices (Fig. 5e). Output EC, represented by a row in the EBC matrix, reflects the propagation range resulting from stimulation of a specific brain region (the source). Conversely, input EC, represented by a column in the EBC matrix, identifies regions capable

https://doi.org/10.1038/s41592-025-02654-x

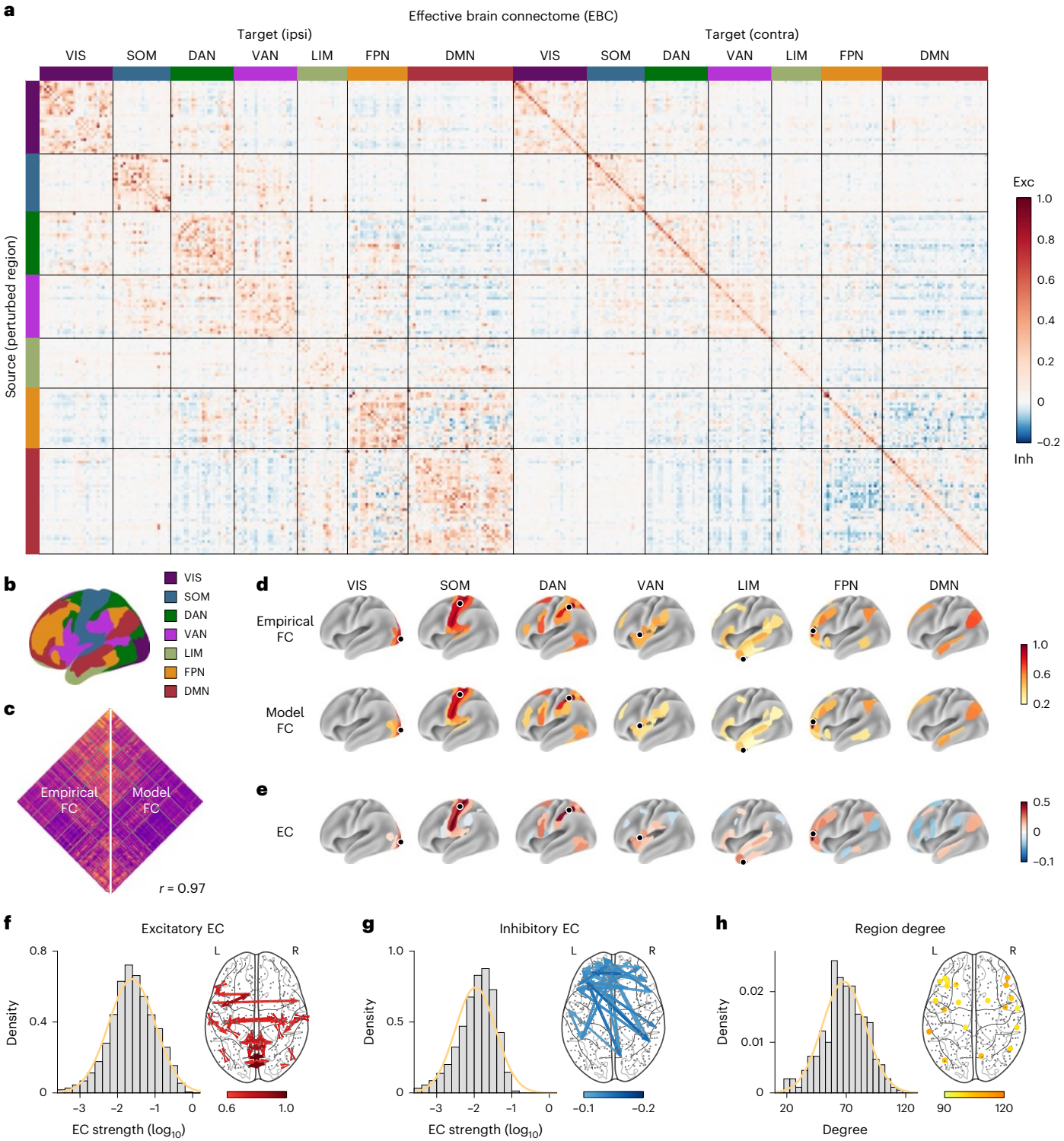


Fig. 3 | Human EBC inferred by NPI. a, Group-averaged EBC across 800 participants, with regions organized by functional networks. Each row represents EC from a source region in the left hemisphere to the left (ipsi) and right (contra) hemisphere, scaled to a maximum response of 1.0. Stronger red and blue colors represent stronger excitatory (Exc) and inhibitory (Inh) connections, respectively. **b**, Cortical regions assigned to seven resting-state functional networks. **c**, Positive correlation ($r = 0.97$) between FC derived from model-generated BOLD signals (model FC) and empirical BOLD signals (empirical FC). **d**, Seed-based FC maps from empirical BOLD data (empirical FC) and ANN-generated BOLD data (model FC). Seeds are located in each of the seven resting-state networks, indicated by a black dot on each map (except for DMN, where the seed is within the brain). Top 10% FC strengths within each resting-state network are plotted. **e**, NPI-inferred EC maps with seeds in each of the seven resting-state networks. Excitatory EC values are shown in red, and inhibitory EC values are shown in blue. Top 10% EC strengths are plotted. **f**, Left: excitatory EC strength follows a log-normal distribution, as shown by the Gaussian fit of log-transformed EC values. Right: the 50 strongest excitatory EC connections. **g**, The same analysis as **f** for inhibitory EC. **h**, Left: degree distribution of brain regions based on EC binarized using an 80% strength threshold. The degree of a region is calculated as the average of its in-degree and out-degree. Right: the 30 brain regions with the largest degrees after binarizing the EBC. VIS, visual network; SOM, somatomotor network; DAN, dorsal attention network; VAN, ventral attention network; LIM, limbic network; FPN, frontoparietal network; DMN, default mode network.

https://doi.org/10.1038/s41592-025-02654-x

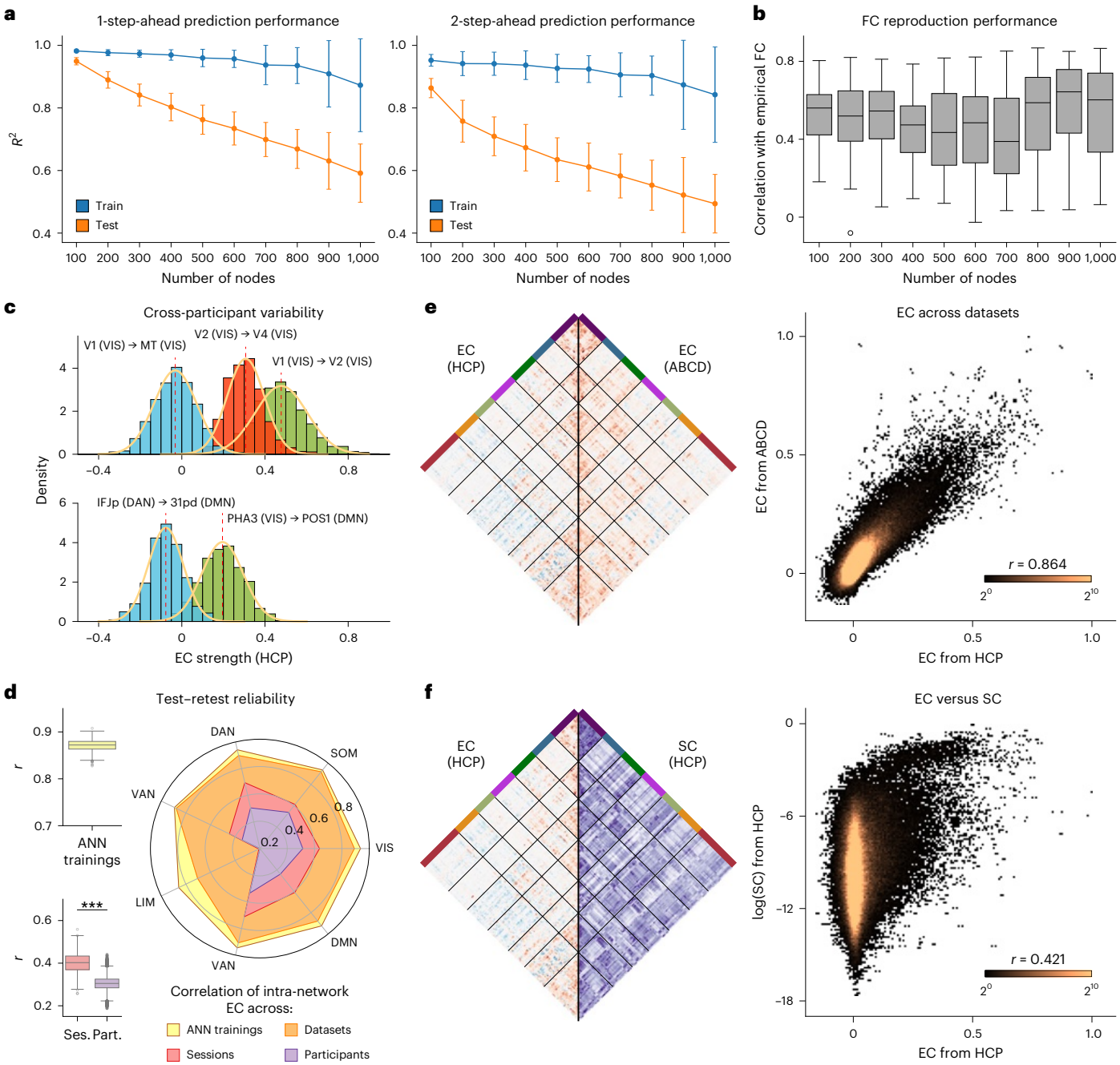


Fig. 4 | NPI-inferred EC is robust and aligns with structural connectivity. **a**, Prediction performance (R^2) for one-step-ahead (left) and two-step-ahead (right) predictions on training (blue) and test (orange) datasets, as the number of brain regions increases ($n = 100$). **b**, FC reproduction performance, measured as the correlation between model FC and empirical FC, across increasing numbers of brain regions ($n = 100$). The central line represents the median performance, and the bounds indicate the 25th and 75th percentiles. **c**, Cross-participant variability in EC strengths from 800 participants. Within-network EC variations, exemplified by EC connections within the visual network (VIS), including V1 to V2 (green), V2 to V4 (red) and V1 to MT (blue) (top). Cross-network EC variations, showing a positive EC from PHA3 in VIS to POS1 in DMN (green) and a negative EC from IFJp in DAN to 31pd in DMN (blue) (bottom). **d**, Test–retest reliability of NPI-inferred EC. Left: whole-brain test–retest reliability. Right: network-level test–retest reliability. Reliability was assessed across ANN trainings, sessions, participants and datasets. 'ANN trainings' refers to running NPI twice on individual data ($n = 800$). 'Sessions' (Ses.) refers to splitting and training on half of the individual data ($n = 800$). 'Participants' (Part.) refers to cross-participant EC correlation from 800 participants in the HCP dataset ($n = 319{,}600$). Inter-session reliability is significantly higher than inter-participant reliability ($P < 10^{-60}$, Mann–Whitney U-test). 'Datasets' refers to the consistency between participant-averaged EC inferred from the HCP and the ABCD datasets ($n = 1$). **e**, Left: EC inferred from the HCP dataset and ABCD dataset. Right: quantitative relationship between them ($r = 0.864$). **f**, Left: NPI-inferred EC compared to log-transformed SC derived from DSI of the HCP dataset. Right: quantitative relationship between them ($r = 0.421$).

of propagating stimulation to a given area (the target). We specifically examined output EC with the dorsolateral prefrontal cortex (dlPFC) as the source and input EC with the posterior cingulate cortex (PCC) as the target, as these regions are frequently studied in neuromodulation research (Fig. 5f,g). Results demonstrated that NPI-inferred EBC accurately captured both output and input patterns, showing stronger correlations with CCEP-derived output and input connectivity than FC.

https://doi.org/10.1038/s41592-025-02654-x

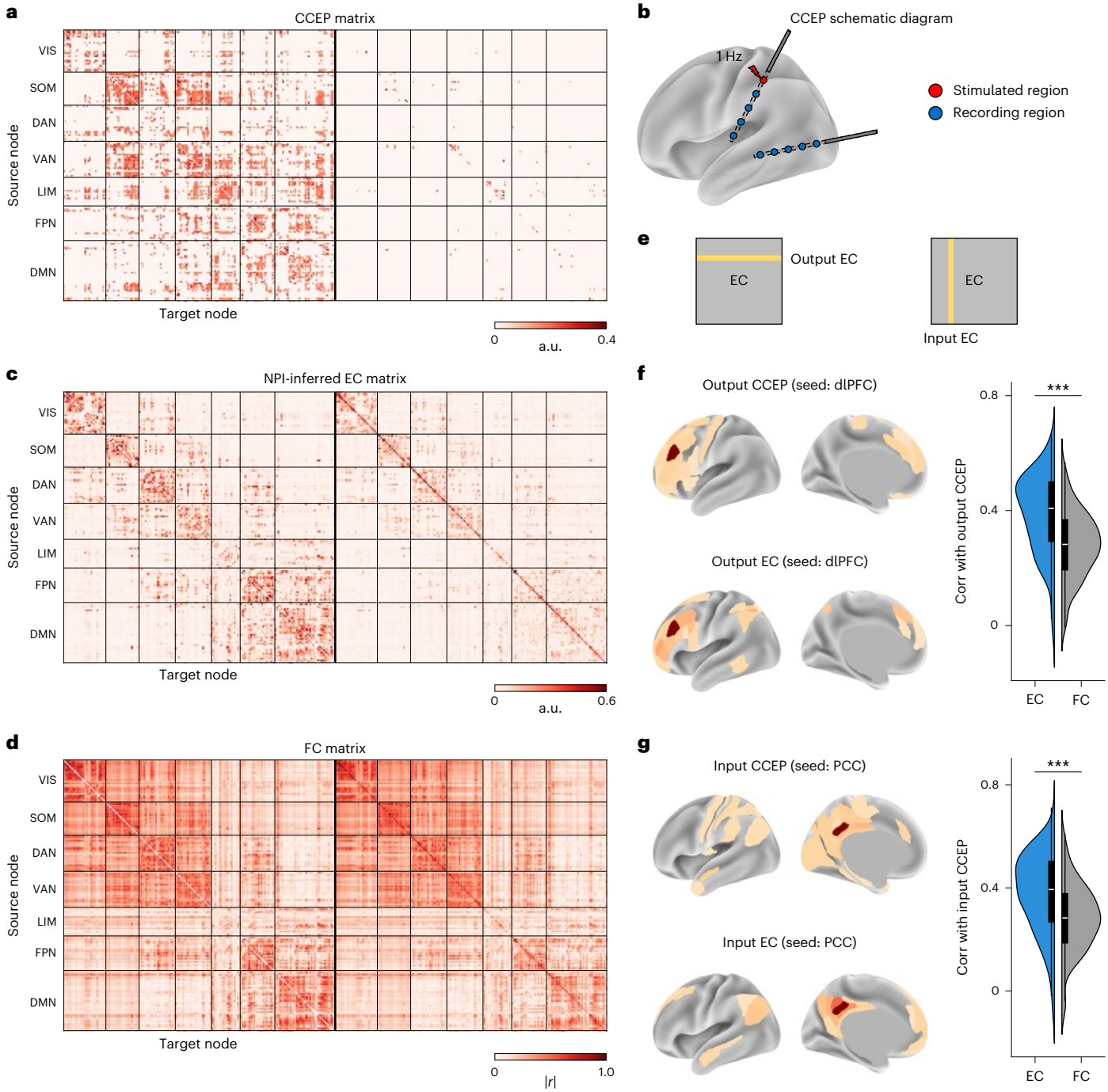


Fig. 5 | Validating EBC with cortico-cortical evoked potentials. a, Group-level CCEP matrix from the F-TRACT project involving 613 patients, depicting evoked responses from the left hemisphere to the entire brain. **b**, Schematic representation of the CCEP experimental setup, showing invasive stimulation and recording locations. **c**,**d**, Matrices of group-level NPI-inferred EC (**c**) and empirical FC (**d**), respectively, from the HCP dataset, each involving 800 participants and organized in order in congruence with the CCEP matrix. **e**, Left: a row in the CCEP or NPI-inferred EC matrix represents output CCEP or EC, indicating connections from a source region to all other regions. Right: a column in the CCEP or NPI-inferred EC matrix represents input CCEP or EC, reflecting connections from all other regions to a target region. **f**, Left: output CCEP and output EC from the dlPFC show high similarity, as evidenced by overlapping distributions. Right: correlations between all rows of the EC and CCEP matrices are significantly higher than those between rows of the FC and CCEP matrices (left hemisphere, $P = 4.55 \times 10^{-24}$, two-sided Wilcoxon signed-rank test, $n = 180$). Violin plots include boxes representing the IQR (Q1–Q3), horizontal lines indicating medians (Q2) and whiskers extending to the minimum and maximum values. **g**, Similar analysis for input CCEP and input EC from the PCC. Left: overlapping distributions of input CCEP and input EC. Right: correlations between all columns of the EC, FC and CCEP matrices (left hemisphere, $P = 4.71 \times 10^{-22}$, two-sided Wilcoxon signed-rank test, $n = 180$).

Notably, the advantages of NPI-inferred EBC extend beyond those of CCEP. CCEP relies on invasive procedures involving electrical stimulation at a single site per patient, necessitating data aggregation across many individuals to construct a group-level connectivity map. In contrast, NPI offers a noninvasive, data-driven alternative. This makes NPI not only easier to implement but also more adaptable for widespread research and clinical applications.

To evaluate the potential of NPI-inferred participant-level EBC as a biomarker, we applied NPI to fMRI data from the Autism Brain Imaging Data Exchange (ABIDE) dataset[34] and the Alzheimer's Disease

Neuroimaging Initiative (ADNI) dataset[35] (Supplementary Note 6 and Supplementary Fig. 12). Our findings showed that EC performed comparably to FC in classifying healthy individuals versus patients with brain disorders, suggesting that NPI-inferred EC could serve as a viable alternative to FC as a biomarker. Moreover, the directionality inherent in EC offers additional insights, holding promise for guiding personalized treatment strategies.

Discussion

The concept of EC is fundamental in neuroscience but varies across methodologies[11,36,37]. For example, GC views EC as the predictive influence of one brain region over another, whereas DCM frames it through coupling coefficients within a state-space model. NPI adopts a 'perturb and record' approach that aligns with the statistical notion of causality: a perturbation in one variable that substantially alters another indicates a causal link[38]. Such a definition is congruent with empirical methods such as optogenetics, where direct regional perturbations are applied and the resultant neural responses are observed to confirm causal interactions[8,10,39].

NPI offers several distinct advantages over traditional methodologies for deriving EC. First, NPI enables noninvasive mapping of EC, a stark contrast to conventional approaches that often require invasive procedures, thereby expanding the applicability to a broader range of participants[10]. Second, unlike traditional computational methods, NPI employs ANNs to directly learn the complex, nonlinear dynamics of brain activity from data. By avoiding predefined model structures or assumptions about neural mechanisms, NPI effectively handles diverse data types and dynamics that parametric models may struggle to capture[17]. The flexibility of ANNs within the NPI framework further allows for advanced machine-learning techniques, such as pre-training to construct group-level surrogate models and fine-tuning to develop individual-level models[13,40]. Finally, NPI's versatility extends to accommodating various forms and scales of perturbations, once the surrogate model is adequately trained. This adaptability, combined with the efficiency of ANNs in processing large fMRI datasets with numerous brain nodes, enhances NPI's practicality across diverse experimental and clinical settings.

Although this study primarily employs NPI with rs-fMRI data and simple impulse perturbations, the framework's versatility extends far beyond this initial application. Potential applications of NPI range from analyzing single-neuron activity to large-scale neuroimaging data, such as electroencephalography and fMRI. NPI's ability to integrate EC findings across these diverse scales not only enhances our understanding of the brain's structural-functional interplay but also holds promise for uncovering the neural mechanisms underlying complex cognitive processes.

NPI holds substantial promise for therapeutic applications. First, EC maps inferred by NPI can serve as biomarkers for neurological disorders, offering mechanistic insights by comparing EC patterns between patients and healthy controls. Additionally, NPI enhances the precision of neurostimulation treatments by generating personalized EC maps[41,42]. The alignment of NPI-inferred EBC with CCEP patterns underscores its potential utility in guiding personalized neurostimulation strategies. Moreover, NPI's ability to model the effects of stimulating multiple regions or adjusting stimulation parameters provides a robust framework for optimizing neurostimulation approaches. This capability enables the customization of interventions based on individual brain connectivity profiles, potentially improving therapeutic outcomes.

As a data-driven approach, NPI leverages the predictive power of ANNs to infer EC but faces the common challenge of requiring large volumes of high-quality data. A critical future direction involves developing surrogate brain models that maintain high predictive accuracy, while reducing data demands. This could include exploring advanced ANN architectures or incorporating domain-specific knowledge to enhance model performance. Beyond EC inference, applying varied interventions to trained surrogate ANN models presents an exciting opportunity to deepen our understanding of brain dynamics. Such analyses could uncover novel insights into brain function, paving the way for innovative therapeutic and research applications.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-025-02654-x.

References

1. Park, H.-J. & Friston, K. Structural and functional brain networks: from connections to cognition. *Science* **342**, 1238411 (2013).
2. Deco, G., Vidaurre, D. & Kringelbach, M. L. Revisiting the global workspace orchestrating the hierarchical organization of the human brain. *Nat. Human Behav.* **5**, 497–511 (2021).
3. Seguin, C., Sporns, O. & Zalesky, A. Brain network communication: concepts, models and applications. *Nat. Rev. Neurosci.* **24**, 557–574 (2023).
4. Yeh, C.-H., Jones, D. K., Liang, X., Descoteaux, M. & Connelly, A. Mapping structural connectivity using diffusion mri: challenges and opportunities. *J. Magn. Reson. Imaging* **53**, 1666–1682 (2021).
5. van den Heuvel, M. P. & Hulshoff Pol, H. E. Exploring the brain network: a review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* **20**, 519–534 (2010).
6. Schippers, M. B., Roebroeck, A., Renken, R., Nanetti, L. & Keysers, C. Mapping the information flow from one brain to another during gestural communication. *Proc. Natl Acad. Sci. USA* **107**, 9388–9393 (2010).
7. Manjunatha, K. K. H. et al. Controlling target brain regions by optimal selection of input nodes. *PLoS Comput. Biol.* **20**, e1011274 (2024).
8. Kim, S. et al. Whole-brain mapping of effective connectivity by fMRI with cortex-wide patterned optogenetics. *Neuron* **111**, 1732–1747 (2023).
9. Randi, F., Sharma, A. K., Dvali, S. & Leifer, A. M. Neural signal propagation atlas of *Caenorhabditis elegans*. *Nature* **623**, 406–414 (2023).
10. Hollunder, B. et al. Mapping dysfunctional circuits in the frontal cortex using deep brain stimulation. *Nat. Neurosci.* **27**, 573–586 (2024).
11. Friston, K. J., Kahan, J., Biswal, B. & Razi, A. A DCM for resting state fMRI. *NeuroImage* **94**, 396–407 (2014).
12. Li, S., Xiao, Y., Zhou, D. & Cai, D. Causal inference in nonlinear systems: Granger causality versus time-delayed mutual information. *Phys. Rev. E* https://doi.org/10.1103/PhysRevE.97.052216 (2018).
13. Liang, Z., Luo, Z., Liu, K., Qiu, J. & Liu, Q. Online learning Koopman operator for closed-loop electrical neurostimulation in epilepsy. *IEEE J. Biomed. Health Inform.* **27**, 492–503 (2022).
14. Abrol, A. et al. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat. Commun.* **12**, 353 (2021).
15. Perich, M. G. et al. Inferring brain-wide interactions using data-constrained recurrent neural network models. Preprint at *bioRxiv* https://doi.org/10.1101/2020.12.18.423348 (2021).
16. Tu, T., Paisley, J., Haufe, S. & Sajda, P. A state-space model for inferring effective connectivity of latent neural dynamics from simultaneous eeg/fmri. *Adv. Neural Inform. Proc. Sys.* **32**, 4662–4671 (2019).

17. Das, A. & Fiete, I. R. Systematic errors in connectivity inferred from activity in strongly recurrent networks. *Nat. Neurosci.* **23**, 1286–1296 (2020).
18. Ivanovs, M., Kadikis, R. & Ozols, K. Perturbation-based methods for explaining deep neural networks: a survey. *Pattern Recog. Lett.* **150**, 228–234 (2021).
19. Dong, M. et al. Causal identification of single-cell experimental perturbation effects with cinema-ot. *Nat. Methods* **20**, 1769–1779 (2023).
20. Veit, M. J. et al. Temporal order of signal propagation within and across intrinsic brain networks. *Proc. Natl Acad. Sci. USA* **118**, e2105031118 (2021).
21. Ozdemir, R. A. et al. Individualized perturbation of the human connectome reveals reproducible biomarkers of network dynamics relevant to cognition. *Proc. Natl Acad. Sci. USA* **117**, 8115–8125 (2020).
22. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage* **59**, 2142–2154 (2012).
23. Sanchez-Romero, R. et al. Estimating feedforward and feedback effective connections from fMRI time series: assessments of statistical methods. *Netw. Neurosci.* **3**, 274–306 (2019).
24. Van Essen, D. C. et al. The wu-minn Human Connectome Project: an overview. *NeuroImage* **80**, 62–79 (2013).
25. Glasser, M. F. et al. A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
26. Thomas Yeo, B. T. et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
27. Oh, S. W. et al. A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).
28. Markov, N. T. et al. A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cereb. Cortex* **24**, 17–36 (2014).
29. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
30. Schaefer, A. et al. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cereb. Cortex* **28**, 3095–3114 (2018).
31. Liu, M., Liu, X., Hildebrandt, A. & Zhou, C. Individual cortical entropy profile: test–retest reliability, predictive power for cognitive ability, and neuroanatomical foundation. *Cereb. Cortex Commun.* **1**, tgaa015 (2020).
32. Saragosa-Harris, N. M. et al. A practical guide for researchers and reviewers using the abcd study and other large longitudinal datasets. *Dev. Cogn. Neurosci.* **55**, 101115 (2022).
33. Lemaréchal, J.-D. et al. A brain atlas of axonal and synaptic delays based on modelling of cortico-cortical evoked potentials. *Brain* **145**, 1653–1667 (2022).
34. Di Martino, A. et al. The Autism Brain Imaging Data Exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry* **19**, 659–667 (2014).
35. Petersen, R. C. et al. Alzheimer's Disease Neuroimaging Initiative (ADNI) clinical characterization. *Neurology* **74**, 201–209 (2010).
36. Barnett, L. & Seth, A. K. The MVGC multivariate granger causality toolbox: a new approach to Granger-causal inference. *J. Neurosci. Methods* **223**, 50–68 (2014).
37. Singh, M. F., Braver, T. S., Cole, M. W. & Ching, S. Estimation and validation of individualized dynamic brain models with resting state fmri. *NeuroImage* **221**, 117046 (2020).
38. Pearl, J. *Causality* (Cambridge University Press, 2009).
39. Bernal-Casas, D., Lee, H. J., Weitz, A. J. & Lee, J. H. Studying brain circuit function with dynamic causal modeling for optogenetic fmri. *Neuron* **93**, 522–532.e5 (2017).
40. Li, D., Wei, C., Li, S., Zou, J. & Liu, Q. Visual decoding and reconstruction via EEG embeddings with guided diffusion. *Adv. Neural Inform. Proc. Sys.* **37**, 102822–102864 (2024).
41. Schuepbach, W. et al. Neurostimulation for parkinson's disease with early motor complications. *N. Engl. J.Med.* **368**, 610–622 (2013).
42. Scangos, K. W. et al. Closed-loop neuromodulation in an individual with treatment-resistant depression. *Nat. Med.* **27**, 1696–1700 (2021).

Methods

The NPI method

Training ANN as a surrogate brain. NPI employs an ANN to replicate the brain's neural dynamics. While various network architectures can be used, this study implements an MLP as the ANN $f(\cdot)$, designed to predict the neural state at the next time step based on the states from the three preceding steps (see Supplementary Note 2 for an alternative one-step-input ANN model). The brain's dynamical system is modeled as

$$\hat{\mathbf{x}}_{t+1} = f(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \theta), \tag{1}$$

where $\mathbf{x}_t$, $\mathbf{x}_{t-1}$ and $\mathbf{x}_{t-2}$ represent the neural states of various brain regions at times t, $t-1$ and $t-2$, respectively. The function f is the MLP with parameters θ, encompassing all trainable weights of the network. $\hat{\mathbf{x}}_{t+1}$ is the predicted neural state at time $t+1$.

For a dataset with N brain regions, the network architecture includes an input layer of size $3N$, two hidden layers with sizes $2N$ and $0.8N$, respectively, and an output layer of size N. This structure was optimized via grid search based on prediction performance on the test set (Extended Data Fig. 1).

The MLP is trained to minimize the one-step-ahead prediction error. Each training sample consists of inputs $\mathbf{x}_t$, $\mathbf{x}_{t-1}$ and $\mathbf{x}_{t-2}$ and output $\mathbf{x}_{t+1}$. The loss function $\mathcal{L}(\theta)$ is defined as the squared error between the predicted and actual next neural states:

$$\mathcal{L}(\theta) = \| f(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \theta) - \mathbf{x}_{t+1} \|_2^2. \tag{2}$$

Training was conducted over 60 epochs with a batch size of 100, using the Adam optimizer with a learning rate of 10^{-3}. The implementation was performed in PyTorch on an NVIDIA GeForce RTX 4080 GPU.

Perturbing the trained ANN to infer EC. The target EC is an $n \times n$ matrix, where the ith row of the EC matrix represents the output EC from region i to all other regions, and the jth column represents the input EC from all other regions to region j.

To infer EC from a specific region i to all other regions, a perturbation is applied to the input on the ith node of the trained MLP, and the resulting changes in the output are observed. Perturbing all n regions sequentially yields the entire EC matrix. Perturbation is implemented by modifying the input $\mathbf{x}_t$ with an additive scaled unit vector $\mathbf{e}_i$, where the ith component is 1 (indicating the perturbed region), and all other components are 0. EC from region i to all others is quantified as the averaged response at time $t+1$ after perturbation:

$$\mathrm{EC}_{i\cdot} = \mathbb{E}_t[f(\mathbf{x}_t + \Delta \times \mathbf{e}_i, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}) - f(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2})]. \tag{3}$$

Here, $\mathbf{e}_i$ is a unit vector with a value of 1 at the ith entry and 0 elsewhere, representing perturbation in region ith. Δ denotes the perturbation magnitude, set to half the standard deviation of the BOLD signals. Given the nonlinear nature of brain dynamics, perturbation responses depend on the brain's state, similar to the state-dependent responses observed in real neural stimulation[43,44]. To account for this variability, virtual perturbations were performed at each time point's neural state. Participant-level EC was obtained by averaging perturbation responses across all states. Group-level EC and FC were derived by averaging connection strengths across all participants.

Ground-truth neural dynamical models

We validated the performance of NPI using three datasets: a public synthetic fMRI dataset and two generative models with known ground-truth EC, including an RNN model and a WBM. In RNN and WBM, ground-truth EC was determined by perturbing the activity of a node and observing the resulting propagation among other nodes.

RNN model. The RNN model consists of n nodes. We denote the state of the ith neuron as x_i and $\mathbf{x} = [x_1, \ldots x_n]^T$ is a n-dimensional vector that represents the states of all the n neurons in the network. The neural dynamics are governed by

$$\mathrm{d}\mathbf{x}(t) = [-\mathbf{x}(t) + Wh(\mathbf{x}(t))]\mathrm{d}t + \sigma \mathrm{d}\boldsymbol{\xi}(t), \tag{4}$$

where W is the weight matrix (representing SC) and $h(\cdot)$ is the tanh activation function, the noise $\boldsymbol{\xi}(t)$ is an n-dimensional standard Wiener process with independent components and σ scales the noise amplitude. The weight matrix W contains entries sampled from $\mathcal{N}(0, n^{-1})$. The initial state is sampled from the Gaussian distribution $\mathcal{N}(0,1)$. The RNN dynamics were simulated using the Euler method with $\Delta t = 0.01$:

$$\mathbf{x}(t+\Delta t) = \mathbf{x}(t) + [-\mathbf{x}(t) + Wh(\mathbf{x}(t))]\Delta t + \sigma\sqrt{\Delta t}\mathbf{Z}(t),\ \mathbf{Z}(t) \sim \mathcal{N}(0, \mathbf{I}_n), \tag{5}$$

where $\mathbf{Z}(t) \sim \mathcal{N}(0, \mathbf{I}_n)$ is a Gaussian white noise. Training data for NPI were extracted by downsampling the dynamics of $\mathbf{x}$ to a temporal resolution (TR) of 1 (taking every 100 points). Ground-truth EC was derived by perturbing the neural states at time t and observing responses at $t+1$. For node i, the initial signal $\mathbf{x}_t$ was perturbed to $\mathbf{x}_t + \Delta \times \mathbf{e}_i$ with $\Delta = 1$, and the system was run to compute $\mathbf{x}_{t+1}$. Ground-truth EC was calculated as the difference between $\mathbf{x}_{t+1}$ mapped from the perturbed and unperturbed $\mathbf{x}_t$ states.

Public synthetic BOLD dataset. This dataset[23] simulates neural firing rates transformed into BOLD signals via a hemodynamic response function (HRF). SC matrices feature sparse connectivity, with most values being zero and a few nonzero values sampled from $\mathcal{N}(0.5, 0.1)$, truncated to the range from 0.3 to 0.7. The dataset encompasses nine network structures with varying degrees of complexity, all of which feature cyclic structures. The dataset includes nine network structures with 5–10 nodes, incorporating unidirectional connections, two cycles and four cycles. Neural dynamics follow

$$\frac{\mathrm{d}\mathbf{z}}{\mathrm{d}t} = \sigma A\,\mathbf{z} + C\mathbf{u}, \tag{6}$$

where $\mathbf{z}$ represents firing rates, σ is a lag constant, A is the SC matrix between nodes and C is a matrix controlling how the external neuronal inputs $\mathbf{u}$ feed into the network. Observed BOLD signals $\bar{\mathbf{y}}$ are generated by

$$\bar{\mathbf{y}} = g(\mathbf{z}, \theta), \tag{7}$$

where $g(\cdot)$ is the HRF and θ represents its parameters. As ground-truth EC is unavailable in this dataset, EC inference performance was measured by the AUC for classifying the presence or absence of SC connections after binarizing inferred EC.

Whole-brain model. The dynamic mean-field model[45] simulates large-scale human brain dynamics with $N = 66$ excitatory neural assemblies. The firing rate r_i of population i is

$$r_i = F(I_i) = \frac{aI_i - b}{1 - \exp\left(-d\left(aI_i - b\right)\right)}, \tag{8}$$

where $F(\cdot)$ is a nonlinear function that maps net current to firing rate, $a = 270$ Hz/nA, $b = 108$ Hz, $d = 0.154$ s. The net current I_i is

$$I_i = wJ_N S_i + GJ_N \sum_{j=1}^{N} C_{ij} S_j + I_{\mathrm{bi}}, \tag{9}$$

where $J_N = 0.2609$, $w = 0.55$ and $G = 3.5$ are coupling parameters, C is the SC matrix derived from DSI[46] and I_{bi} is the background input modeled as an Ornstein–Uhlenbeck process. Synaptic drive S_i follows

$$\frac{dS_i}{dt} = F(I_i)\gamma(1 - S_i) - \frac{1}{\tau_s}S_i, \tag{10}$$

with $\tau_s = 100$ ms and $\gamma = 0.641$. BOLD signals $B_i(t)$ are obtained by convolving $S_i(t)$ with a Boynton γ kernel[47]. Training data for NPI were extracted with TR = 0.72 s, matching HCP data. Ground-truth EC was determined by perturbing I_t to $I_t + \Delta \times \mathbf{e}_i$ ($\Delta = 5$) and observing BOLD responses at $t + 4$ TR due to HRF lag. Ground-truth EC was calculated as the difference between perturbed and unperturbed BOLD signals at $t + 4$ TR.

Data processing

This study utilized rs-fMRI data from multiple datasets, including 800 healthy participants from the HCP dataset[24] (Figs. 3–5) and 2,600 healthy participants from the ABCD dataset[32] (Fig. 4). In both datasets, rs-fMRI data were recorded with or resampled to a TR of 0.72 s. The data from the HCP and ABCD datasets were preprocessed using the HCP minimal preprocessing pipeline[48], which included motion correction, brain extraction and spatial normalization. Denoising was performed using ICA-FIX, a method combining independent component analysis with the FSL tool FIX. The denoised data were then processed using the Nilearn package[49] to extract regional-level signals in the 0.01–0.1-Hz frequency range. For analyses involving diseased individuals (Supplementary Fig. 12), we used rs-fMRI data from 234 patients with autism and 285 healthy controls in the ABIDE dataset[34] and 60 patients with Alzheimer's disease and 60 healthy controls in the ADNI dataset[35]. Details of the preprocessing and analysis for the ABIDE and ADNI datasets are provided in Supplementary Note 6.

When evaluating the signal prediction performance of the surrogate models, each model is trained on 90% of the individual's fMRI data (the full first three sessions and 60% of the fourth session) and tested on the remaining 10% (the final 40% of the fourth session). When deriving individual EC and FC, all four sessions of each participant are used. When studying the relationship between the SC and EC, we used the SC matrix provided by Demirtaş et al.[50], derived using FSL's bedpostx and probtrackx2 workflows. The SC matrix was scaled to a range from zero to one and log-transformed. EC matrices were separately obtained for each participant using the NPI framework, trained on four fMRI runs per participant. The ECs were then averaged across 800 participants and scaled such that the strongest connection in the averaged EC matrix had a value of one. All individual ECs were then scaled by the same factor.

For analyses of the HCP and ABCD datasets, the brain was parcellated into 379 regions using the MMP 1.0 atlas[25], which includes 180 cortical regions per hemisphere and 19 subcortical regions. Our analysis primarily focused on the EC among the 360 cortical regions, with subcortical regions included during training to minimize bias in EC inference from unobserved areas. Parcellation was performed by averaging BOLD signals across voxels within each cortical region.

The 360 cortical regions were grouped into seven functional networks based on the resting-state networks defined by Yeo et al.[26]. These networks are the visual network (VIS), somatomotor network (SOM), dorsal attention network (DAN), ventral attention network (VAN), limbic network (LIM), frontoparietal control network (FPN) and default mode network (DMN). Each cortical region was assigned to the network with which it shared the most voxels. Seed regions were selected in the left-hemisphere core regions of each of the seven networks (Supplementary Table 3). Seed-based FC was calculated as the Pearson's correlation between the seed region and all other regions.

CCEP data were obtained from the F-TRACT project[51]. For comparison with the CCEP matrix, we used the NPI-inferred EBC matrix derived from HCP rs-fMRI data under the same MMP parcellation scheme.

Quantitative metrics and statistical analyses

To evaluate the quality of brain signal predictions, the coefficient of determination (R^2) was calculated between the ground-truth and predicted signals for each brain region, using the formula

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{11}$$

where y_i represents the actual signals, $\hat{y}_i$ represents the predicted signals, $\bar{y}$ is the mean of the actual signals and n is the number of time points. Overall R^2 is the averaged R^2 across all brain regions.

To assess ANN's ability to learn inter-regional relations, Pearson's correlation coefficient (r) was calculated between model FC and empirical FC. Model FC was obtained from data generated by the ANN with 1,200 TRs, where the ANN's output was recursively fed as input to simulate BOLD signals. Empirical FC was derived by calculating inter-regional correlation coefficients from the ground-truth data. The performance of EC inference was assessed by calculating Pearson's correlation coefficient (r) between ground-truth EC and NPI-inferred EC. For matrices with binary weights (Fig. 2i,j), we calculated the AUC to assess the model's ability to distinguish the presence or absence of specific connections correctly.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The synthetic data generated using the ground-truth RNN and WBM are publicly available on GitHub at https://github.com/ncclab-sustech/NPI/. The following datasets used in this study are accessible via their respective repositories: HCP dataset (https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release), ABCD dataset (https://abcdstudy.org/scientists/data-sharing/), CCEP dataset (https://f-tract.eu/atlas/), ABIDE dataset (https://fcon_1000.projects.nitrc.org/indi/abide/) and ADNI dataset (http://adni.loni.usc.edu). The brain atlases used in this study are also publicly available: MMP atlas (https://github.com/mbedini/The-HCP-MMP1.0-atlas-in-FSL), Schaefer atlases (https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal/Parcellations/MNI) and the AAL atlas (available from the Nilearn Python package). Source data are provided with this paper.

Code availability

The code supporting this study is available on GitHub at https://github.com/ncclab-sustech/NPI/, under the Apache License, v.2.0 (Apache-2.0). The main Python packages used in this study are numpy (v.1.26.4), torch (v.2.2.2), scipy (v.1.12.0), Nilearn (v.0.10.3), matplotlib (v.3.8.3), seaborn (v.0.13.2) and jupyter (v.1.1.1).

References

43. Scangos, K. W., Makhoul, G. S., Sugrue, L. P., Chang, E. F. & Krystal, A. D. State-dependent responses to intracranial brain stimulation in a patient with depression. *Nat. Med.* **27**, 229–231 (2021).
44. Lurie, D. J. et al. Questions and controversies in the study of time-varying functional connectivity in resting fmri. *Netw. Neurosci.* **4**, 30–69 (2020).
45. Deco, G. et al. Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. *J. Neurosci.* **33**, 11239–11252 (2013).
46. Hagmann, P. et al. Mapping the structural core of human cerebral cortex. *PLoS Biol.* **6**, e159 (2008).
47. Boynton, G. M., Engel, S. A., Glover, G. H. & Heeger, D. J. Linear systems analysis of functional magnetic resonance imaging in human v1. *J. Neurosci.* **16**, 4207–4221 (1996).

48. Glasser, M. F. et al. The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage* **80**, 105–124 (2013).
49. Abraham, A. et al. Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform*. https://doi.org/10.3389/fninf.2014.00014 (2014).
50. Demirtaş, M. et al. Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics. *Neuron* **101**, 1181–1194. e13 (2019).
51. Jedynak, M. et al. F-tract: a probabilistic atlas of anatomo-functional connectivity of the human brain. *Ebrains* https://doi.org/10.25493/5AM4-J3F (2023).

Acknowledgements

Q.L. was supported by the National Natural Science Foundation of China (62472206), the National Key R&D Program of China (2021YFF1200804), Shenzhen Excellent Youth Project (RCYX20231211090405003), Shenzhen Science and Technology Innovation Committee (2022410129, KJZD20230923115221044, KCXFZ20201221173400001), Guangdong Provincial Key Laboratory of Advanced Biomaterials (2022B1212010003), and the Center for Computational Science and Engineering at Southern University of Science and Technology. C.Z. was supported by Hong Kong RGC Senior Research Fellowship Scheme (SRFS2324-2S05). Y.H. was partly supported by ECS-26303921 from the Research Grants Council of Hong Kong. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the paper. Z. Liang was supported by GuangDong Basic and Applied Basic Research Foundation (2025A1515011645). We thank the support from the Swarma-TCCI scholarship to Z. Luo and Z. Liang. We also thank H. Wu, J. Jiang, K. Du, Y. Mu, P. Zhou, S. Gu and Z. Cui, and members of the NCC laboratory including C. Wei, K. Lou, Z. Li, X. Xu and S. Wang for their valuable discussions.

Author contributions

Z. Luo designed the study, developed the NPI framework, conducted the primary analyses and drafted the paper. K.P. contributed to validating the NPI framework on synthetic and real datasets, analyzing results and preparing the paper. Z. Liang and S.C. supported the comparison of NPI with competing methods and assisted in figure generation. C.X. provided technical expertise in designing the ANN architecture and predicting fMRI signals. D.L. contributed to fMRI data preprocessing. Y.H. and C.Z. offered critical guidance on NPI validation and result analysis. Q.L., as the corresponding author, conceptualized the study, supervised the research and drafted the paper. All authors reviewed and approved the final paper.

Competing interests

The authors declare no competing interests.

Additional information

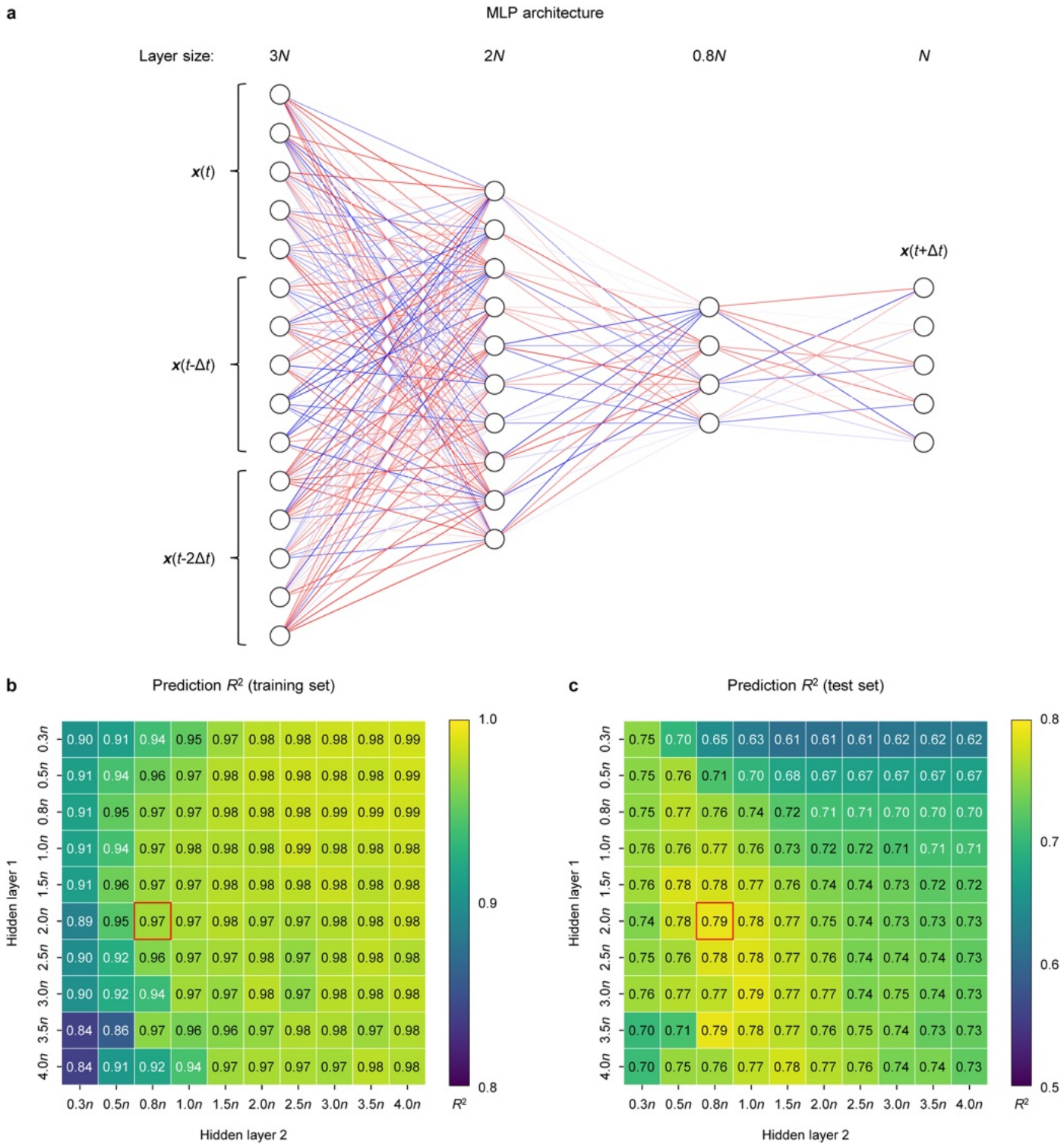
Extended data is available for this paper at https://doi.org/10.1038/s41592-025-02654-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-025-02654-x.

Correspondence and requests for materials should be addressed to Quanying Liu.

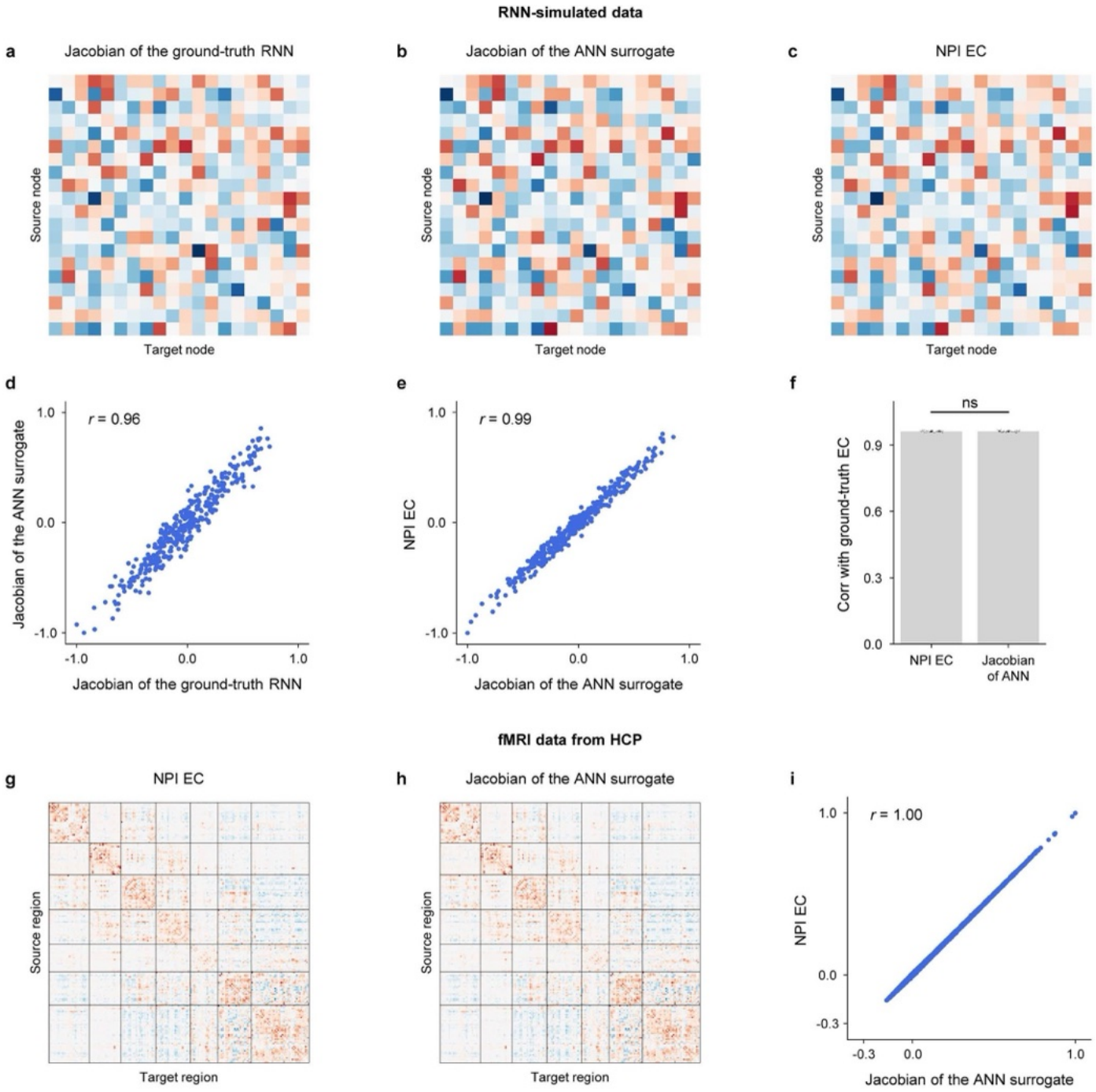
Peer review information *Nature Methods* thanks Enrico Amico, Thomas Bolton and Matthew Singh for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editor: Nina Vogt, in collaboration with the *Nature Methods* team.

https://doi.org/10.1038/s41592-025-02654-x



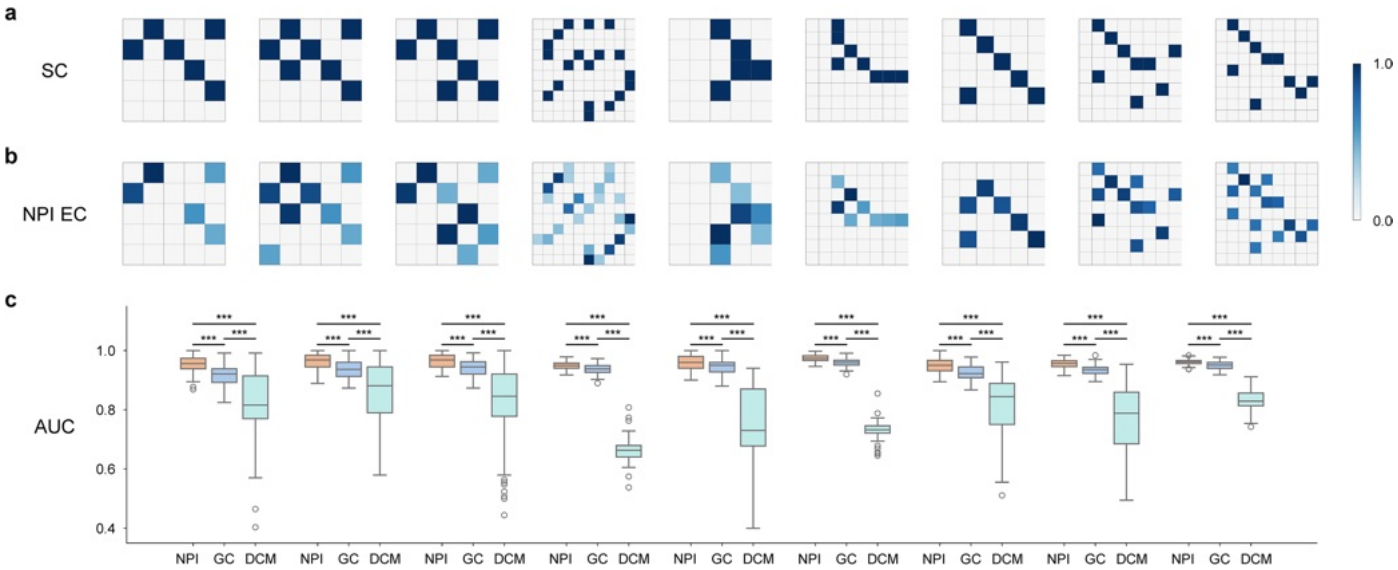
Extended Data Fig. 1 | Optimizing MLP architecture via grid search. (**a**) The MLP architecture used in our study, which includes an input layer, two hidden layers, and an output layer, derived from grid search. (**b**) The R^2 of one-step-ahead prediction on the training set under various sizes of hidden layer configurations, averaged across 20 participants. (**c**) The R^2 of one-step-ahead prediction on the test set under various sizes of hidden layer configurations, averaged across 20 participants.

https://doi.org/10.1038/s41592-025-02654-x



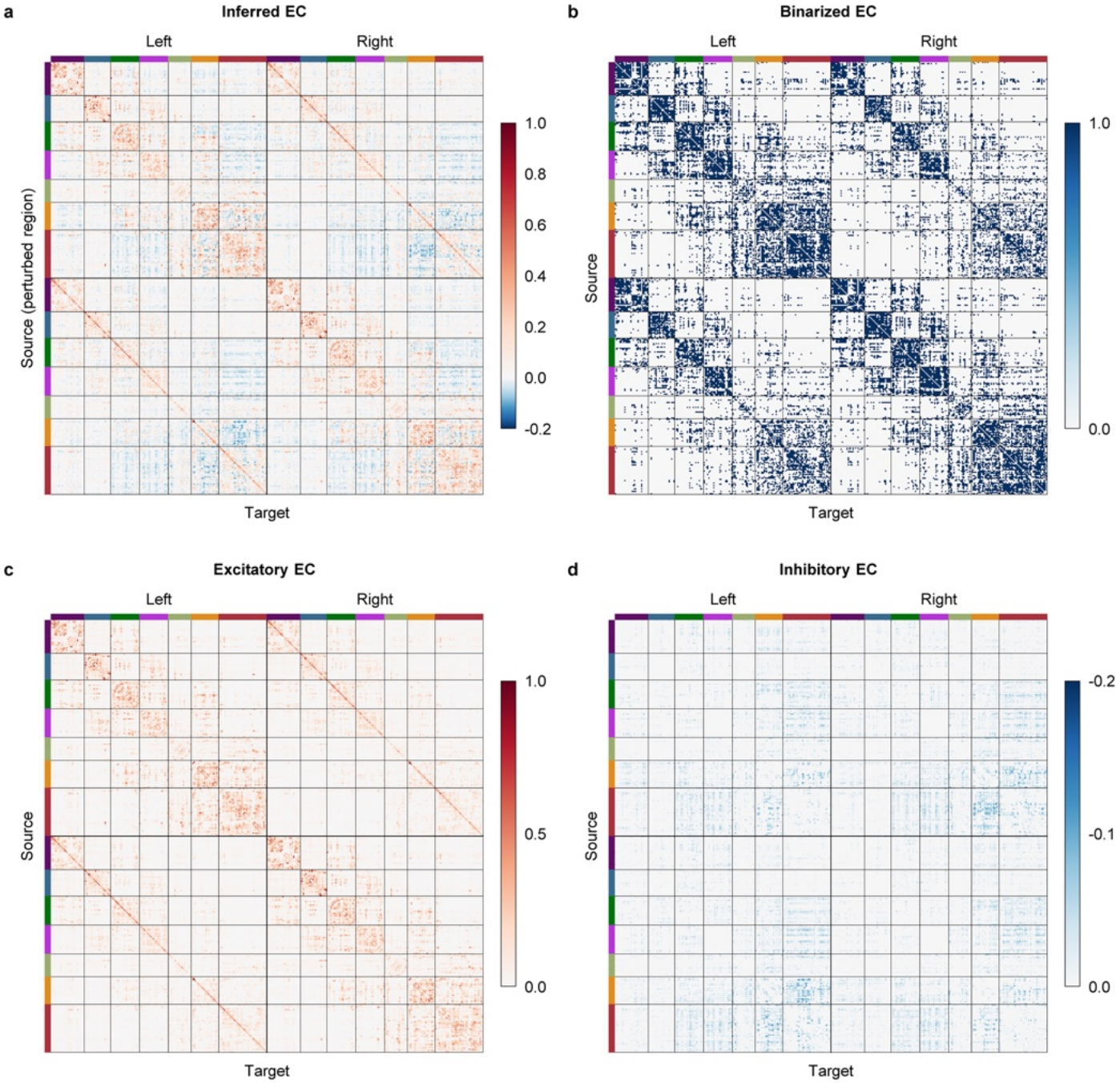
Extended Data Fig. 2 | NPI-inferred EC is consistent with the Jacobian matrix of the trained ANN model. (**a**) Jacobian matrix of an example RNN, numerically calculated using Pytorch. (**b**) Jacobian matrix of an ANN trained to predict the synthetic signals. (**c**) NPI-inferred EC by perturbing the trained ANN. (**d**) Jacobian of the trained ANN vs. Jacobian matrix of the ground-truth RNN across connection pairs. (**e**) NPI-inferred EC vs. Jacobian matrix of trained ANN across connection pairs. (**f**) Correlation coefficients between the NPI-inferred EC and the ground-truth EC, and between the Jacobian matrix and the ground-truth EC. There is no significant difference (P=0.87, two-sided Wilcoxon signed-rank test, n=50). Error bars represent standard deviation. (**g**) NPI-inferred EC on resting-state fMRI data from the HCP dataset, averaged across 800 participants. (**h**) Jacobian matrix of the ANN model trained on resting-state fMRI data from the HCP dataset, averaged across 800 participants. (**i**) NPI-inferred EC vs. Jacobian matrix of the trained ANN across connection pairs.

https://doi.org/10.1038/s41592-025-02654-x



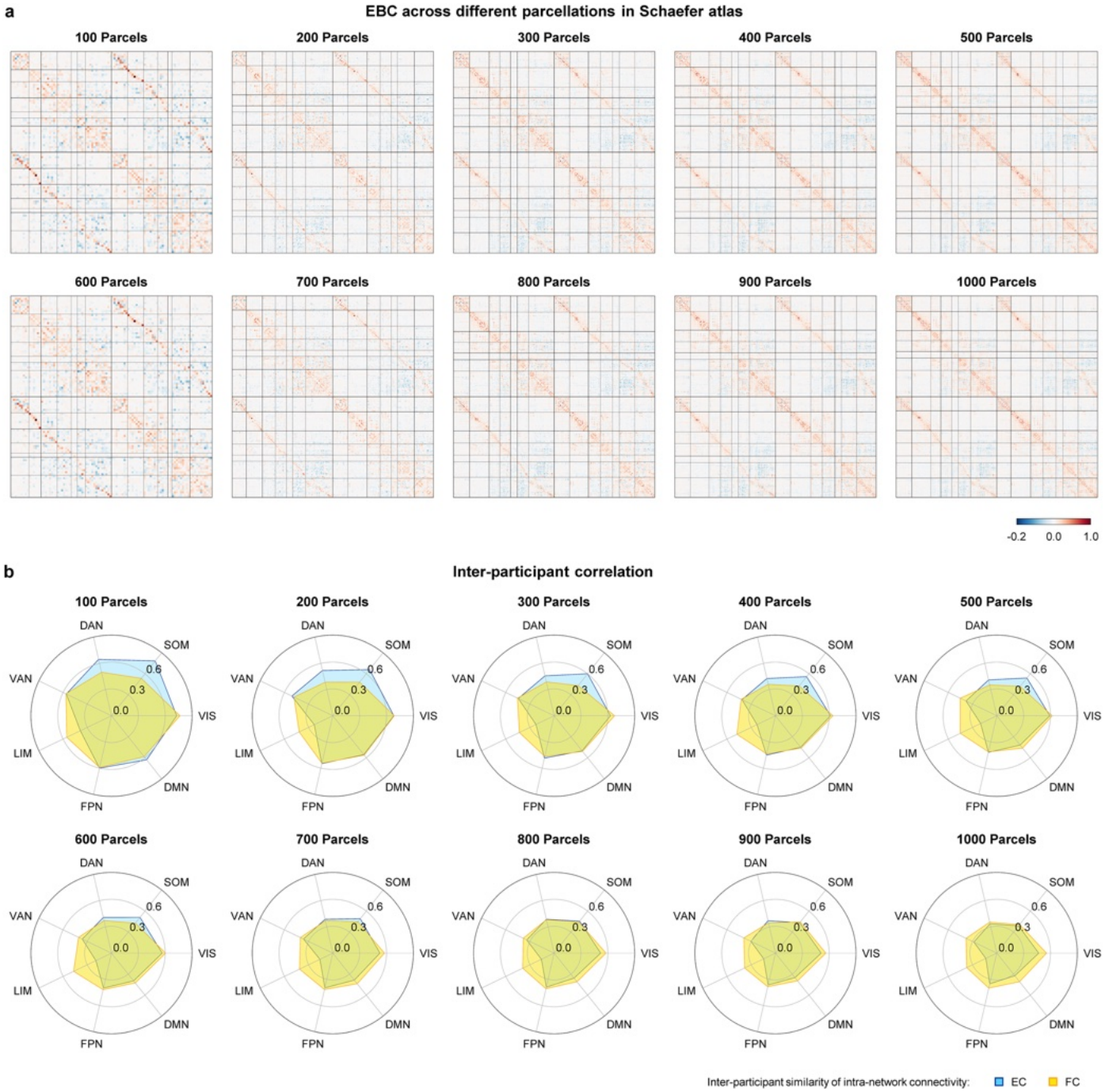
Extended Data Fig. 3 | The performance of NPI is reliable across different network topographies. (**a**) The test data are generated by generative models with predefined directed, binary structural connectivities (SC), from a public dataset[23] (**b**) NPI is utilized to map the EC from synthetic BOLD signals. (**c**) Comparisons of the AUC scores of EC inference with NPI across nine different SC configurations with the inferences obtained with GC and DCM. Error bars represent standard deviations. P values for nine structures (n=60 for each bar), NPI vs. GC, GC vs. DCM, and NPI vs. DCM respectively: Net1: $P = 6.25 \times 10^{-11}$, 7.48×10^{-9}, 8.40×10^{-11}, Net2: $P = 8.54 \times 10^{-9}$, 1.59×10^{-7}, 4.40×10^{-10}, Net3: $P = 1.57 \times 10^{-10}$, 2.23×10^{-10}, 5.82×10^{-11}, Net4: $P = 1.20 \times 10^{-6}$, 1.63×10^{-11}, 1.63×10^{-11}, Net5: $P = 2.08 \times 10^{-5}$, 1.63×10^{-11}, 1.62×10^{-11}, Net6: $P = 2.45 \times 10^{-10}$, 1.63×10^{-11}, 1.62×10^{-11}, Net7: $P = 7.85 \times 10^{-9}$, 6.32×10^{-10}, 3.37×10^{-11}, Net8: $P = 1.43 \times 10^{-10}$, 1.63×10^{-11}, 1.63×10^{-11}, Net9: $P = 4.80 \times 10^{-9}$, 1.63×10^{-11}, 1.63×10^{-11}.

https://doi.org/10.1038/s41592-025-02654-x



Extended Data Fig. 4 | The EBC, binarized EBC, and excitatory and inhibitory part of EBC. (**a**) The whole-brain EBC. (**b**) EBC binarized by a threshold of 80% strength of EC. The entries larger than the threshold are set to 1, while the rest are set to 0. (**c**, **d**) The excitatory (**c**) and inhibitory (**d**) parts of EBC.

https://doi.org/10.1038/s41592-025-02654-x



Extended Data Fig. 5 | The NPI-inferred EC derived from brain atlases with increasing numbers of regions. (**a**) EBC across different resolution of parcellations in Schaefer atlases. The group-level EBC are averaged across 100 participants. (**b**) Inter-participant correlation of individual EC and FC across different parcellations in Schaefer atlases. Results are averaged across 100 participants.

nature portfolio

Corresponding author(s): Quanying Liu

Last updated by author(s): Mar 20, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The code for data simulation is available at \url{https://github.com/ncclab-sustech/NPI/}, under the Apache License, Version 2.0 (Apache-2.0).
Data analysis	The code for data analysis is available at \url{https://github.com/ncclab-sustech/NPI/}, under the Apache License, Version 2.0 (Apache-2.0). The main Python packages used in this study are: numpy (1.26.4), torch (2.2.2), scipy (1.12.0), nilearn (0.10.3), matplotlib (3.8.3), seaborn (0.13.2), jupyter (1.1.1).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The synthetic data generated using the ground-truth RNN and whole-brain model are publicly available at \url{https://github.com/ncclab-sustech/NPI/}. The following datasets used in this study are accessible via their respective repositories: HCP dataset: \url{https://www.humanconnectome.org/study/hcp-young-adult/

document/1200-subjects-data-release}, ABCD dataset: \url{https://abcdstudy.org/scientists/data-sharing/}, CCEP dataset: \url{https://f-tract.eu/atlas/}, ABIDE dataset: \url{http://fcon_1000.projects.nitrc.org/indi/abide/}, ADNI dataset: \url{http://adni.loni.usc.edu}. The brain atlases used in this study are also publicly available: MMP atlas: \url{https://github.com/mbedini/The-HCP-MMP1.0-atlas-in-FSL}, Schaefer atlases: \url{https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal/Parcellations/MNI}, AAL atlas: available from the Nilearn Python package.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Data of 800 subjects from HCP dataset were used. Data of 234 autism patients and 285 healthy controls from ABIDE dataset were used. Data of 60 patients with Alzheimer's disease and 60 healthy controls from ADNI dataset were used.
Data exclusions	No data were excluded.
Replication	The study can be reproduced using public codes.
Randomization	For HCP, ABIDE, and ADNI datasets, data we used were randomly chose from all available data.
Blinding	Not relevant. We only used public datasets.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging