nature research

https://doi.org/10.1038/s41551-020-00666-w

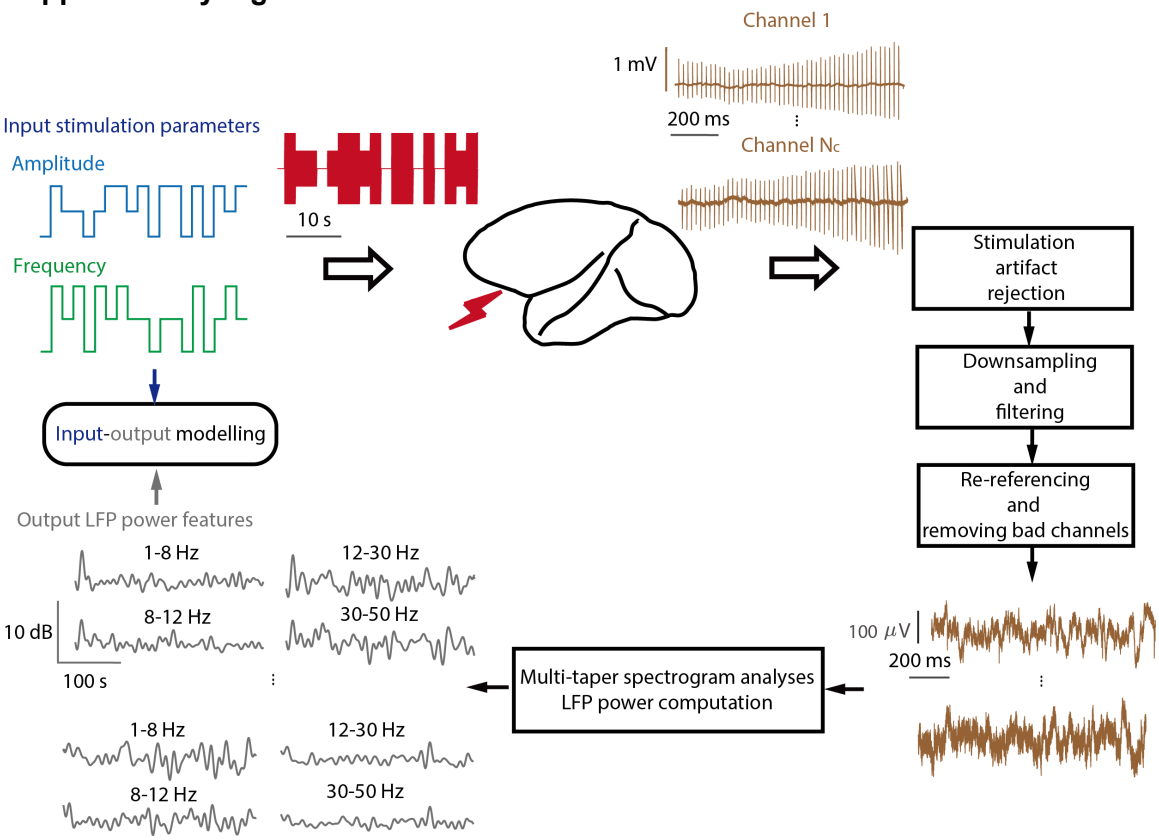
Supplementary information

Modelling and prediction of the dynamic responses of large-scale brain networks during direct electrical stimulation

In the format provided by the authors and unedited

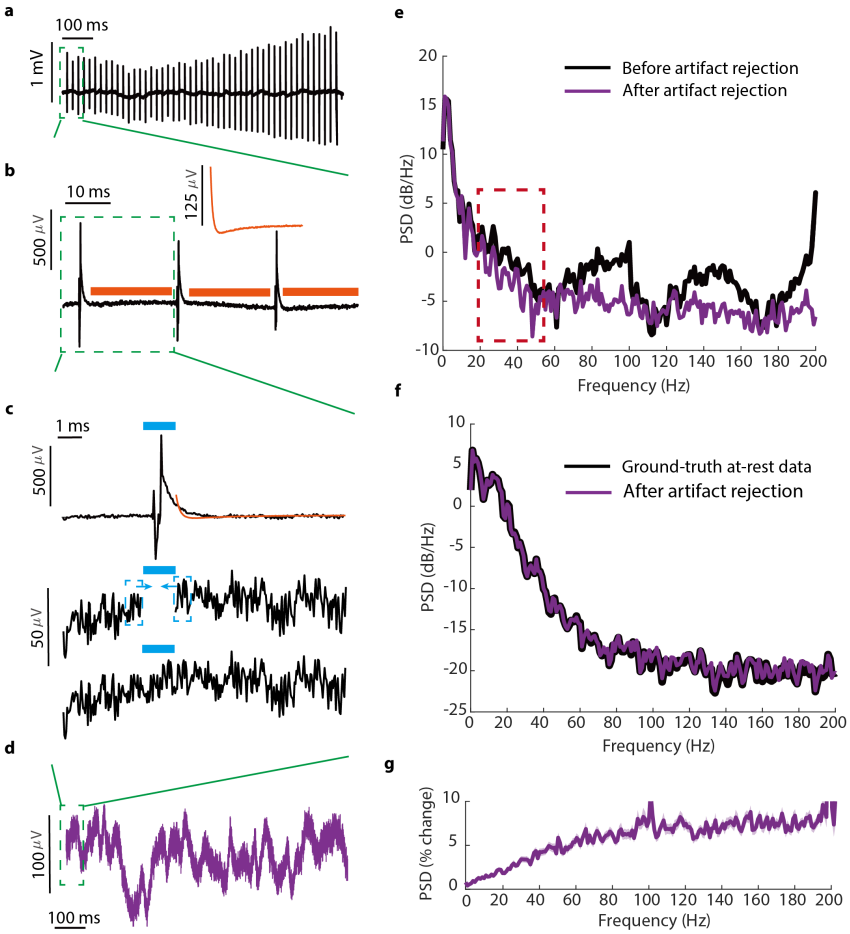
Supplementary Information Contents

Supplementary Figures



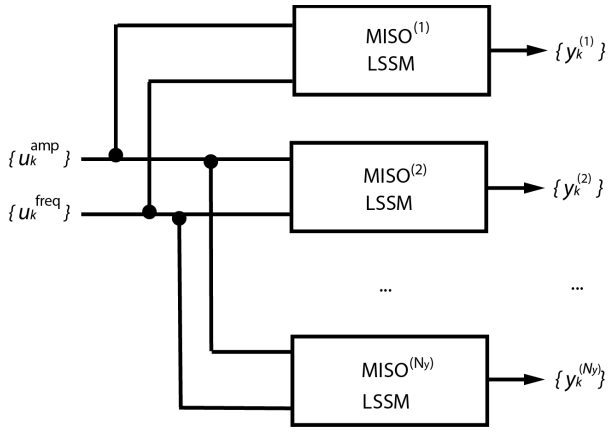
Supplementary Fig. 1 | Details of neural data collection and preprocessing.

Multi-level noise (MN)-modulated pulse train was used to stimulate at various stimulation sites using bipolar stimulation (Supplementary Table 3), and raw signals across multiple brain regions (Supplementary Tables 1 and 2) were simultaneously recorded at 30kHz. Raw signals were processed offline: 1) stimulation artifacts were rejected; 2) stimulation artifact-rejected raw signals were low-pass filtered, downsampled to 200Hz and further filtered to remove non-neural noise such as line noise (see Methods); 3) filtered local field potential (LFP) signals were re-referenced to their closest channels and channels in white matter and with large non-neural noise were removed. Output LFP power features from 4 frequency bands, i.e., 1–8 Hz (delta+theta), 8–12 Hz (alpha), 12–30 Hz (beta), 30–50 Hz (low-gamma), were then calculated from each LFP signal using multi-taper spectrogram analyses (Supplementary Note 2). Finally, the output LFP power features and the input MN stimulation parameters formed the input-output (IO) dataset to be used for IO modelling.



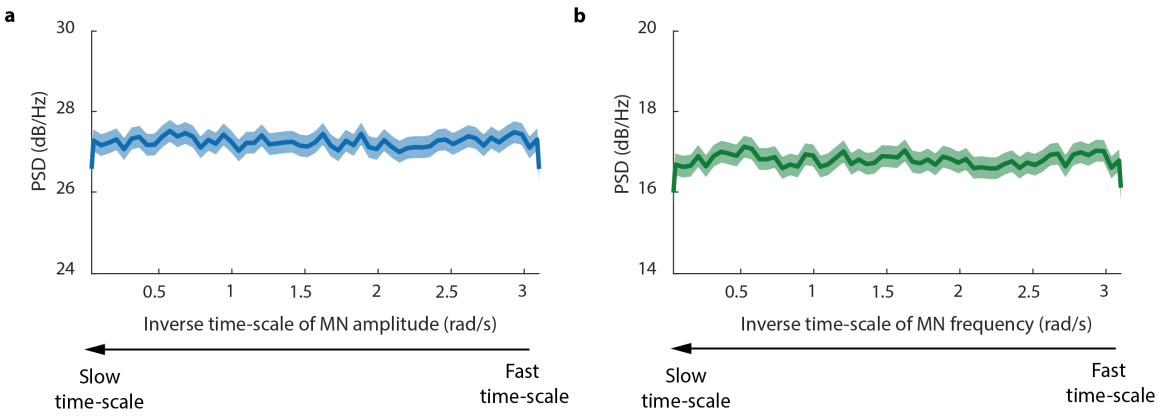
Supplementary Fig. 2 | Stimulation artifact rejection.

a, an example channel shows the procedure of stimulation artifact rejection. Raw signal recorded during stimulation was masked by large stimulation artifacts consisting of a large “instantaneous artifact spike” and a long “artifact tail”. **b**, obtaining the template of the “artifact tail”. Orange bars represent the time window for obtaining the template, which was chosen as the ending point of the current “instantaneous artifact spike” to the starting point of the next “instantaneous artifact spike” (see Supplementary Note 1). The orange signal represents the obtained template of the “artifact tail”. **c**, removal of the “artifact tail” using template subtraction and removal of the “instantaneous artifact spike” using interpolation. The “artifact tail” template (orange) was subtracted from the raw signal. The data during the “instantaneous artifact spike” window (blue bar) was discarded and treated as missing data. The missing data was then interpolated by the raw data immediately before and after the “instantaneous artifact spike” window (dashed blue boxes). **d**, clean raw signal after stimulation artifact rejection. **e**, power spectral density (PSD) of the stimulation artifact-rejected signal compared with the raw signal. Dashed red boxes show an example where stimulation artifacts introduced artificial increase of powers. f, control analyses using at-rest data (see Supplementary Note 1) show that the stimulation artifact rejection algorithm introduced negligible distortion of the signal spectrum. g, in the control analyses, the percentage of change of the PSD of the stimulation artifact rejected signal relative to the PSD of the true at-rest data was smaller than 5% below 50Hz (which is the main frequency range of interest in this study, Supplementary Notes 1 and 2). Solid line represents mean and shaded area represents s.e.m.



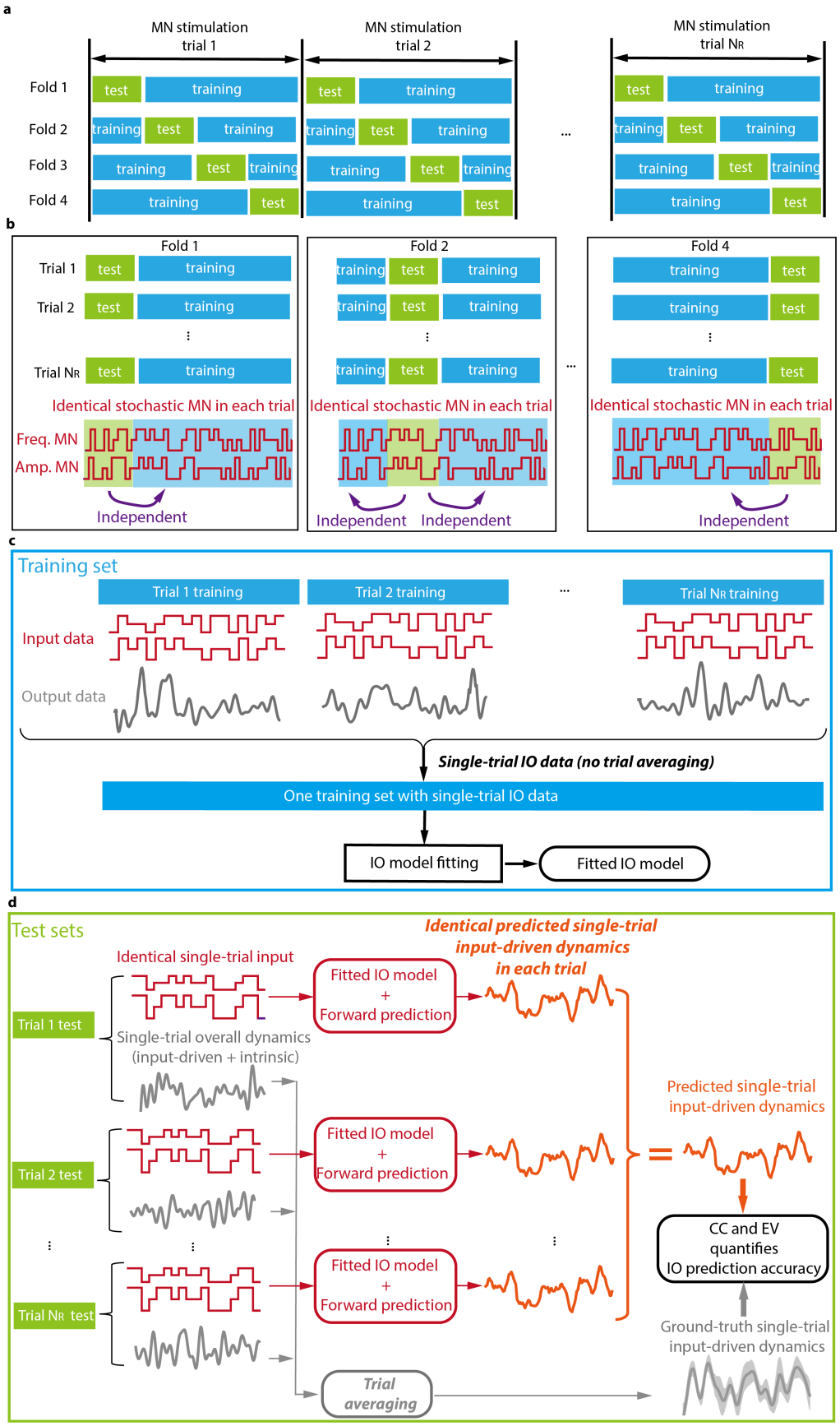
Supplementary Fig. 3 | The multiple-input-multiple-output (MIMO) linear state-space model (LSSM) is constructed as a series of multiple-input-single-output (MISO) LSSMs.

The MIMO LSSM describes how the time-evolution of stimulation amplitude $\{u_k^{\text{amp}}\}$ and frequency $\{u_k^{\text{freq}}\}$ affects the time-evolution of the LFP power features $\{y_k\}$. For each component in $\{y_k\}$, e.g., the ith component $\{y_k^{(i)}\}$, we modeled its dynamic response to stimulation amplitude and frequency using a MISO LSSM, e.g., MISO$^{(i)}$.



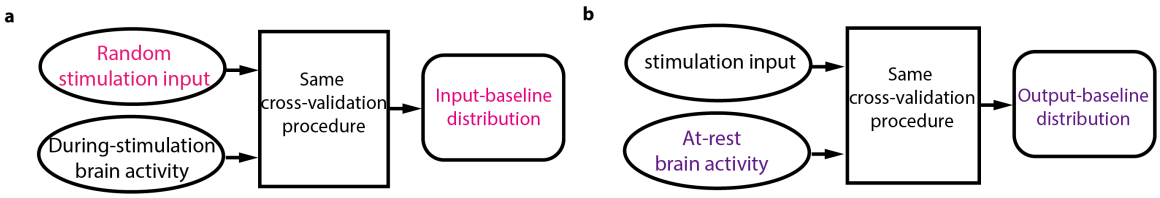
Supplementary Fig. 4 | MN stimulation amplitude and frequency have white spectrum across all time-scales.

a, The spectrum of MN stimulation amplitude. Solid line represents the mean PSD of 1000 independently generated 240-sample long MN amplitude sequences. Shaded area represents s.e.m. **b**, The spectrum of MN stimulation frequency. Figure convention is the same as (a).


a
MN stimulation trial 1
MN stimulation trial 2
MN stimulation trial NR
Fold 1
Fold 2
Fold 3
Fold 4
test
training
b
Fold 1
Fold 2
Fold 4
Trial 1
Trial 2
Trial NR
Identical stochastic MN in each trial
Freq. MN
Amp. MN
Independent
c
Training set
Trial 1 training
Trial 2 training
Trial NR training
Input data
Output data
Single-trial IO data (no trial averaging)
One training set with single-trial IO data
IO model fitting
Fitted IO model
d
Test sets
Identical single-trial input
Identical predicted single-trial input-driven dynamics in each trial
Trial 1 test
Trial 2 test
Trial NR test
Single-trial overall dynamics (input-driven + intrinsic)
Fitted IO model + Forward prediction
Predicted single-trial input-driven dynamics
CC and EV quantifies IO prediction accuracy
Ground-truth single-trial input-driven dynamics
Trial averaging

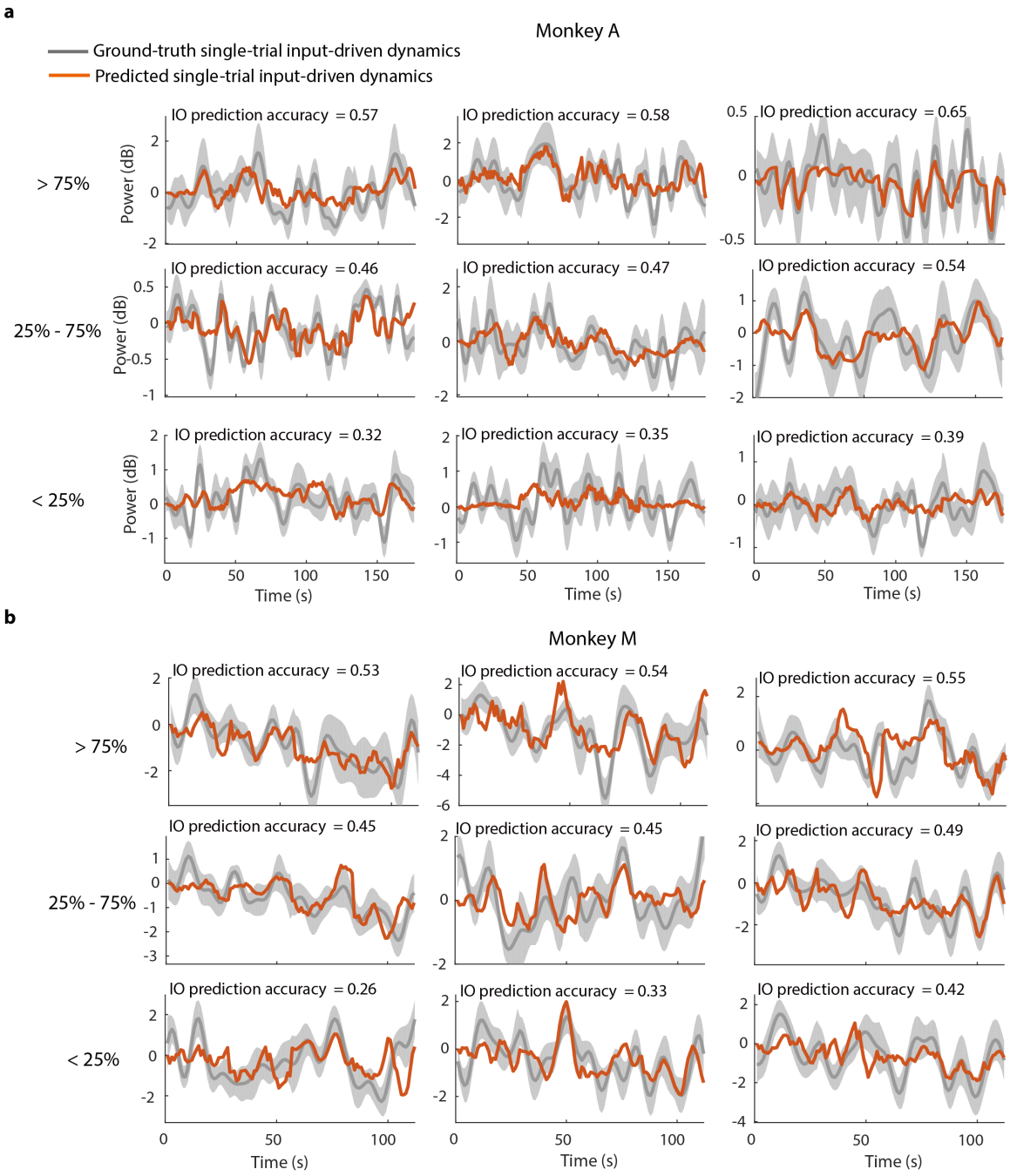
Supplementary Fig. 5 | Details of multi-trial experimental design and the cross-validation procedure.

a, forming the training and test sets in four-fold cross-validation. In each cross-validation fold (each row), the test sets (green) were taken from the same quarter of IO data in each trial and the rest of the data (blue) was used as the training set. Within each cross-validation fold, the green segment in each trial was taken as one test set; thus in total we had N_R test sets (see (d)). Also, the blue segments of all trials (all blue segments across the row) were concatenated as one training set (see (c)). The training set thus consists of concatenated single-trial data. **b**, independence of test and training sets in each cross-validation fold. The training set did not see any data from the test sets. The input in all test sets was identical since our experimental design used the same stochastic MN input for all trials. The input in the test sets was independent of the input in the training set because MN is a stochastic white-spectrum pattern that evolves randomly over time (Supplementary Note 9). **c**, model fitting in an example cross-validation fold. IO model was fitted only using the training set. The stimulation input and LFP power features (output) used in training were not averaged across trials. Model was fitted to *single-trial* IO data, which was concatenated across trials (and not averaged). **d**, model evaluation in an example cross-validation fold. IO model was evaluated in the test sets. Model evaluation was based on forward prediction to predict the single-trial input-driven dynamics, and then compare with the ground-truth of single-trial input-driven dynamics. Since an identical stochastic input was used for all test sets and since forward prediction only uses the past input, the single-trial forward predictions in all test sets were also identical. We thus used this single-trial forward prediction to get the predicted single-trial input-driven dynamics (note that trial-averaging the predicted single-trial input-driven dynamics keeps them intact because the input and thus these predicted single-trial input-driven dynamics are identical across all trials for test sets). To obtain the ground-truth of single-trial input-driven dynamics, we averaged the measured overall brain network dynamics—i.e. measured LFP power feature time-series—across the test sets (this is the only time averaging is done in the evaluation and is needed to recover the ground-truth by reducing intrinsic dynamics that are input-irrelevant). We finally computed the cross-validated correlation coefficient (CC) between the predicted single-trial input-driven dynamics and their ground-truth. Note that the ground-truth was also the ground-truth of the *single-trial* input-driven dynamics because the input was the same across trials and thus trial-averaging the measured LFP power features kept the single-trial input-driven dynamics intact while suppressing intrinsic dynamics that are input-irrelevant. The cross-validated CC quantified the IO prediction accuracy.



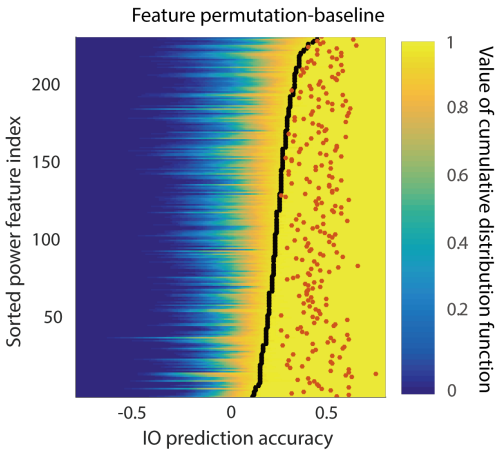
Supplementary Fig. 6 | Input-baseline test and output-baseline test.

a, input-baseline test where the input of the IO dataset was replaced with randomly generated MN patterns and the same cross-validation procedure and model fitting was repeated to obtain the input-baseline distribution of IO prediction accuracy. **b**, output-baseline test where the output of the IO dataset was replaced with the at-rest data on which stimulation artifact rejection control was also applied. The same cross-validation procedure and model fitting was then repeated to obtain the output-baseline distribution of IO prediction accuracy.



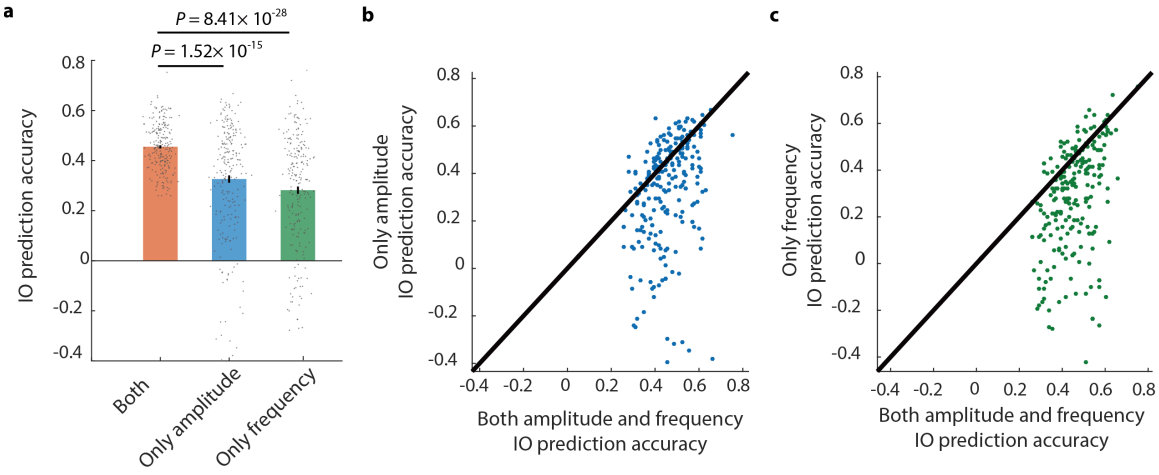
Supplementary Fig. 7 | Example forward predictions of single-trial input-driven dynamics in cross-validation.

a, nine example forward predictions of single-trial input-driven dynamics in cross-validation for Monkey A. The IO prediction accuracies in the three rows are within > 75% (top), 25%-75% (middle) and < 25% (bottom) quantiles of the distribution of IO prediction accuracy. Grey shaded area provides the s.e.m. in the ground-truth. Other figure conventions are the same as Fig. 2a. b, same as (a) but for Monkey M. A subset of these are also provided in Fig. 3.



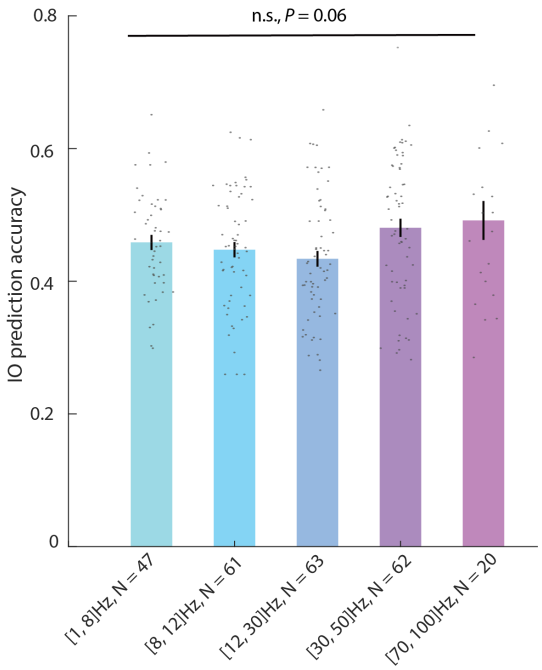
Supplementary Fig. 8 | Visualization of the feature permutation-baseline test.

To show that IO model predictions are specific to the LFP power feature from the recorded site, we conducted a new statistical test termed feature permutation-baseline test. To compute the feature permutation-baseline distribution, for each predictable power feature, we used its fitted IO model to forward predict all other LFP power features within the network; we then computed the corresponding IO prediction accuracy for these other LFP power features and used their distribution as the feature permutation-baseline distribution. We define the *P* value as the probability that IO prediction accuracy from the feature permutation-baseline distribution is greater than the IO prediction accuracy from the actual IO dataset. The results of this feature permutation-baseline tests for all predictable power features are shown in the figure. Figure conventions are the same as in Fig. 2b except that the cumulative distribution function of the feature permutation-baseline is shown. The actual IO prediction accuracy value for each predictable power feature is shown by the red dot, which was larger than the value associated with the black dot (i.e., larger than the value associated with the significance level of the feature permutation-baseline).



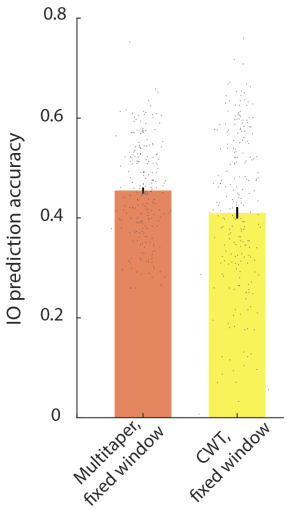
Supplementary Fig. 9 | Brain network dynamics respond to real-time changes in both stimulation amplitude and frequency.

a, the IO prediction accuracy of modelling both the input stimulation amplitude and frequency was significantly larger than only modelling the input stimulation amplitude or only modelling the input stimulation frequency. The bar represents the mean, and the black error bar represents s.e.m. Raw IO prediction accuracy is shown with dots (N = 233 independent samples, i.e., LFP power features, for all bars). Two-sided Wilcoxon signed-rank test P values are also shown. **b**, each colored dot shows the IO prediction accuracy of a predictable power feature when only modelling the stimulation amplitude vs. when modelling both the input stimulation amplitude and frequency (N = 233 independent samples, i.e., LFP power features). The dots are largely below the 45-degree equality line showing the benefit of modelling frequency. **c**, same as (a) but showing only modelling the stimulation frequency. The dots are largely below the 45-degree equality line showing the benefit of modelling amplitude (N = 233 independent samples, i.e., LFP power features).



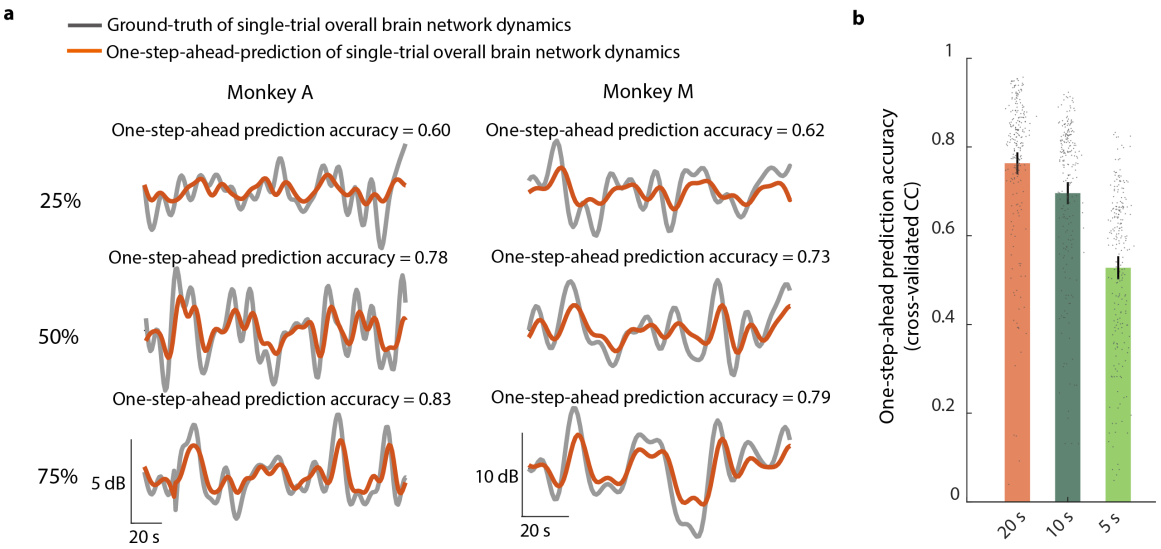
Supplementary Fig. 10 | IO prediction accuracy is similar across LFP frequency bands, including the high gamma band.

The bar represents the mean, and the black error bar represents s.e.m. Raw IO prediction accuracies are shown with dots. The sample number *N* (number of LFP power features) is indicated for each bar. Kruskal-Wallis test *P* value is also shown. n.s.: not significant. Note that we were able to compare the high gamma band 70-100 Hz to the other four bands because we used the output-baseline test, which controlled for the effect of stimulation artifacts. Indeed, compared with the other four bands and among power features that pass the input-baseline test, fewer high gamma power features passed the output-baseline test (percentage not passing the output-baseline test: high gamma band 81.66% ± 7.04% vs. the other four bands 28.46% ± 6.84%, mean ± s.e.m., two-sided Wilcoxon signed-rank test, $P = 6\times10^{-5}$). This higher percentage for the high gamma band was because we delivered MN stimulation at 50Hz and 100Hz, which were near/within the high gamma band. Thus the stimulation artifact had larger residues in the gamma band after stimulation artifact rejection (Supplementary Fig. 2g). Output baseline test allowed us to remove power features that had residues and thus successfully make the above comparison.



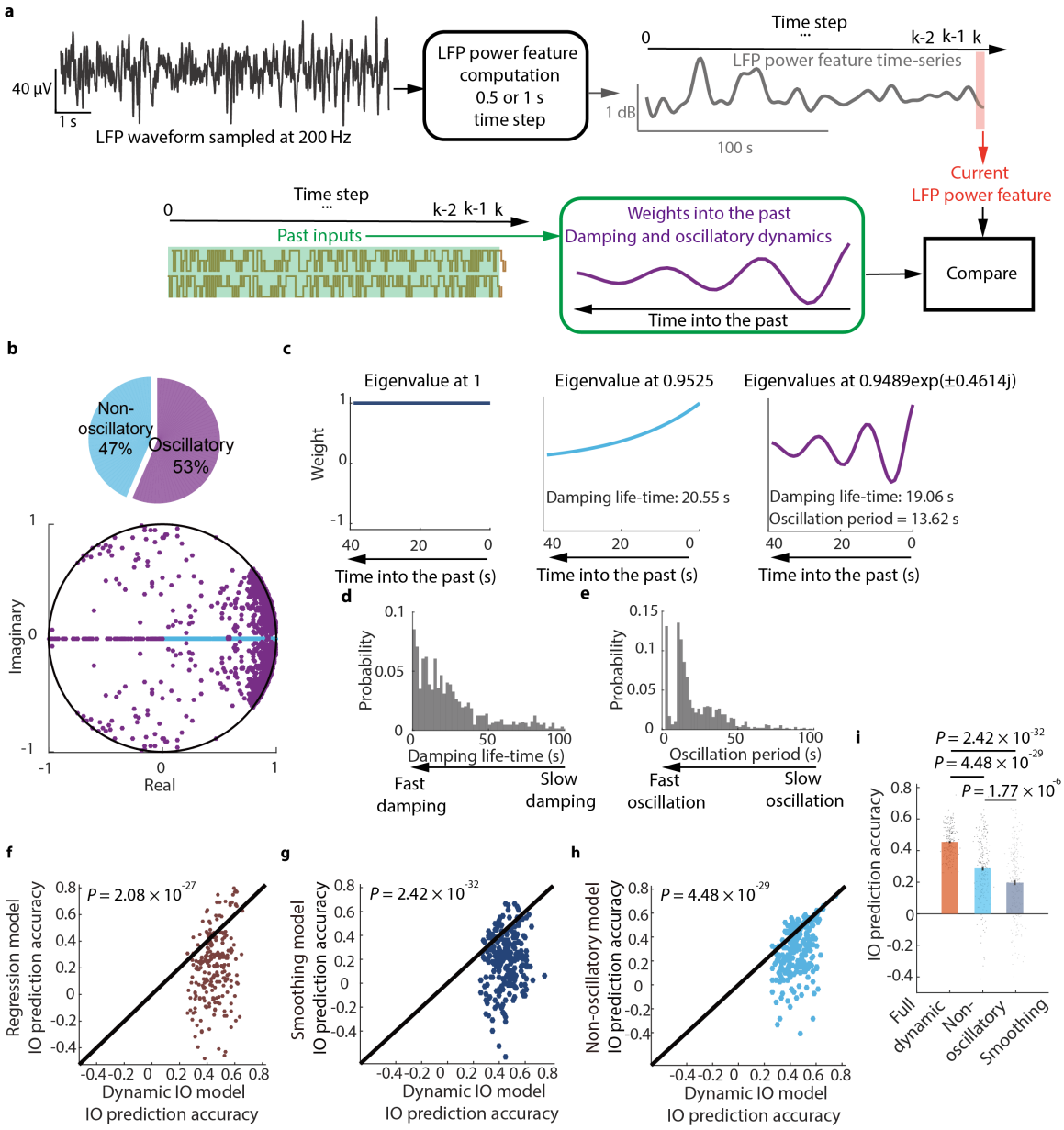
Supplementary Fig. 11 | IO modelling is robust to the method used to compute the LFP power features.

The bar represents the mean, and the black error bar represents s.e.m. Raw IO prediction accuracy is shown with dots (*N* = 233 independent samples, i.e., LFP power features, for both bars). The IO prediction accuracy using LFP power features computed using the continuous wavelet transform (CWT) method was similar to that of the LFP power features computed using the multitaper method in the main analyses (0.41 ± 0.01 for CWT vs. 0.45 ± 0.006 in the main analyses; mean ± s.e.m.).



Supplementary Fig. 12 | The dynamic IO model enables one-step-ahead prediction of single-trial overall brain network dynamics.

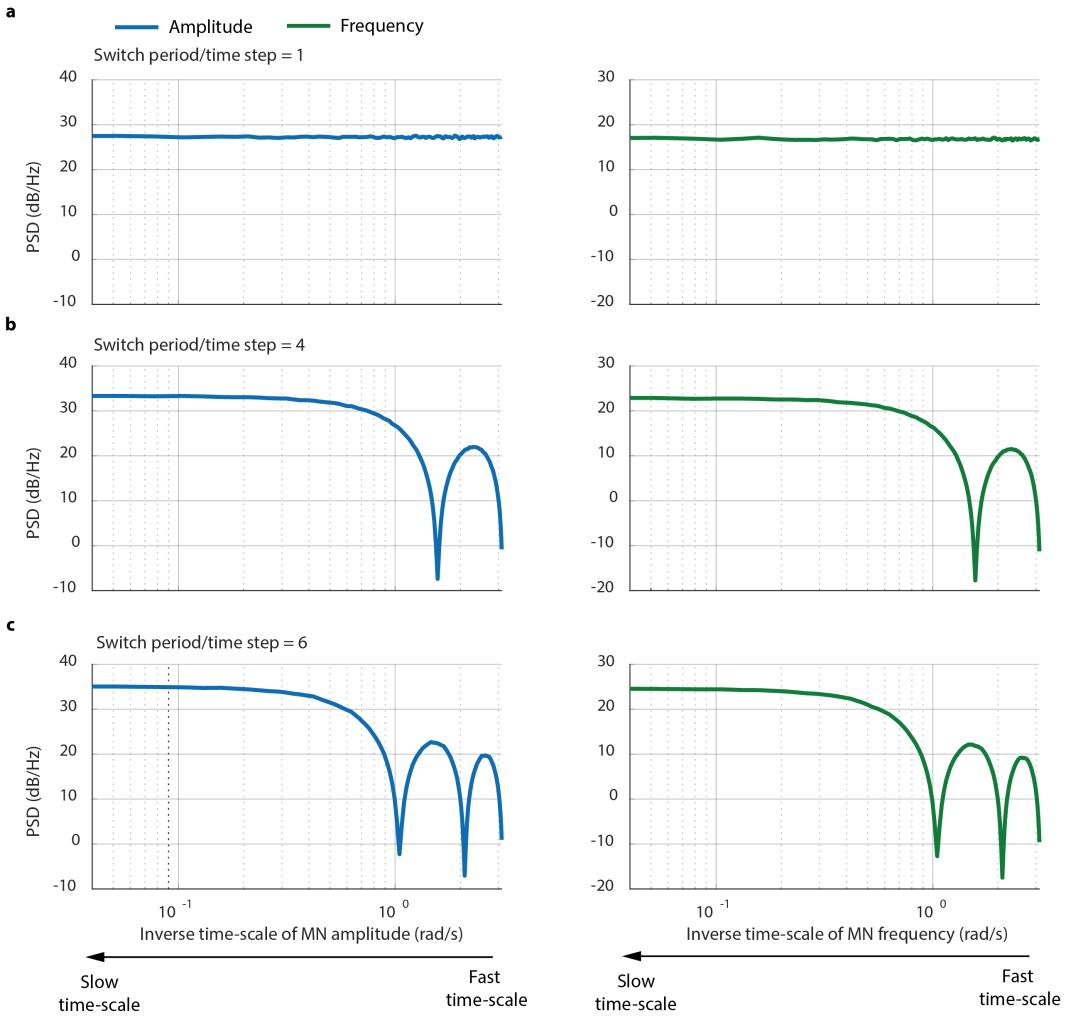
a, three example one-step-ahead predictions of single-trial overall brain network dynamics—i.e., of measured LFP power feature time-series—in the test sets using the fitted IO model from the training set in Monkey A (left) and in Monkey M (right). The one-step-ahead prediction accuracy (cross-validated correlation coefficient (CC)) of the three examples are around the 25% (top), 50% (middle) and 75% (bottom) quantiles of the distribution of one-step-ahead prediction accuracy across all predictable power features (indicated to the left of each row). **b**, robustness of the one-step-ahead prediction accuracy to the time-scale used for computing the correlation coefficient. The bar represents the mean, and the black error bar represents s.e.m. Raw IO prediction accuracies are shown with dots (N = 233 independent samples, i.e., LFP power features, for all bars). Within the time-scale of the input effect (around 20 s and slower, see Supplementary Fig. 15), the cross-validated CC between the one-step-ahead prediction of the LFP power feature and its measured single-trial value was 0.76 ± 0.02 (mean ± s.e.m., across all predictable power features; output-permutation test, FDR corrected $P < 10^{-16}$). This result shows that the models can also predict single-trial overall brain network dynamics. This prediction was also achieved for faster time-scales that were twice or even four times faster (0.69 ± 0.02 for 10 s and 0.53 ± 0.03 for 5 s, FDR corrected $P < 10^{-16}$ for all predictable power features in both cases).



Supplementary Fig. 13 | Eigenvalues of the state transition matrices of the fitted IO LSSMs and oscillatory dynamics.

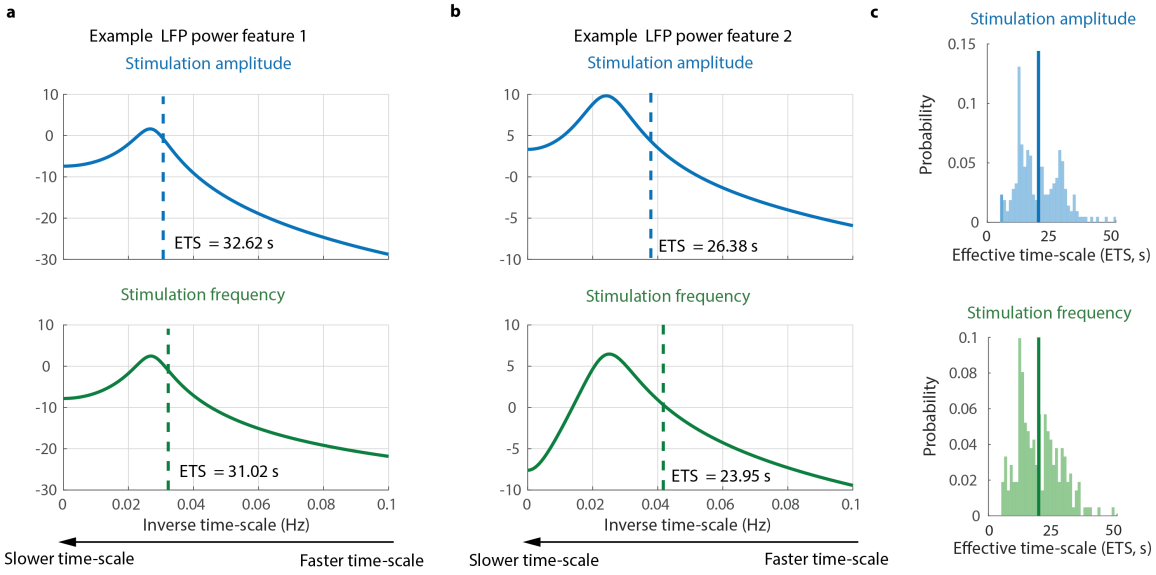
a, the LFP power feature time-series is computed from LFP waveform using a fixed time step of 0.5 or 1 s (see Methods). The predicted single-trial input-driven dynamics of the LFP power feature represent how past stimulation inputs are weighted to predict the current output LFP power feature. **b**, lower panel: Complex domain eigenvalue distribution across all fitted LSSMs from the 16 IO datasets. Eigenvalues on the positive real axis are colored blue and elsewhere purple. All eigenvalues were within the unit circle (black circle), representing stable dynamics of the brain network response to stimulation. Upper panel: Pie chart of the number of non-oscillatory eigenvalues (the ones on positive real axis) and oscillatory eigenvalues (the ones on negative real axis and the complex conjugate ones). **c**, example weights of the past inputs in predicting the current output LFP power feature associated with different eigenvalues. Eigenvalue at 1 represents smoothing dynamics without any damping (left). Eigenvalue on the positive real

axis but smaller than 1 represents non-oscillatory dynamics with exponential damping (middle). Pure damping dynamics (non-oscillatory dynamics) mean that the farther into the past the input is, the smaller its weight/contribution is in predicting the current LFP power features. In other words, the effect of the input farther into the past 'washes out' more compared with the input closer to the current time step. Conjugate complex eigenvalues represent oscillatory dynamics with exponential damping (right). Oscillatory dynamics mean that the weighting of past inputs oscillates at a certain period. Here j is the unit imaginary number. **d**, distribution of the damping life-time of oscillatory eigenvalues (see Supplementary Note 13 for the definition of damping life-time). **e**, distribution of the oscillation period of oscillatory eigenvalues (see Supplementary Note 13 for the definition of oscillation period). **f**, the IO prediction accuracy of the dynamic IO model was significantly larger than the regression model. In the left scatter plot, each colored dot shows the IO prediction accuracy of the regression model vs. that of the dynamic IO model for one predictable power feature, with the solid 45-degree line representing equality. The dots are largely below the 45-degree line, showing the benefit of the dynamic IO model vs. regression (N = 233 independent samples, i.e., LFP power features). Two-sided Wilcoxon signed-rank test P value is also shown. **g**, the prediction accuracy of the dynamic IO model was significantly larger than the smoothing model. Figure convention is the same as (**f**). h, the prediction accuracy of the dynamic IO model was significantly larger than the non-oscillatory model. Figure convention is the same as (**g**). i, the bar plot summarizes panels (**f, g, h**). The bar represents the mean, and the black error bar represents s.e.m. Raw IO prediction accuracies are shown with dots (N = 233 independent samples, i.e., LFP power features, for all bars). Two-sided Wilcoxon signed-rank test P values are also shown.



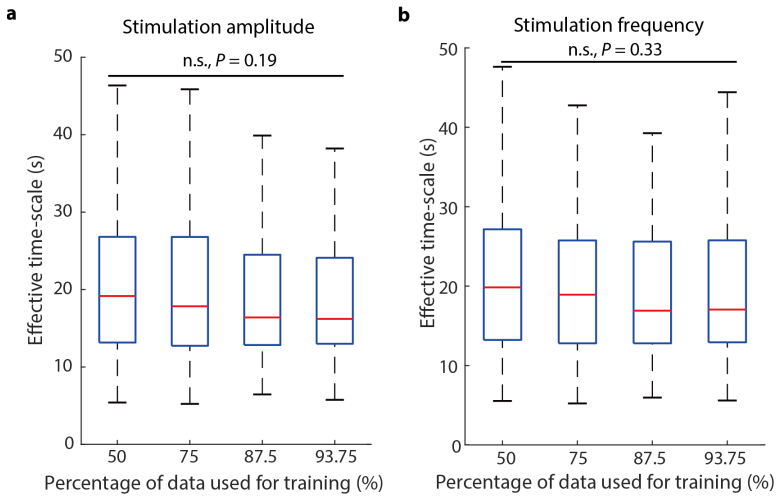
Supplementary Fig. 14 | The spectrum of MN amplitude or frequency patterns varies with an MN design parameter, which is the ratio of the switch period over time step (T_{sw}/T_s).

a, the spectrum of MN amplitude and frequency with T_{sw}/T_s = 1, which had white spectrums across all time-scales. Solid line represents the mean PSD of 1000 independently generated 240-sample long MN patterns. The x-axis unit is in rad/s and represents $2\pi T_s f$, with f the inverse time-scale and T_s the time-step used in modelling. **b**, same as (a) but with T_{sw}/T_s = 4. In this case, the fast time-scale range of the spectrum was suppressed, but the spectrum remained intact across the slow time-scales (x-axis is log-scaled to emphasize intact spectrum at slow time-scales). **c**, Same as (b) but with T_{sw}/T_s = 6. In this case, the fast time-scale ranges of the spectrum were further suppressed. This design parameter allows the experimenter to focus the modelling on the time-scales of interest for the brain network response. We found that while we kept the MN spectrum white for T_{sw}/T_s = 1 and suppressed the fast-time scale range of the MN spectrum for T_{sw}/T_s = 4 and 6, the IO prediction accuracy corresponding to different T_{sw}/T_s 's (1, 4, and 6) were not significantly different (Kruskal-Wallis test, P = 0.1361), confirming that the LFP response to stimulation mainly happened at slow time-scales (see Supplementary Note 14 and Supplementary Fig. 15).



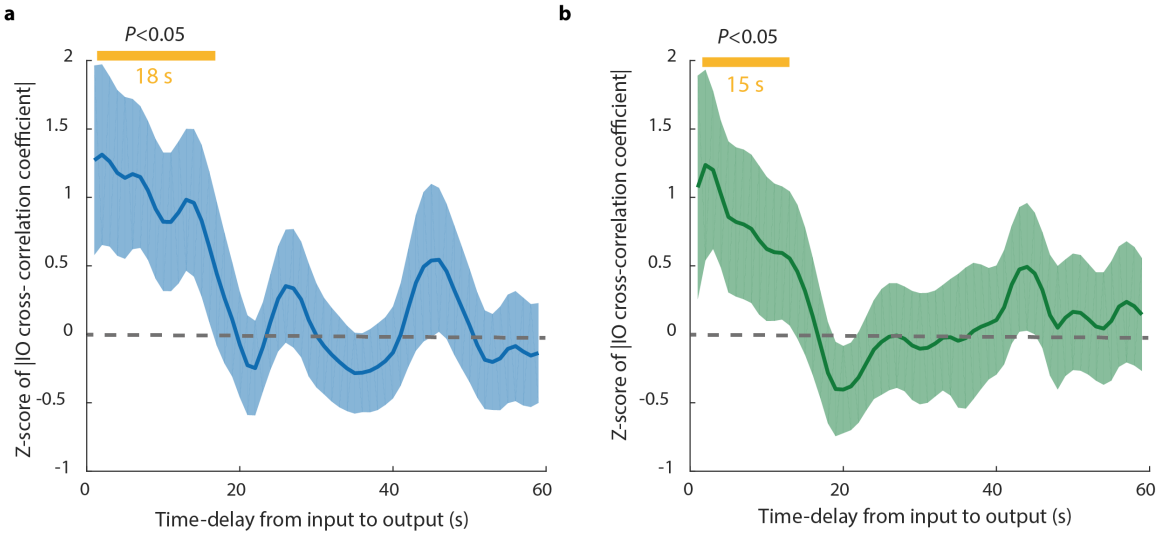
Supplementary Fig. 15 | Time-scale of the input-driven dynamics in response to stimulation amplitude and frequency.

a, examples of the frequency domain transfer function from the stimulation amplitude to the output LFP power feature. The solid line represents the energy of the transfer function (i.e., the magnitude of the squared transfer function) from the stimulation amplitude to the output LFP power feature. The upper limit of x-axis represents the fastest time-scale examined in this analysis (see Supplementary Note 14). The dashed line represents the time-scale slower than which 80% of the energy of the transfer function is concentrated at. This time-scale is defined as the effective time-scale (ETS). This example shows that the energy of the transfer function concentrated within slow time-scales (see Supplementary Note 14). **b**, same as (**a**) but for stimulation frequency. **c**, the distribution of the ETS of the input-driven dynamics in response to stimulation amplitude (top) and frequency (bottom). The average ETS across LFP power features was 19.89 ± 0.58 s for stimulation amplitude and 19.52 ± 0.56 s for stimulation frequency (mean ± s.e.m.). The average ETS was much larger than the fastest time-scale examined in this analysis, which was 10 s or 5 s, again showing that the energy of the transfer function concentrated within slow time-scales (see Supplementary Note 14).



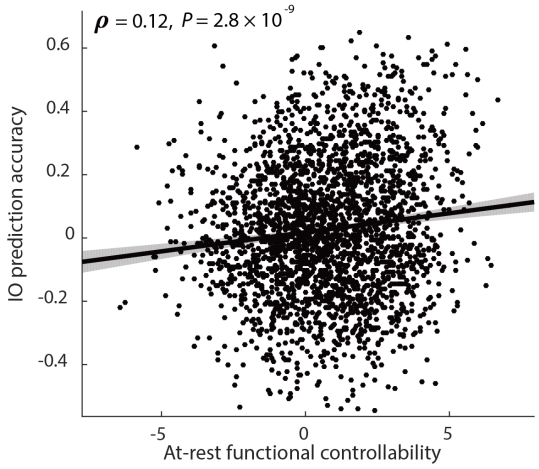
Supplementary Fig. 16 | Computation of effective time-scale is robust to the amount of training data used for fitting the IO models as long as the training data length is longer than the effective time-scale.

a, box plots of the effective time-scale of input-driven dynamics in response to stimulation amplitude. Each boxplot shows the results computed from IO models fitted using different percentages of the IO data. In the 16 IO datasets (Supplementary Table 3), the average amount of IO data per dataset was 3055 s (s.e.m. was 576 s). We used different percentages of the IO data in each dataset to fit the IO model (i.e. 50%, 75%, 87.5%, and 93.75%). We then computed the effective time-scale from the fitted IO model. Note that the shortest length of training data included in this analysis for a dataset was on average 1528 s (s.e.m. was 288 s) (corresponding to 50% of IO data). Even this shortest length was much longer than the computed effective time-scales (maximum effective time-scale < 60 s, see y axis). Red line represents median. Box lines represent 25% and 75% quantiles. Whiskers represent 5% and 95% quantiles. N = 233 independent samples, i.e., LFP power features, are included in each box. Kruskal-Wallis test P value is also shown. n.s.: not significant. **b**, same as (**a**) but for stimulation frequency.



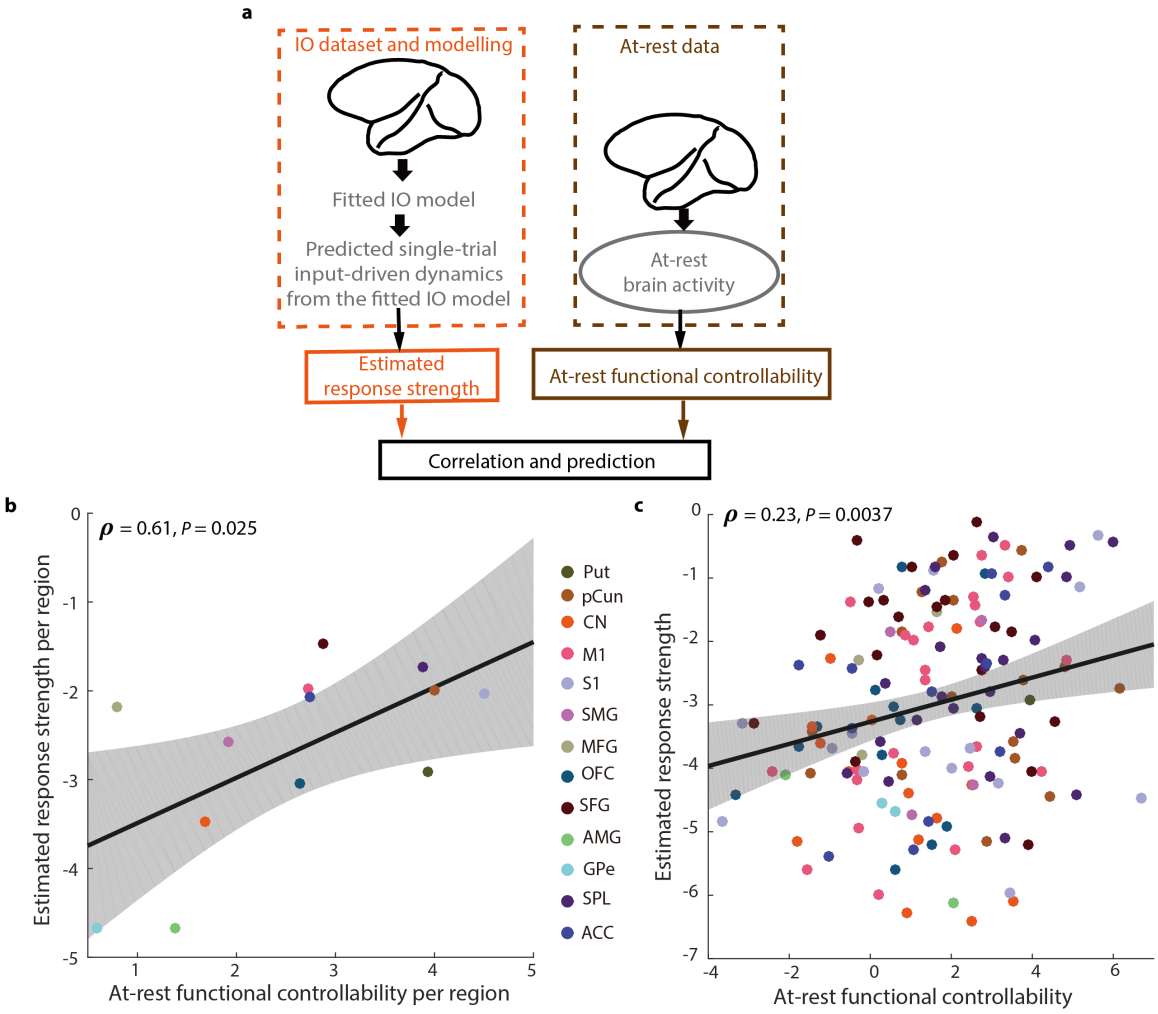
Supplementary Fig. 17 | Input-output (IO) cross-correlation between the stimulation amplitude/frequency (input) and the ground-truth single-trial input-driven dynamics (output).

a, the Z-scored absolute value of the IO cross-correlation coefficient between the input stimulation amplitude and the output. Z-scoring is done based on the distribution of chance level absolute values of the cross-correlation coefficient, which are computed between randomly generated MN input and the same output as in the actual IO dataset. Line represents mean and shaded area represents 95% confidence interval. The mean value of the chance level is zero and indicated by dashed grey line. IO cross-correlation coefficients that are significantly larger than the chance level after FDR correction for multiple time-delays are indicated by the horizontal orange bar on top (FDR-corrected random-test $P < 0.05$, see Supplementary Note 15). Also, the time-delays that had a significant IO cross-correlation coefficient are indicated below the orange bar. The absolute value of the IO cross-correlation coefficient was significantly larger than the chance level at time-delays from $\tau = 1\ s$ to $\tau = 18\ s$, showing that the input stimulation amplitude significantly correlated with the output at time-delays up to 18 s. This maximum time-delay for significant IO cross-correlation was close to the effective time-scale of the effect of stimulation amplitude on the output, which was 19.89 ± 0.58 s (mean ± s.e.m., Supplementary Fig. 15c). For time-delays larger than the maximum time-delay with a significant IO cross-correlation coefficient (end of orange bar), the confidence interval of the IO cross-correlation coefficient contains or is very close to the chance level of zero, and the IO cross-correlation coefficient is not significantly different from chance after FDR correction (FDR-corrected random-test $P > 0.05$, see Supplementary Note 15). **b**, same as (**a**) but for stimulation frequency. The absolute value of the IO cross-correlation coefficient was significantly larger than the chance level at time-delays from $\tau = 1s$ to $\tau = 15s$. The maximum time delay of 15 s was again close to the effective time-scale of the effect of stimulation frequency on the output, which was 19.52 ± 0.56 s (Supplementary Fig. 15c).



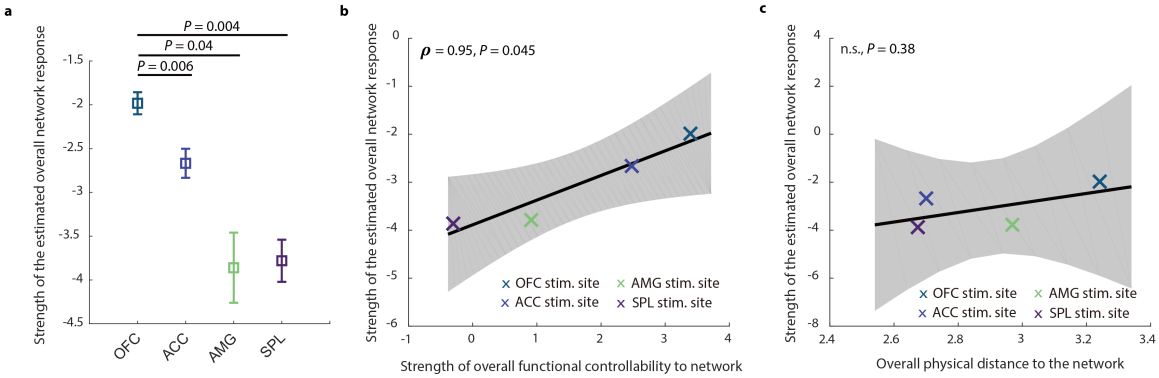
Supplementary Fig. 18 | At-rest functional controllability correlates with the IO prediction accuracy across all modeled LFP power features.

Across all modeled LFP power features (regardless of whether they were predictable or not), at-rest functional controllability is still significantly correlated with the IO prediction accuracy. Each dot corresponds to one LFP power feature (N = 2640 independent samples, i.e., LFP power features). The least-squares fitted linear line is shown as solid line and its 95% confidence bound shown as the shaded area. The linear correlation coefficient ρ and Pearson's P value are shown on the top left.



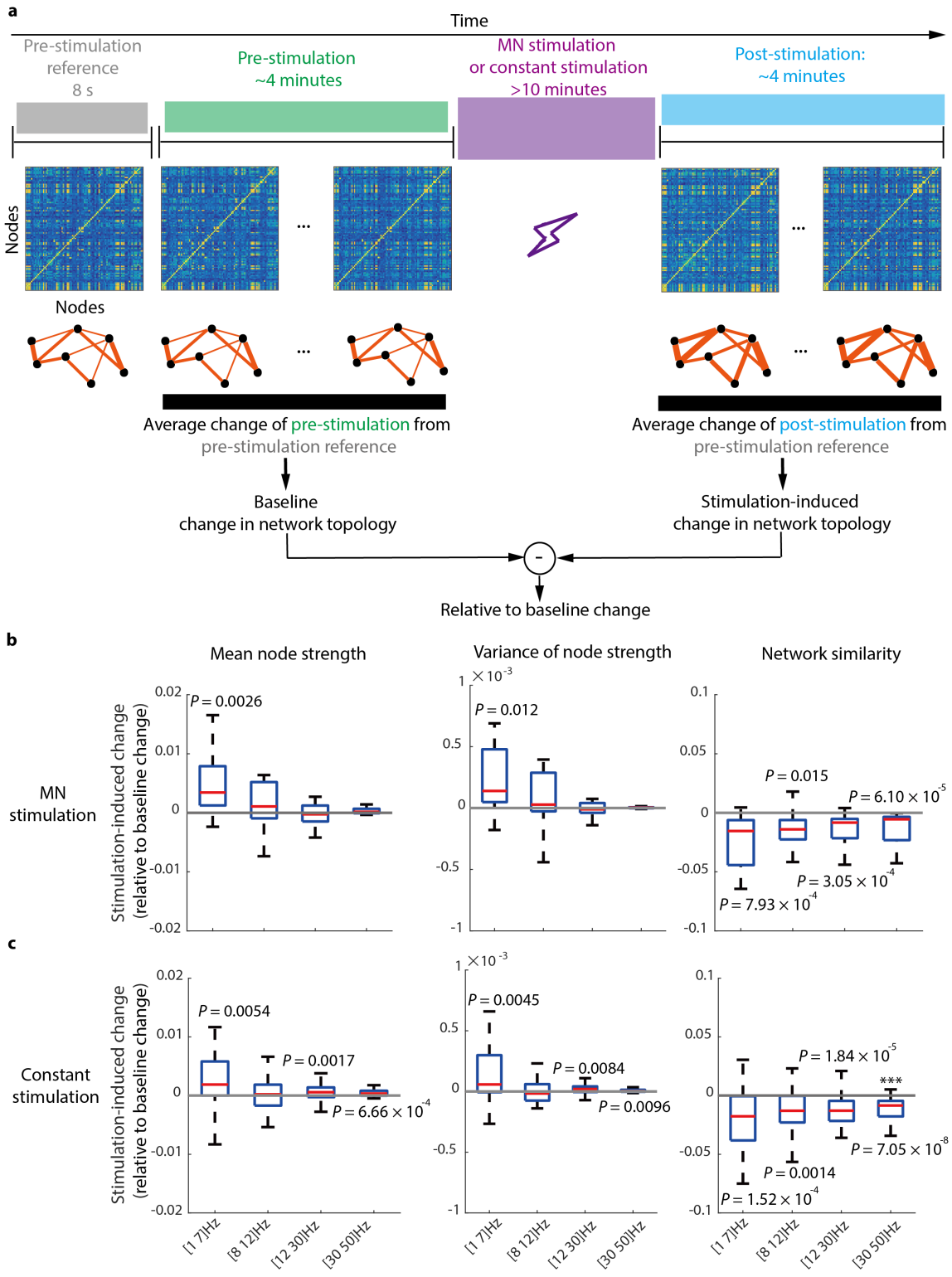
Supplementary Fig. 19 | At-rest functional controllability explains the variability in the estimated response strength at different network nodes.

a, at-rest functional controllability was calculated from at-rest LFP data without stimulation. We computed the average energy in the temporal variations of the predicted single-trial input-driven dynamics from the IO model to obtain the estimated response strength for OFC stimulation. **b**, at-rest functional controllability of each brain region significantly correlated with the estimated response strength at that region (in total N = 13 independent samples, i.e., brain regions). The least-squares fitted linear line is shown as solid line and its 95% confidence bound shown as the shaded area. The linear correlation coefficient ρ and Pearson's P value are shown on the top left. See Supplementary Table 1 for brain region abbreviations. **c**, at-rest functional controllability significantly correlated with estimated response strength across LFP power features. Dots present raw power feature data (in total N = 153 independent samples, i.e., LFP power features) and are color-coded according to the brain region from which they are recorded. The least-squares fitted linear line is shown as solid line and its 95% confidence bound shown as the shaded area. The linear correlation coefficient ρ and Pearson's P value are shown on the top left. Color codes are the same as (**b**).



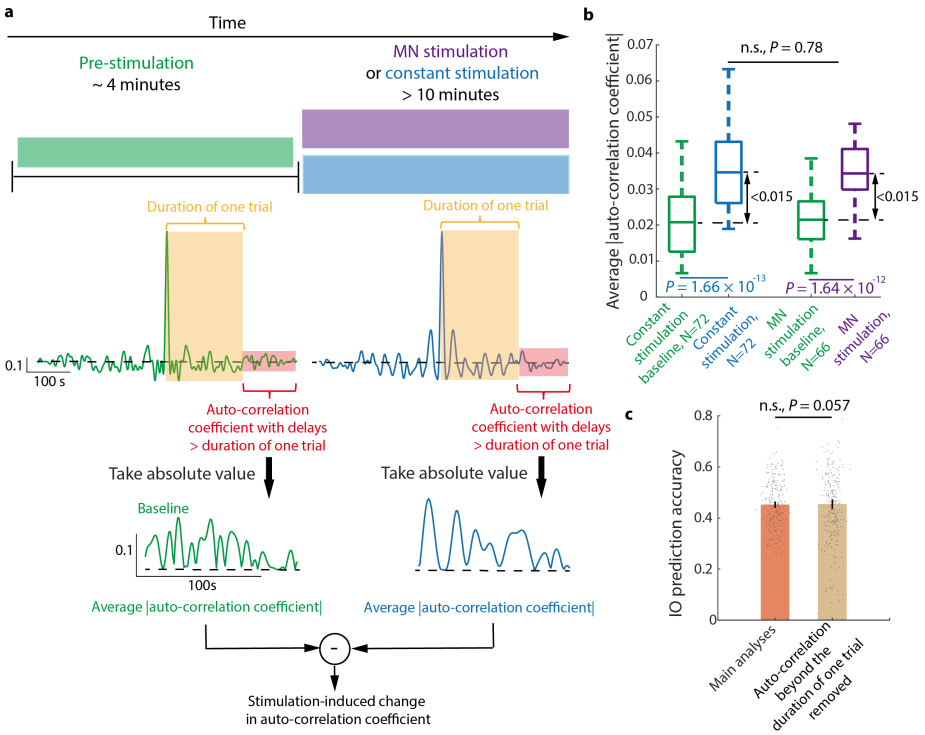
Supplementary Fig. 20 | Across all four stimulation sites, the strength of the estimated overall network response to a stimulation site correlates with the strength of the overall functional controllability from that site to the rest of the network, but not with the overall physical distance.

a, the strength of the estimated overall network response to OFC stimulation was significantly larger than the other three sites. The square represents mean and the whisker represents s.e.m. The sample sizes *N* for OFC, ACC, AMG, SPL are 153, 63, 3, and 8, respectively. Kruskal-Wallis test *P* values are also shown. **b**, the strength of the estimated overall network response to a stimulation site correlated significantly with the overall functional controllability from that site to the rest of the network (*N* = 4 independent samples, i.e., stimulation sites). The least-squares fitted linear line is shown as solid line and its 95% confidence bound shown as the shaded area. The linear correlation coefficient *ρ* and Pearson's *P* value are shown on the top left. The stimulation site is colored coded (same color code as Fig. 3). Crosses are mean values. **c**, the strength of the estimated overall network response to a stimulation site did not correlate significantly with the overall physical distance from that site to the rest of the network (*N* = 4 independent samples, i.e., stimulation sites). The least-squares fitted linear line is shown as solid line and its 95% confidence bound shown as the shaded area. The linear correlation coefficient *ρ* and Pearson's *P* value are shown on the top left. n.s.: not significant. Color code is the same as (**b**).


a
Time
Pre-stimulation reference 8 s
Pre-stimulation ~4 minutes
MN stimulation or constant stimulation >10 minutes
Post-stimulation: ~4 minutes
Nodes
Nodes
Average change of pre-stimulation from pre-stimulation reference
Average change of post-stimulation from pre-stimulation reference
Baseline change in network topology
Stimulation-induced change in network topology
Relative to baseline change
b
Mean node strength
Variance of node strength
Network similarity
MN stimulation
Stimulation-induced change (relative to baseline change)
P = 0.0026
P = 0.012
P = 0.015
P = 6.10 × 10-5
P = 3.05 × 10-4
P = 7.93 × 10-4
c
Constant stimulation
P = 0.0054
P = 0.0017
P = 6.66 × 10-4
P = 0.0045
P = 0.0084
P = 0.0096
P = 1.84 × 10-5
***
P = 7.05 × 10-8
P = 0.0014
P = 1.52 × 10-4
[1 7]Hz
[8 12]Hz
[12 30]Hz
[30 50]Hz

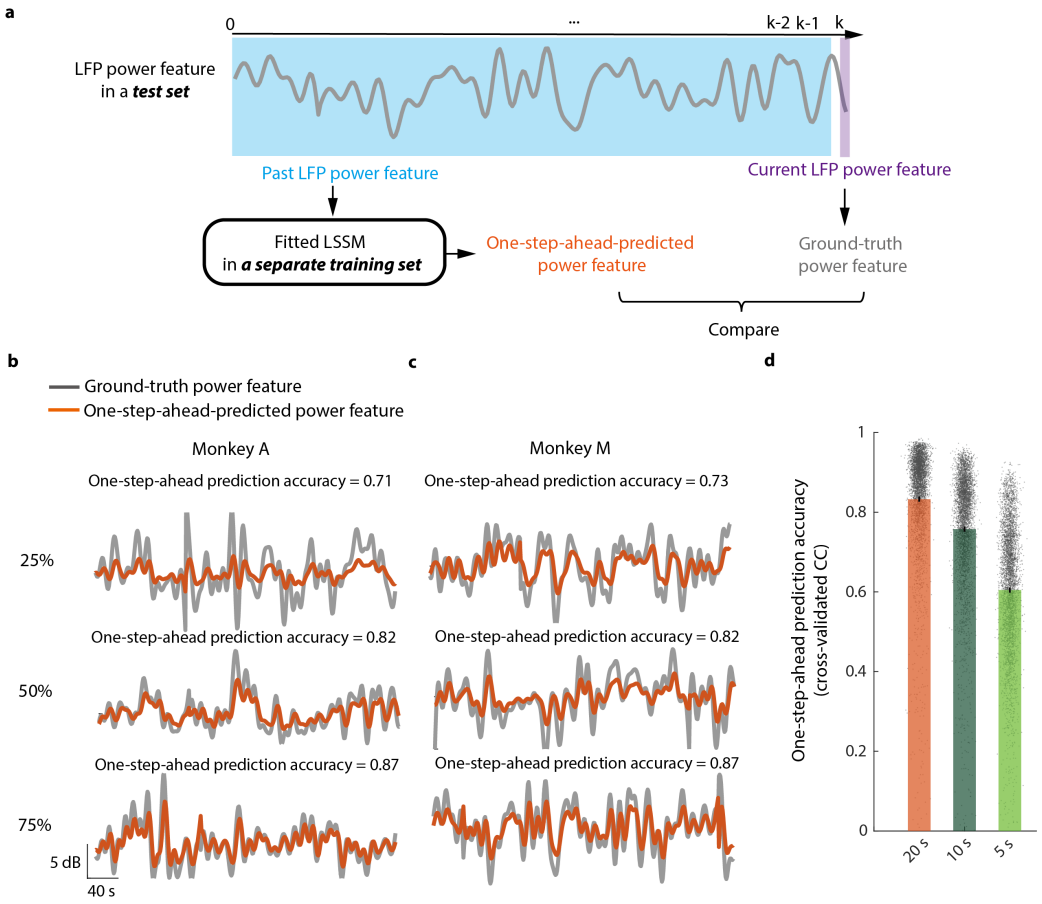
Supplementary Fig. 21 | Functional brain network topology adapts/changes by a relatively small amount after stimulation ends in both MN and constant stimulation datasets.

a, computation of coherence networks at different time periods (see Supplementary Note 18 for details). Stimulation-induced changes in network topology are relative to baseline changes that are computed purely using pre-stimulation data. Difference between the stimulation-induced changes and the baseline changes is shown as the relative change. **b**, change in network topology measures relative to baseline (baseline is shown as zero) for MN stimulation datasets. In each panel, grey horizontal line represents no change from baseline. Red line represents median. Box lines represent 25% and 75% quantiles. Whiskers represent 5% and 95% quantiles. N = 16 independent samples, i.e., MN stimulation datasets (see Supplementary Table 3), are included in each box. Two-sided Wilcoxon signed-rank test P values that are significant after FDR correction are also shown. **c**, same as (**b**) but for constant stimulation datasets and with the same boxplot convention as in (**b**). In each panel, grey horizontal line represents no change from baseline. N = 48 independent samples, i.e., constant stimulation datasets (see Supplementary Table 4), are included in each box. Two-sided Wilcoxon signed-rank test P values that are significant after FDR correction are also shown.



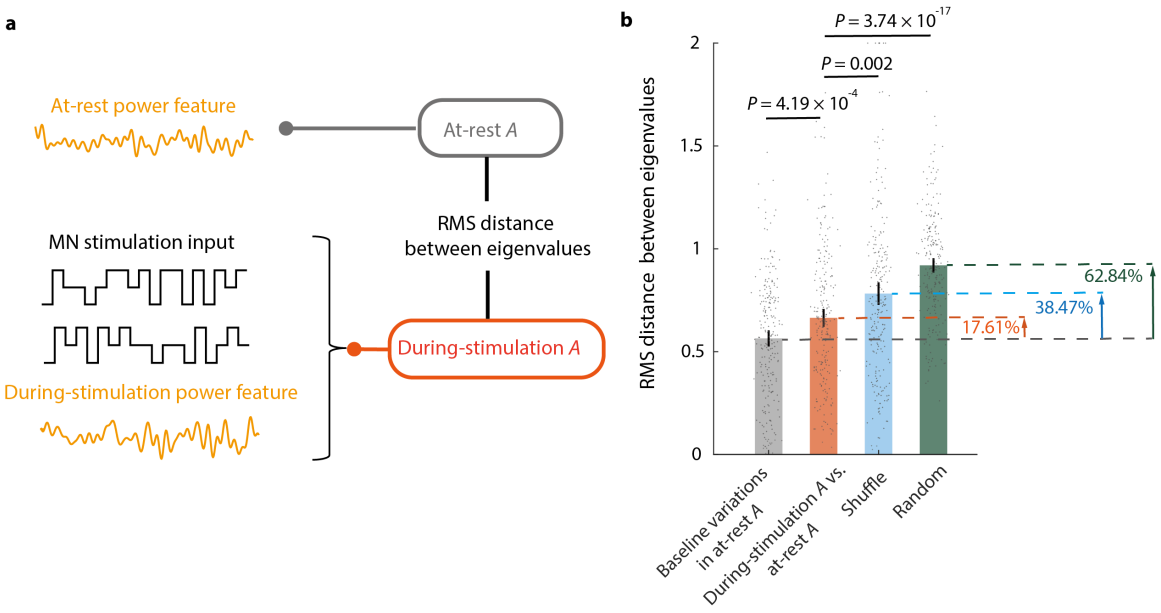
Supplementary Fig. 22 | Stimulation induces relatively small changes/adaptations in auto-correlation of LFP power features during stimulation but does not significantly influence the computed IO prediction accuracy in cross-validation.

a, computation of auto-correlation coefficients of LFP power features at delays longer than the duration of the MN experiment trial for during stimulation data and for at-rest data (as baseline) (see Supplementary Note 19 for details). We compute the average absolute value of the auto-correlation coefficients both for during stimulation data and for at-rest data and take the difference between these two average values as the stimulation-induced change in the auto-correlation coefficient. **b**, in MN stimulation datasets, 28% of predictable LFP power features showed a significant stimulation-induced change in the auto-correlation coefficient. In constant stimulation datasets, 31% of predictable LFP power features showed a significant stimulation-induced change in the auto-correlation coefficient. For these predictable power features that show a significant stimulation-induced change in the auto-correlation coefficient, the statistics of the absolute values of the auto-correlation coefficients for during stimulation data and for at-rest data are shown. Median of the stimulation-induced change in the auto-correlation coefficient for these features is shown by vertical arrows. Red line represents median. Box lines represent 25% and 75% quantiles. Whiskers represent 5% and 95% quantiles. The sample number *N*, i.e., number of power features that show a significant stimulation-induced change, is indicated for each box. In comparing with baseline, two-sided Wilcoxon signed-rank test *P* values are shown (blue and purple numbers). In comparing MN stimulation and constant stimulation, two-sided Wilcoxon rank-sum test *P* value is shown (black number). n.s.: not significant. **c**, IO prediction accuracy in cross-validation when autocorrelations for delays that are longer than the duration of the MN experiment trial are removed. The bar represents the mean, and the black error bar represents s.e.m. Raw IO prediction accuracies are shown with dots (*N* = 233 independent samples, i.e., LFP power features, for both bars). Two-sided Wilcoxon signed-rank test P values are also shown. n.s.: not significant.



Supplementary Fig. 23 | The at-rest LSSM can predict the intrinsic dynamics of at-rest LFP power features.

a, one-step-ahead prediction in the test set. Based on the at-rest LSSM fitted in a separate training set, one-step-ahead prediction uses the past at-rest LFP power features to predict the current at-rest LFP power feature in the test set (Supplementary Note 11). **b**, three example one-step-ahead predictions of intrinsic dynamics of at-rest LFP power features in the test set using the fitted at-rest LSSM in the training set in Monkey A. The one-step-ahead prediction accuracy of the three examples are around the 25% (top), 50% (middle) and 75% (bottom) quantiles of the distribution of one-step-ahead prediction accuracy (cross-validated correlation coefficient (CC)) across all at-rest LFP power features. **c**, same as (**b**) but for Monkey M. **d**, the prediction of intrinsic dynamics was robust to the choice of time-scales considered. Intrinsic dynamics can be predicted at time-scales even twice (dark green) or four times (light green) faster than the effective time-scale (orange) of the dynamical effect of stimulation. The bar represents the mean, and the black error bar represents s.e.m. Raw IO prediction accuracies are shown with dots (N = 4780 independent samples, i.e., at-rest LFP power features, for all bars). Across all the at-rest LFP power features, the one-step-ahead prediction accuracy (CC) was 0.83 ± 0.006 by using the effective time-scale (i.e. 20 s and slower), which was significantly larger than chance level (FDR-corrected output-permutation test $P < 10^{-16}$ for all at-rest LFP power features). For time-scales that were even twice and four times faster, the one-step-ahead prediction accuracies were 0.76 ± 0.006 and 0.61 ± 0.006 (mean ± s.e.m.), respectively, and significantly larger than chance level (FDR-corrected output-permutation test $P < 10^{-16}$ for all at-rest LFP power features in both cases, corresponding to time-scale of 10 s and slower, or 5s and slower, respectively).



Supplementary Fig. 24 | At-rest dynamics constrain during-stimulation dynamics.

a, for the same predictable power feature, at-rest A was fitted from at-rest data without stimulation and during-stimulation A was fitted from MN stimulation data. The root mean square (RMS) distance between the eigenvalues of at-rest A and during-stimulation A quantifies the difference between the two state transition matrices. **b**, difference between the state transition matrix fitted with at-rest data (at-rest A) and the state transition matrix fitted with during-stimulation data (during-stimulation A) was relatively small for the same LFP power feature. RMS distance between the eigenvalues of various state transition matrices are shown in the bar plot. The bar represents the mean, and the black error bar represents s.e.m. Raw IO prediction accuracy is shown with dots (N = 233 independent samples, i.e., LFP power features, for all bars). To control for the effect of noise in fitting the LSSM parameters, as baseline, we compute the baseline variations in the at-rest A using at-rest LFP data. To do so, we fit two at-rest A's, one using the first half and one using the second half of at-rest LFP data. We then compute the difference between these two at-rest A's and use this difference as the baseline variations in at-rest A. Shuffle represents the statistics of the difference between the at-rest A of any LFP power feature and the during-stimulation A of another LFP power feature. Random represents the statistics of the difference between a random stable A and the at-rest A. The dashed line and arrows represent the average percentage increase from the baseline for each bar. The difference between the during-stimulation A and the at-rest A for the same LFP power feature was significantly smaller than the shuffle difference (two-sided Wilcoxon signed-rank test, P = 0.002) and the random difference (two-sided Wilcoxon signed-rank test, $P = 3.74\times10^{-17}$). The two-sided Wilcoxon signed-rank test P values are also shown in the figure. Also, the percentage increase of the shuffle difference from the baseline was 38.47% and the percentage increase of the random difference from the baseline was 62.84%; both of these percentages were more than twice as large as that for the difference between the during-stimulation A and at-rest A for the same LFP power feature (17.61%). This result shows that the at-rest A constrains the during-stimulation A because the during-stimulation A is much more similar to the at-rest A than it is to a random A or to the A of another LFP power feature.

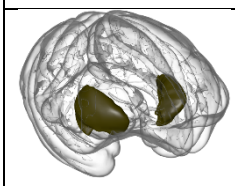
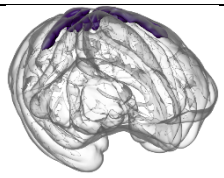
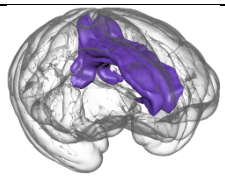
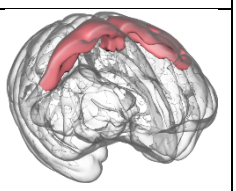
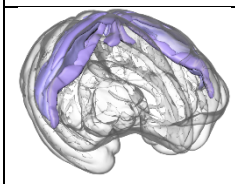
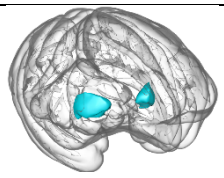
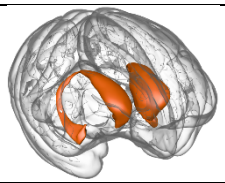
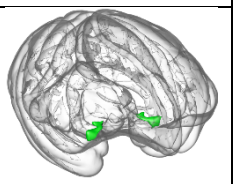
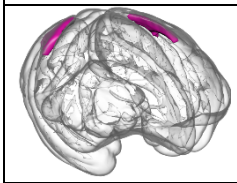
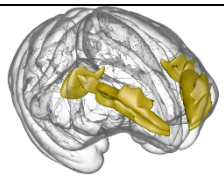
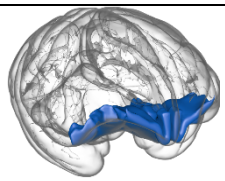
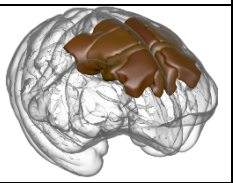
Supplementary Tables

Supplementary Table 1 | Brain region abbreviations.

Brain region	Abbreviation
Orbitofrontal cortex	OFC
Anterior cingulate cortex	ACC
Posterior cingulate cortex	PCC
Superior parietal lobule	SPL
Supramarginal gyrus	SMG
Middle frontal gyrus	MFG
Superior frontal gyrus	SFG
Precentral gyrus	M1
Postcentral gyrus	S1
Amygdala	AMG
Caudate nucleus	CN
External globus pallidus	GPe
Putamen	Put
Precuneus	pCun

Supplementary Table 2 | Anatomical locations of brain regions.

Each brain region included in Fig. 3 is shown on a template monkey brain[120]. The colors of the anatomical region in the template monkey brain are the same as the color codes for brain regions in Fig. 3. Note that for better visualization on the template monkey brain, ACC and PCC are pooled as one single region covering cingulate. Also, pCun and SPL are pooled as one single region covering parietal cortex. See Supplementary Table 1 for brain region abbreviations.

Put	SPL + pCun	ACC + PCC	M1
			
S1	GPe	CN	AMG
			
SMG	MFG	OFC	SFG
			

Supplementary Table 3 | IO Dataset information (continue to next page).

Note that for each IO dataset $T = L \times N_R$ is the total number of data samples used for IO modelling. See Supplementary Table 1 for brain region abbreviations.

IO Dataset	Subject	Stim. site	# of trials	Stim. duration per trial (s)	MN levels	Basic switching time (T_{sw}, s)	Time step used in modelling (T_s, s)	$\frac{T_{sw}}{T_s}$	Number of data samples per trial (L)	Number of modeled power features after preprocessing (N_y)	Number of modeled channels after preprocessing ($N_y/4$)	Figure panels that use data from this dataset	Supplementary Figure panels that use data from this dataset
1	Monkey A	OFC	10	180	3	4	1	4	176	356	89	2b,c,3a, 4b-d,f, 5, 6a-c, 8c-d	2g, 7a, 8-11,12b, 13, 15-22, 23d, 24
2	Monkey A	OFC	10	180	3	1	1	1	176	356	89	2a, b, c, 3 4b-f, 5, 6b-c, 8b-d	2e,f,g, 7a, 8-11,12a-b, 13, 15-22, 23b,d, 24
3	Monkey A	SPL	20	270	3	6	1	6	176	248	62	2b,c, 3a, 4a-d,f 6b-c, 8c-d	2g, 7a, 8-11, 12a-b, 13, 15-17, 20-22, 23d, 24
4	Monkey A	ACC	20	270	3	6	1	6	256	240	60	2b,c,3a, 4b-f, 6b-c, 8c-d	2g, 7a, 8-11, 12b, 13, 15-17, 20-22, 23b,d, 24
5	Monkey A	OFC	20	270	3	6	1	6	256	292	73	2b,c, 4b,c,f, 5, 6b-c, 8c-d	2g, 8-11, 12b, 13, 15-22, 23d, 24
6	Monkey A	ACC	20	180	3	4	1	4	176	288	72	2b,c,3a 4a-c,f 6b-c, 8b-d	2g, 7a, 8-11, 12b, 13, 15-17, 20-22, 23d, 24
7	Monkey A	OFC	20	180	3	4	1	4	176	304	76	2b,c, 4b,c,e,f, 5, 6b-c, 8c-d	2g, 8-11, 12b, 13, 15-22, 23d, 24
8	Monkey A	AMG	20	180	3	1	1	1	176	300	75	2b,c, 4b,c,e,f, 5, 6b-c, 8c-d	2g, 8-11, 12b, 13, 15-17, 20-22, 23d, 24

IO Dataset	Subject	Stim. site	# of trials	Stim. duration per trial (s)	MN levels	Basic switching time (T_{sw}, s)	Time step used in modelling (T_s, s)	$\frac{T_{sw}}{T_s}$	Number of data samples per trial (L)	Number of modeled power features after preprocessing (N_y)	Number of modeled channels after preprocessing ($N_y/4$)	Figure panels that use data from this dataset	Supplementary Figure panels that use data from this dataset
9	Monkey A	OFC	20	180	3	1	1	1	176	192	48	2b,c, 4a-c, f, 5, 6b-c, 8c-d	2g, 8-11, 12b, 13, 15-22, 23d, 24
10	Monkey A	OFC	30	240	2	1	1	1	240	288	72	2b,c, 4b-d, f, 5, 6b-c, 8c-d	2g, 8-11, 12b, 13, 15-22, 23d, 24
11	Monkey A	OFC	30	240	2	1	1	1	240	224	56	2b,c, 4b,c,f, 5 6b-c, 8c-d	2g, 8-11, 12b, 13, 15-22, 23d, 24
12	Monkey A	OFC	10	60	2	0.5	0.5	1	112	264	66	2b,c,e 4b,c,f, 5 6b-c, 8c-d	2g, 7a, 8-11, 12b, 13, 15-22, 23d, 24
13	Monkey A	AMG	10	60	2	0.5	0.5	1	112	268	67	2b,c, 4b,c,f, 6b-c, 8c-d	2g, 8-11,12b, 13, 15-17, 20-22, 23d, 24
14	Monkey A	OFC	10	60	2	0.5	0.5	1	112	220	55	2b,c,e 4b,c,f, 5, 6b-c, 8c-d	2g, 7a, 8-11, 12b, 13, 15-22, 23d, 24
15	Monkey M	OFC	10	60	2	0.5	0.5	1	112	144	44	2b,d,f, 4a-c,f, 5, 6a-c, 8c-d	2g, 7b, 8-11, 12a-b, 13, 15-22, 23c,d, 24,
16	Monkey M	ACC	10	60	2	0.5	0.5	1	112	196	105	2b,d,f, 4b,c,f, 6b-c, 8c-d	2g, 7b, 8-11, 12a-b, 13, 15-17, 20-22, 23c,d, 24

Supplementary Table 4 | Constant stimulation dataset information.

The dataset number (first column) is in the same order as in the MN stimulation datasets (Supplementary Table 3). The structure of these constant stimulation datasets was: 5 minutes pre-stimulation at-rest data + 10 minutes constant stimulation data + 5 minutes post-stimulation at-rest data

Datasets	Subject	Stim. site	Constant stimulation frequency (Hz)	Three possible constant stimulation amplitudes (μA)	Supplementary figures that use data from this dataset
1	Monkey A	OFC	100	15, 22, 30	21 - 24
2	Monkey A	OFC	100	15, 22, 30	21 - 24
3	Monkey A	SPL	100	15, 22, 30	21 - 24
4	Monkey A	ACC	100	15, 22, 30	21 - 24
5	Monkey A	OFC	100	15, 22, 30	21 - 24
6	Monkey A	ACC	100	15, 22, 30	21 - 24
7	Monkey A	OFC	100	15, 22, 30	21 - 24
8	Monkey A	AMG	100	15, 22, 30	21 - 24
9	Monkey A	OFC	100	15, 22, 30	21 - 24
10	Monkey A	OFC	100	10, 30, 50	21 - 24
11	Monkey A	OFC	100	10, 30, 50	21 - 24
12	Monkey A	OFC	100	10, 30, 50	21 - 24
13	Monkey A	AMG	100	10, 30, 50	21 - 24
14	Monkey A	OFC	100	10, 30, 50	21 - 24
15	Monkey M	OFC	100	10, 30, 50	21 - 24
16	Monkey M	ACC	100	10, 30, 50	21 - 24

.

Supplementary Notes

Supplementary Note 1. Stimulation artifact rejection

The stimulation artifact typically consists of an “instantaneous artifact spike” and an “artifact tail”[17,103,104]. The “instantaneous artifact spike” immediately following the stimulation pulse can have an amplitude of more than 10 times larger than at rest raw signal (hundreds vs. tens of microvolts) and the “artifact tail” usually has a consistent shape lasting tens of milliseconds following each “instantaneous artifact spike”[17,103,104]. Both types of artifacts can introduce artificial power changes in neural signals. Our stimulation artifact rejection algorithm removes both types of artifacts using four steps described below (Supplementary Fig. 2).

First, we detected the timings of the stimulation artifact by thresholding the broadband raw signal recorded from the stimulation site and detecting its peaks. The threshold was selected based on visual examination.

Second, based on these detected timings, we used a temporal template subtraction method to remove the “artifact tail”[103] as follows: 1) based on visual inspection, we used a window of 1.17 ms (starting from 0.67 ms before the timing of the current artifact to 0.5 ms after the timing of the current artifact) to extract the “instantaneous artifact spike”. We then used a window following this “instantaneous artifact spike” to extract the “artifact tail”, i.e., a window starting from 0.5 ms after the timing of the current artifact to 0.67 ms before the timing of the next artifact (Supplementary Fig. 2b); 2) we collected all “artifact tails” and categorized them according to its stimulation amplitude and frequency, e.g., for a 2-level MN stimulation, we had four categories of “artifact tail” corresponding to $\{(20\mu A, 50Hz), (40\mu A, 50Hz), (20\mu A, 100Hz), (40\mu A, 100Hz)\}$); 3) for each category of “artifact tail”, we calculated the mean of the “artifact tails” belonging to that category, and took it as the “artifact tail” template for that category (Supplementary Fig. 2b); 4) we subtracted the corresponding template from each “artifact tail” (Supplementary Fig. 2c).

Third, we rejected the “instantaneous artifact spike” by interpolation as follows: 1) we discarded the “instantaneous artifact spike” portion (the period starting 0.67 ms before the timing of the artifact and ending 0.5 ms after the timing of the artifact, see previous paragraph) of the raw data and treated this portion as missing data since this portion of the raw data was typically not useful, e.g., because of saturation[17,103,104] (Supplementary Fig. 2c). The proportion of the duration of this missing data over the entire stimulation duration was less than 9%; 3) we interpolated the above portion of missing data using the raw signals immediately before the “instantaneous artifact spike” window and immediately after the “instantaneous artifact spike” window. In particular, we took the 0.67 ms raw data immediately before the “instantaneous artifact spike” window to fill in the first 0.67 ms of the missing data, and similarly took the 0.5 ms raw data immediately after the “instantaneous artifact spike” window to fill in the last 0.5 ms of the missing data (Supplementary Fig. 2c). We used this interpolation instead of linear/spline interpolation because in this way the amplitude and spectral distribution of the interpolated data was more similar to that of background raw signals[17].

Fourth, to show that the above stimulation artifact rejection algorithm only introduces a negligible distortion of the underlying signal spectrum, we conducted a control analysis. For each channel, we repeatedly concatenated the pre-stimulation raw data enough times to form the “at-rest control raw data” that had the same length as the during-stimulation raw data. Then we used the artifact timings identified from the actual during-stimulation raw data to add artifacts to the “at-rest control

raw data" and then repeat the same stimulation artifact rejection algorithm on the artifact-added "at-rest control raw data". More specifically, for each actual artifact timing, we replaced the "instantaneous artifact spike" window of the "at-rest control raw data" with the "instantaneous artifact spike" from the actual during-stimulation raw data, and replaced the "artifact tail" window of the "at-rest control raw data" with the fitted "artifact tail" template computed from the actual during-stimulation raw data. Next, we used the same artifact removal algorithm on this artifact-added "at-rest control raw data", where the added artifact is treated as unknown to the artifact removal algorithm. This control analysis accounts for both the effect of interpolation during the spike (missing data) and the effect of artifact tail fitting and subtraction on the "at-rest control raw data". Overall, this analysis allowed us to assess the effect of the artifact rejection algorithm because in this case we knew the ground truth PSD of the "at-rest control raw data" before the algorithm was applied to it as this data had no artifacts to begin with. We compared the ground-truth PSD of the "at-rest control raw data" and its PSD after the artifact rejection algorithm was applied to it (Supplementary Fig. 2f). We found that across all the 1020 recording channels used in modelling, the relative difference of the two was < 5% for frequency ranges below 50Hz (Supplementary Fig. 2g). This result shows that the artifact rejection algorithm only introduced a negligible distortion of neural signal PSD below 50Hz, which covered the frequencies of interest in this study. Further, as a control analysis, we showed that even for the higher gamma band 70-100 Hz with larger distortions of neural signal PSD (Supplementary Fig. 2g), after controlling the effect of stimulation artifacts, our IO modelling still successfully predicted the response at this band (Supplementary Fig. 10).

Supplementary Note 2. Computation of the LFP power features: choice of the frequency band, time step and power extraction method

Choice of the frequency band

We focus our analyses on the four frequency bands below 50 Hz—i.e., 1–8 Hz (delta+theta), 8–12 Hz (alpha), 12–30 Hz (beta), 30–50 Hz (low-gamma)—because of their important roles in many brain functions and behaviors such as somatosensory processing[121], memory[122], attention[123], learning[124], mood processing[13,22] and movement[1,43,106–109]. These four bands have also been implicated in many brain dysfunctions: for example, subthalamic nucleus (STN) beta band power has been used as a biomarker for Parkinson's disease that correlates with bradykinesia and rigidity levels[125,126]; orbitofrontal cortex (OFC) theta and alpha powers have been shown to be correlated with mood in patients with moderate to severe depression traits[13]; thalamic theta and alpha powers have been shown to be related to the pathophysiology of tics in Tourette syndrome[127]. These lower frequency bands could thus serve as neural biomarkers to be controlled within a closed-loop neuromodulation system, for example for control of rigidity level in Parkinson's disease[20], control of mood state in depression[1,2] and control of motor tics in Tourette syndrome[2]. In addition, we test our IO models in predicting the response at the higher frequency gamma band of 70-100 Hz given that higher frequency gamma band can also show changes in many brain functions and behaviors[128]. Finally, given that typically LFP spectral features rather than the LFP waveform are used to study brain functions and to provide biomarkers for brain disorders as explained above, in this work we focus on the IO modelling of these spectral features.

Choice of the time step

We used a time step T_s of 0.5 s or 1 s to compute the LFP power features for two main reasons as expanded on below.

First, many internal brain states that are likely the target for closed-loop control in many neuropsychiatric applications have dynamics on the order of seconds to minutes. Slow dynamics of neural features on the order of seconds to minutes have been previously shown to be relevant to internal brain states such as affective states[13,22,23], exploratory and defensive states[129], motivational and homeostatic states (e.g., thirst[130]), and arousal[131], and to be relevant during other spontaneous behaviors without specific external sensory inputs[132]. Thus, such slow dynamics are likely to become the target for many neuromodulation systems that aim to modulate brain functions/dysfunctions. For example, the slow-changing dynamics relevant to affective and mood states[13,22,23] are likely to become the target for closed-loop stimulation therapies for many neuropsychiatric disorders. That is why here we focus on IO modelling of the dynamics of LFP power features computed with a time step on the order of seconds.

Second, we need to use time steps on the order of seconds because we are also interested in modelling the LFP power features at the low-frequency bands, which fundamentally require longer time steps to be computed[51,133]. Intuitively, this is because we need to fully observe at least one period of the low-frequency band signals to compute their power. As these signals have longer periods, this naturally requires a longer time step. We aim to model the brain network dynamics at multiple frequency bands including low frequency theta band (1-7 Hz) to show that our framework provides a general enabling technology. Low-frequency LFP power bands have been shown to be important neural biomarkers to be controlled in many brain disorders, for example OFC theta and alpha power in mood disorders[13] and thalamic theta and alpha power in Tourette syndrome[127]. These bands are also important in many brain functions such as mood processing[22,23], memory[122] and attention[123]. In these applications, for example, to compute the theta band powers (1-7 Hz) and model their changes, a time step of more than 1 s is needed. This is because the slowest-changing 1 Hz component in this band requires one complete period (1 s) to compute its power. Therefore, any meaningful variation of the 1Hz power only happens slower than 1 s.

Choice of the power extraction method

We use the multitaper method[105] to compute the LFP power features with a given time step because the multitaper method can reduce the variance in extracting the spectral powers of stochastic LFP signals[105], leading to less noisy power computation and benefiting IO modelling. To confirm that our IO modelling method is robust to the method of power computation, we conducted a control analysis to compare the multitaper method and the continuous wavelet transform (CWT)[134] method in terms of the IO prediction accuracy. In the control analysis, we used CWT to compute continuous estimates of LFP power features at 200 Hz and then averaged the CWT powers over time the same way as in the main analyses (i.e. with the same time steps). We then repeated the entire IO modelling with this new way of computing LFP power features. We found that the IO prediction accuracy using CWT power features was similar to that of the multitaper power features (Supplementary Fig. 11). While similar, the performance of IO modelling using CWT was a bit worse because the multitaper method specifically aims to optimally reduce the variance in computing power features of stochastic signals[105] and thus can reduce the noise in IO modelling; in comparison, the CWT method aims to improve the frequency

resolution in computing the power features[134] without optimizing the variance in computing power features. This result also confirms that for IO modelling purposes at the time steps we consider (0.5 to 1 s), multitaper method is preferred over the CWT method for computing the LFP power features.

Supplementary Note 3. Design of the MN-modulated stimulation waveform: choice of the stimulation parameter values

The stimulation parameter values in MN are a design choice. In this work, we chose the stimulation parameter values based on typical ranges used in current DBS protocols. More specifically, we chose the stimulation frequency values to be between 50 Hz and 100 Hz, which is within the stimulation frequencies used in direct electrical stimulation for improving mood[13] (e.g., stimulation frequencies at 60Hz, 80Hz and 100Hz have been shown to be effective[13]) and overlaps with the stimulation frequency range in some studies of DBS for Parkinson's Disease[4,102] (e.g., stimulation frequencies at 70 and 80 Hz have been shown to be effective[102]) for example. We chose the MN stimulation amplitude to include 15 μA, 20 μA, 30 μA and 40 μA, which covered the range of stimulation amplitude that has been shown to be effective in symptom relief in an non-human primate model of neurological disorders[135]. More specifically, a prior study[135] shows that the typical voltage used for alleviating Parkinson's Disease symptoms in a non-human primate model was 3 V and the impedance of the electrodes used in the study was 100–150 kΩ. Therefore, the corresponding current was around 3 V/100 kΩ = 30 μA, which was within the current range 15 μA to 40 μA that we tested in our MN stimulation. Moreover, in the three-level MN stimulation, we additionally included stimulation off (0 μA amplitude and 0 Hz stimulation) because prior work has shown significant on-off stimulation effects on power features[13].

Note that in principle other desired ranges of stimulation amplitude and frequency can also be used in our modelling framework. To demonstrate an enabling technology, we chose the above stimulation amplitude and frequency range to show the feasibility of model-based prediction of large-scale multiregional brain network responses; this choice of amplitude and frequency range was especially motivated by its use in neuropsychiatric disorders such as depression. This range can be tuned for any specific application. For example, our approach may be extended to model other higher frequency stimulation ranges that have been shown to be effective in neurological disorders such as Parkinson's Disease, e.g., DBS in the range of 130-185 Hz[96] (also see Discussion section "Future extensions and directions").

Supplementary Note 4. Design of the MN-modulated stimulation waveform: choice of the switch period of MN

MN has white spectrum across all time-scales when the switch period T_{sw} (defined as the shortest time between switches in the MN, see Methods) equals the time step T_s. When $\frac{T_{sw}}{T_s} > 1$, the fast time-scale component of the MN is suppressed, thus allowing the experimenter to focus the modelling on the slower time-scale elements of brain network dynamics (i.e., the dynamics of the measured LFP power features). On the other hand, T_s should not be chosen smaller than T_{sw} to ensure that at most one switch of stimulation parameters happens within one time step. In this work, in 10 out of the 16 IO datasets, we chose $T_{sw} = T_s = 1$ s or 0.5 s such that MN has white spectrum across all time-scales. In the other 6 datasets, we varied T_{sw} to be 4 s or 6 s and used $T_s = 1$ s to have larger values of $\frac{T_{sw}}{T_s}$ so that the fast time-scale component of the MN can be tuned (Supplementary Fig. 14). We used these 6 datasets as additional control analyses to confirm our

finding that mainly the slow time-scale component of the temporally-varying stimulation amplitude and frequency drove the brain network dynamics (Supplementary Note 14).

Supplementary Note 5. State transition matrix characterizes the input-driven dynamics as well as the intrinsic dynamics

The state transition matrix A characterizes both the input-driven dynamics and the intrinsic dynamics. We can see this from equations (2) and (3) in Methods. First, from equation (2), the effect of the stimulation input on the input-driven part of the latent state $x_{k,I}$ is dictated by the state transition matrix A since this matrix weights how past inputs affect the current state as expanded below:

$$x_{k,I} = Bu_{k-1} + ABu_{k-2} + A^2Bu_{k-3} + \cdots + A^{k-1}Bu_{0,} \qquad \text{(S1)}$$

We obtained the above equation by recursively evaluating equation (2), that is by recursively replacing the equation for $x_{i,I}$'s for time steps $i = 1, \ldots, k$ into equation (2). Accordingly, the state transition matrix A is directly used in forward prediction as expanded in equation (6) in Methods. Second and similarly, from equation (3), the evolution of the intrinsic part of the latent state $x_{k,N}$ over time is also characterized by the state transition matrix A as expanded below:

$$x_{k,N} = w_{k-1} + Aw_{k-2} + A^2w_{k-3} + \cdots + A^{k-1}w_{0.} \qquad \text{(S2)}$$

This is obtained by recursively evaluating equation (3). Therefore, we see that the input-driven dynamics in equation (S1) and the intrinsic dynamics in equation (S2) are both dictated by the same critical state transition matrix A. Consequently, the state transition matrix A has an important role in the one-step-ahead prediction of the overall brain network dynamics during stimulation as expanded in equation (14) in Methods.

Supplementary Note 6. Details on single-trial forward prediction, single-trial one-step-ahead prediction and experimental design

Here, we provide further details on the difference between forward prediction and one-step-ahead prediction, and how our experimental design allows us to evaluate these two different predictions for single trials.

As stated in Methods (equations (1)-(3)), the measured overall brain network dynamics y_k (i.e., the measured LFP power feature time-series) can be decomposed into two parts: (i) input-driven dynamics $y_{k,I}$ and (ii) intrinsic dynamics $y_{k,N}$. To evaluate the ability of the model to predict different aspects of neural dynamics, two different prediction measures need to be used as detailed below.

Forward prediction to predict single-trial input-driven dynamics.

This is the prediction we use in the main analyses (see Supplementary Fig. 5). Forward prediction aims to quantify how well the IO model can predict the input-driven dynamics. Dissociating input-driven dynamics from all the intrinsic (input-irrelevant) dynamics is difficult and requires a careful new experiment design in this study. To achieve this dissociation, we repeat the same stochastic MN input across all trials (Fig. 1b and Supplementary Fig. 5b). In addition, in each cross-validation fold, we take the *same* time-segment in each trial as one test set (e.g., first quarter) and

take the rest of each trial for inclusion in the training set (e.g. last three quarters) (Supplementary Fig. 5b). Note that in each cross-validation fold, there is one training set consisting of concatenated training data from all trials, and there is one test set per trial and thus a total of N_R test sets where N_R is the number of trials (Supplementary Fig. 5b-5d). Doing the above achieves two goals for evaluating our IO model in a given cross-validation fold:

(a) As the first goal, in each fold, we can obtain the ground-truth of single-trial input-driven dynamics in the test sets by averaging the measured overall brain network dynamics across test-sets (each test set corresponds to one trial; Supplementary Fig. 5d). This is because for a given fold, as the input is the same across trials, the single-trial input-driven dynamics are also the same across all trials and thus remain intact after averaging (averaged = single, Supplementary Fig. 5d). However, as intended, averaging suppresses the intrinsic dynamics because they are input-irrelevant and change from trial to trial.

(b) As the second goal, this design allows forward prediction to predict and assess single-trial input-driven dynamics. Indeed, as the input is the same across trials (equivalently across test sets) for the given cross-validation fold, forward prediction is also the same for every trial in that fold (thus trial-averaged prediction = single-trial prediction; Supplementary Fig. 5d). This can be seen from equation (6) in Methods, showing that forward prediction of single-trial input-driven dynamics, i.e., $\hat{y}_k$, is only a function of the past inputs $\{u_0, u_1, \dots, u_{k-1}\}$.

Together, achieving these two goals allows us to assess IO prediction accuracy by comparing the single-trial forward prediction of the model with the ground-truth of single-trial input-driven dynamics obtained through averaging.

We emphasize that for different cross-validation folds of an experiment or for different experiments, we use different stochastic inputs (Supplementary Fig. 5b) and no averaging is ever done across cross-validation folds or experiments. So overall, we are evaluating the single-trial forward prediction across different stochastic waveforms. Taken together, the aim for single-trial forward prediction is to evaluate how well the currently measured LFP power feature can be predicted using only past stimulation inputs. Thus, forward prediction assesses single-trial *input-driven* dynamics, which is the main measure of performance for an IO model. That is why we need to compare the forward prediction with the ground-truth of single-trial input-driven dynamics; this ground-truth is obtained by averaging the measured LFP power features to reduce their input-irrelevant dynamics.

One-step-ahead prediction to predict single-trial overall brain network dynamics

If we aim to evaluate how well the fitted IO model can predict the measured overall brain network dynamics in a single trial, we should also use the past LFP power features measured in that trial in our prediction. The aim for single-trial prediction is thus to evaluate how well the currently measured LFP power feature can be predicted by both past inputs and past measured LFP power features in that trial. This allows us to include the prediction of both input-driven and intrinsic dynamics. The ground-truth in this case is thus the measured single-trial LFP power feature time-series and no averaging is needed to obtain the ground-truth.

Since our IO models are fitted directly from single-trial training data without averaging (Supplementary Fig. 5c), the fitted IO models capture single-trial noise statistics and can be

directly used to predict single-trial brain network dynamics as we show in equations (13) and (14) in Methods. Different from forward prediction, the one-step-ahead prediction $\tilde{y}_k$ is a function of both the past inputs $\{u_0, u_1, \dots, u_{k-1}\}$ and the past LFP power features $\{y_0, y_1, \dots, y_{k-1}\}$ as it should by design.

Finally, in our analysis of predicting single-trial overall brain network dynamics, to compute the one-step-ahead prediction for single-trial dynamics that happen at and slower than a given time-scale, we low-pass filter both $\tilde{y}_k$ and y_k below the inverse frequency corresponding to that time-scale first and then compute the correlation coefficient (e.g. below 0.05 Hz=1/20 s for 20 s). To show the robustness of the one-step-ahead prediction accuracy to the time-scale, we test both the time-scale of the stimulation input effect (around 20 s; Supplementary Note 14 and Supplementary Fig. 15) as well as time-scales that are twice or four times faster, i.e., 10 s and 5 s (see Supplementary Fig. 12 for details).

Supplementary Note 7. Appropriateness of the multi-trial experiment design

Here we briefly summarize the reasons and results that show the appropriateness of our multi-trial experiment design:

(1) The main measure of an IO model is its ability to predict single-trial input-driven dynamics, that is to predict how stimulation input drives brain network dynamics. Thus, to evaluate this prediction for our IO models, we design a multi-trial MN experiment where the same stochastic MN input is delivered in each trial. This experiment design allows us to evaluate the IO model by achieving two goals (Supplementary Note 6): (i) Predicting the single-trial input-driven dynamics: we train our IO model using single-trial, non-averaged IO data such that we can forward predict the single-trial input-driven dynamics. (ii) Obtaining the ground-truth of the single-trial input-driven dynamics: the multi-trial MN experiment allows us to use averaging over the test sets to suppress the input-irrelevant intrinsic dynamics and reveal the ground-truth of single-trial input-driven dynamics to compare with our forward prediction (Supplementary Fig. 5 and Supplementary Note 6 above).

(2) In our experiment design, we conduct the cross-validation by leaving out a segment of each trial as the test sets rather than leaving out an entire trial. Thus, this cross-validation design ensures that the input in the training and the test set are completely independent regardless of adaptation effects (Supplementary Note 9 and Supplementary Fig. 5b) and rules out the confounding influence of overfitting to the input pattern (Supplementary Note 9).

(3) Because of the independence of inputs in item (2) above, adaptation effects do not affect our IO prediction accuracy computations. Indeed, because of this input independence, even when there exit adaptations in the LFP power feature dynamics that differentially affect trials, such adaptation will not affect the independence of training and test sets. Consistent with this, we showed that removing the relatively small adaptation effects—removing the relatively small stimulation-induced auto-correlation beyond the duration of the trial—did not affect the IO prediction accuracy (Supplementary Fig. 22c).

Supplementary Note 8. Determination of the number of data points used in IO modelling in each trial

In Methods, we conducted a four-fold cross-validation to evaluate the model prediction performance and further used another four-fold inner cross-validation to select the hyper-parameters within each training set. To ensure that we have an integer number of data points to conduct model fitting and evaluation, we required the number of data points in each trial be a multiple of $4 \times 4 = 16$. Therefore, for a particular IO dataset, if the total number of data points of one trial (denoted by $J_0 = \frac{T_R}{T_s}$) could not be divided by 16, we removed the last mod(J_0,16) data points in each trial, and used the remaining L data points in each trial (Supplementary Table 3) for subsequent IO modelling.

Supplementary Note 9. Control analysis to show the independence of test and training sets

We emphasize that the input in the test set is completely independent of both the input and output in the training set by the stochastic nature of an MN pattern. Also, in forward prediction, the output LFP power feature is predicted purely based on the stimulation input history, not the history of the output itself. Thus the cross-validation procedure ensures that the forward prediction in the test set is truly tested on novel IO data. To further show this, we conducted the following control analysis. For each LFP power feature, we replaced the output LFP power feature in the test with the output LFP power feature in a training set (the training set right next to it), while keeping the test set input intact. In this way, we artificially increased the output correlation with training set to its maximum value of 1. At the same time, we kept the input stimulation in the test set the same. Then we repeated the same cross-validation procedure. We found that among the predictable power features, the IO prediction accuracy did not artificially improve; instead, it was significantly smaller than the IO prediction accuracy obtained from the main analyses (-0.02 ± 0.01 vs. 0.45 ± 0.006, mean ± s.e.m, two-sided Wilcoxon signed-rank test, $P = 5.99 \times 10^{-40}$) and actually dropped to chance level (the IO prediction accuracy -0.02 ± 0.01 was close to 0). This is because the test set input stimulation was completely independent of the training set input and output data given the stochastic nature of MN. Thus replacing the actual test set output data with training set output data disrupted the IO relationship in the test set, leading to small IO prediction accuracy. This further confirms the independence of the test and training sets and the validity of the cross-validation procedure and our experimental design (see Supplementary Note 7).

Supplementary Note 10. *P* value calculation from a baseline distribution

To calculate a *P* value from a baseline distribution (e.g., input- or output-baseline distribution), we first use the largest 50 samples of the baseline distribution to fit a parametric generalized Pareto distribution (GPD)[136]. This is because the distribution of the extreme values of any set of independent and identically distributed random variables can be approximated by a GPD[136]. We test the goodness-of-fit of the GPD by using the Anderson-Darling test. If the Anderson-Darling *P* value (P_{AD}) > 0.05, GPD is a good match to the data; thus we use the fitted GDP distribution to calculate the *P* value as the probability of the test statistic being larger than the actual statistic according to the fitted GDP. If P_{AD} < 0.05, we gradually reduce the number of samples used to fit the GPD by a step size of 10 and repeat the above fitting procedure. If none of the above fitting procedures result in P_{AD} > 0.05, we simply calculate the *P* value as $(N_L + 1)/(N_b + 1)$ where N_b is

the total number of samples in the baseline distribution, and N_L is the number of samples in the baseline distribution that are larger than the actual statistic[136].

Supplementary Note 11. Linear state-space modelling (LSSM) of intrinsic dynamics of at-rest LFP power features

For each MN stimulation dataset and its associated three constant stimulation datasets (see Supplementary Tables 3 and 4), we collected 5 minutes pre-stimulation and 5 minutes post-stimulation at-rest LFP data. After removing 50 s data right after device turns on/off to avoid electronic artifacts (Supplementary Note 16), we concatenate these pre-stimulation and post-stimulation at-rest data to model their intrinsic dynamics. We had 5 minutes – 50 s = 250 s for each pre/post-stimulation at-rest period. We had 2 pre/post-stimulation at-rest periods for each MN stimulation dataset and its associated three constant stimulation datasets. In total we had 250*2*4=2000 s or around 33 minutes of at-rest LFP data from each day. We computed the at-rest LFP power features from the same LFP channels and same bands using the same Method as the main analyses.

We use the same LSSM framework but without the input term, Bu_k, to model the at-rest LFP power features y_k and term the model at-rest LSSM, which is given as:

$$\begin{cases} x_{k+1} = Ax_k + w_k \\ y_k = Cx_k + v_k \end{cases}, \qquad \text{(S3)}$$

Intrinsic dynamics relate to the dynamics of LFP power feature itself: how the current value of LFP power feature depends on its past values. These intrinsic dynamics in the model are driven by the state noise term (w_k) and the observation noise term (v_k) as shown in equation (S3). To assess the prediction of intrinsic dynamics, we need to predict the current value of LFP power features using all past LFP power features measured (unlike what we did in forward prediction to assess input-driven dynamics, see Supplementary Fig. 23a). Utilizing the statistics of the model noise is key in the above prediction of intrinsic dynamics.

We thus use a four-fold cross-validation framework to train and evaluate the at-rest LSSM. The test set again does not see any data from the training set. We use the fitted at-rest LSSM in the train set to perform one-step-ahead prediction in the test set. In one-step-ahead prediction, we predict the value of the current LFP power feature as a function of the LFP power features at all previous time steps. With our trained at-rest LSSM, the optimal one-step-ahead prediction (to minimize mean-squared error) is given by the Kalman predictor[137] that takes the following recursive form

$$\begin{cases} p_{k+1} = Ap_k + K(y_k - Cp_k) \\ \tilde{y}_k = Cp_k \end{cases}, \qquad \text{(S4)}$$

where p_k is the one-step-ahead prediction of the latent state x_k, $\tilde{y}_k$ is the one-step-ahead prediction of y_k, and K is the Kalman gain computed from A, C and noise covariances of w_k and v_k[137]. Different from forward-prediction, in one-step-ahead prediction, the current one-step-ahead predicted state p_k is expressed as a linear function of past LFP power features $y_{k-1}, y_k, \dots, y_0$ and its initial state p_0:

$$p_k = Ky_{k-1} + (A - KC)Ky_{k-2} + \cdots + (A - KC)^{k-1}Ky_0 + (A - KC)^k p_0. \qquad \text{(S5)}$$

This can be seen through expanding equation (S4) by recursively replacing the equation for p_i from time $i = 0, \dots, k$. Similar with forward-prediction, to consider the most challenging case, we also assume we do not have any prior information about the initial state p_0. We thus set the initial state $p_0 = 0$ for all one-step ahead predictions. Finally, we note that the critical Kalman gain that enables the prediction (no prediction is possible with a zero gain) is a function of the model noise statistics[137], showing that modelling the noise is key in the prediction of intrinsic dynamics.

Informed by our time-scale analysis (see Supplementary Fig. 15), we focus on evaluating the prediction of dynamics within the prevalent time-scales we found for the effect of stimulation (20 s and slower). We also assess the robustness of our finding at time-scales that are faster than the prevalent time-scales. For a given time-scale T, after obtaining the one-step ahead prediction $\tilde{y}_k$, we low pass filter both $\tilde{y}_k$ and y_k below $\frac{1}{T}$ Hz to assess time-scales of T s and slower; we then compute the correlation coefficient between them to quantify the one-step-ahead prediction accuracy. The chance level one-step-ahead prediction for an at-rest LFP power feature is given by randomly permuting the outputs in time and repeating the modelling procedure (termed output-permutation test, see Methods).

Supplementary Note 12. Separating the effect of stimulation amplitude and frequency

To separately study the effect of stimulation frequency and amplitude on the output LFP power features, we compared with two special cases of the LSSMs that either just modeled the effect of the amplitude or just modeled the effect of the frequency. First, we only modeled the amplitude effect using the following LSSM:

$$\begin{cases} x_{k+1} = Ax_k + bu_k^{\text{amp}} + w_k \\ y_k = Cx_k + v_k \end{cases}, \qquad \text{(S6)}$$

where $u_k^{\text{amp}} \in \mathbb{R}$ represents the MN stimulation amplitude and $b \in \mathbb{R}$. The model parameters are $\theta = \{A, b, C, Q\}$, where Q is the noise covariance defined as in equation (1) in Methods. Then the same cross-validation procedure was repeated. We compared the IO prediction accuracy resulting from using the above model that only considered the stimulation amplitude with the IO prediction accuracy from the full dynamic IO LSSM in equation (1) that considered both the stimulation amplitude and frequency. We repeated the above modelling and comparison for the case of stimulation frequency alone as well, by replacing u_k^{amp} with $u_k^{\text{freq}} \in \mathbb{R}$ in the above LSSM. This allowed us to also investigate only modelling the frequency effect.

Supplementary Note 13. Oscillatory dynamics in the brain network response to stimulation

The eigenvalues of the state transition matrix A in the LSSM in equation (1) characterize the dynamics of the effect of stimulation on the output LFP power features[52]. Eigenvalues are in general complex conjugate numbers and can be written as $re^{\pm j\delta} \in \mathbb{C}$, where j is the unit imaginary number, $r \in [0,1]$ is the amplitude (representing a stable eigenvalue within the unit disk[138]) and $\delta \in [0, \pi]$ (in rad/s) is the oscillation frequency and we define its associated period as $\tau = \frac{2\pi T_s}{\delta}$ (in seconds, inverse of the oscillation frequency), where T_s is the time step or sampling period of the output LFP power features. Eigenvalues at different locations in the complex conjugate plane represent different types of dynamic responses. An eigenvalue equal to 1 ($r = 1, \delta = 0$) represents dynamic responses in which the effect of stimulation at each time is simply an unweighted smoothed average of the past values of the stimulation inputs[52] (Supplementary Fig. 13c). In addition, an eigenvalue on the positive real axis but different from 1 ($r \in (0,1), \delta = 0$)

represents the dynamic response associated with weighting the past stimulation inputs with an exponential damping[138] (weighting older samples much less that recent samples, Supplementary Fig. 13c). We term the eigenvalues on the positive real axis non-oscillatory eigenvalues. The speed of the exponential damping can be quantified by the damping life-time defined as $\chi = \frac{-1}{\ln r} T_s$ (the time it takes for the signal to decay to $\frac{1}{e}$ or 36.79% of its initial value over time), with larger χ representing slower damping. Finally, a pair of complex conjugate eigenvalues or an eigenvalue on the negative real axis (the eigenvalues with $\delta \neq 0$) represent a more complex form of dynamic response in which the effect of stimulation at each time not only involves an exponential damping of the past stimulation inputs, but also oscillatory changes (Supplementary Fig. 13c). We characterize these with damping life-time χ and oscillatory period τ. We term such a negative eigenvalue or a pair of complex conjugate eigenvalues, oscillatory eigenvalues.

To further confirm the role that oscillatory dynamics play in predicting the brain network response, we conducted a control analysis where we compared the full LSSM in equation (1) in Methods with another special case of LSSM in which we constrained the state transition matrix to only have positive real eigenvalues when fitting it, i.e., to only have non-oscillatory eigenvalues. We call this LSSM a non-oscillatory model. The response corresponding to this non-oscillatory model is thus a weighted average of the previous stimulation inputs with exponential damping (each LFP feature can have a different damping life-time though)[138] but no oscillation. We then repeated the same cross-validation procedure using the non-oscillatory model.

To study the dynamical characteristics of the brain network response to stimulation, we focus on comparing the models on predictable LFP power features, which have a significant IO predication accuracy and thus a predictable response whose dynamics can be studied. The IO modelling results for the other (non-predictable) power features are likely just noise since they do not pass the input/output-baseline tests and will only add noise into the comparisons. Nevertheless, we found that even by including all LFP power features (whether predictable or not) and in the presence of this added noise, the dynamic IO model significantly outperformed the other IO models (see Results), again confirming the existence of complex damping and oscillatory dynamics.

Supplementary Note 14. Time-scale analysis of the brain network dynamics in response to stimulation

Computation of effective time-scale of the effect of stimulation

Since the IO model in equation (1) in Methods is a dynamic parametric linear model, it facilitates the calculation of the frequency domain transfer function from the input stimulation to each output LFP power feature. A closed-form representation of the transfer function $K^{(i)}(jf) \in \mathbb{R}^{N_u \times 1}$ from the input stimulation u_k to an output power feature $y_k^{(i)}$ can be derived as:

$$K^{(i)}(jf) = C^{(i)}\left(e^{j2\pi f} I - A^{(i)}\right)^{-1} B^{(i)}, \qquad \text{(S7)}$$

where $f \in [0, \frac{1}{T_s}]$ is the frequency (T_s is the time step used in the modelling) and j is the unit imaginary number. Note that this frequency f should not be confused with the frequency band of each LFP power feature. After an LFP power feature is extracted from the LFP signal, it constitutes a time-series. The frequency f refers to how fast this LFP power feature time-series and the MN time-series change over time—not how fast the raw LFP signals change over time—,

and thus we refer to f as inverse time-scale. Accordingly, the time-scale (TS, in seconds) is defined as:

$$\mathrm{TS} = \frac{1}{f}. \qquad \text{(S8)}$$

The first element of the transfer function above indicates the effect of the stimulation amplitude on the power feature response and the second element of it shows the effect of the stimulation frequency. The square of the amplitude of the first element of the transfer function—$|K^{(i)}_{\mathrm{amp}}(f)|^2 = [1,0]K^{(i)}(jf)\left([1,0]K^{(i)}(jf)\right)^{\dagger}$ ($\cdot^{\dagger}$ represents taking conjugate transpose)—then shows how the energy of the effect of stimulation amplitude on the power feature response is distributed across different time-scales. Similarly $|K^{(i)}_{\mathrm{freq}}(f)|^2 = [0,1]K^{(i)}(jf)\left([0,1]K^{(i)}(jf)\right)^{\dagger}$ shows the energy distribution of the effect of stimulation frequency across different time-scales. Since we use a moving window of $10T_s$ in the LFP power feature computation, the components of faster time-scales (inverse time-scale f>$\frac{1}{10Ts}$Hz) are suppressed by this moving window. Therefore, in this analysis, we focus on analyzing the energy distribution below $\frac{1}{10Ts}$.

From the energy distribution $|K^{(i)}_{\mathrm{amp}}(f)|^2$, we can for example find the inverse time-scale f_0 below which 80% of the energy is concentrated, i.e., f_0 satisfies:

$$\int_0^{f_0} |K^{(i)}_{\mathrm{amp}}(f)|^2 df = 0.8 \times \int_0^{f_{\max}} |K^{(i)}_{\mathrm{amp}}(f)|^2 df, \qquad \text{(S9)}$$

where $f_{\max} = \frac{1}{10Ts}$ is the inverse of the fastest time-scale that we examine in this analysis. We thus define the effective time-scale (ETS) as the inverse of f_0:

$$\mathrm{ETS}^{(i)}_{\mathrm{amp}} = \frac{1}{f_0}, \qquad \text{(S10)}$$

This means that most of the effect (80%) that the stimulation amplitude had on this power feature ($y_k^{(i)}$) happens at time-scales slower than the $\mathrm{ETS}^{(i)}_{\mathrm{amp}}$. Similarly, most of the effect of the stimulation frequency on this power feature happens at time-scales slower than $\mathrm{ETS}^{(i)}_{\mathrm{freq}}$. We calculated $\mathrm{ETS}^{(i)}_{\mathrm{amp}}$ and $\mathrm{ETS}^{(i)}_{\mathrm{freq}}$ using the fitted LSSM in each of the cross-validation folds, and then averaged the effective time-scales over the four cross-validation folds as the final $\mathrm{ETS}^{(i)}_{\mathrm{amp}}$ and $\mathrm{ETS}^{(i)}_{\mathrm{freq}}$ for the *i*th power feature in y_k.

We calculated the ETS of the effect of stimulation amplitude and frequency on predictable power features using all fitted IO models in the training set in the four-fold cross-validation. The average ETS across LFP power features was 19.89 ± 0.58 s and 19.52 ± 0.56 s (mean ± s.e.m.) for stimulation amplitude and frequency, respectively (Supplementary Fig. 15). The ETS values were much slower (i.e., larger) than the fastest time-scale we examined here, i.e., than $\frac{1}{f_{\max}} = 10T_s = 10\ s$ for a time step of 1 s, or than $\frac{1}{f_{\max}} = 10T_s = 5\ s$ for a time step of 0.5 s. Thus our results show that the energy of the transfer function concentrated within slow time-scales (Supplementary Fig. 15).

Varying the time-scale component in the MN amplitude and frequency to further confirm slow time-scale responses

To further confirm that the LFP response to stimulation mainly happened at slow time-scales, we gradually suppressed the fast time-scale component in the MN amplitude and frequency and compared how the corresponding IO prediction accuracy changed. The spectrum of the MN amplitude and frequency depends on the ratio between the switching period T_{sw} and the discretization time step T_s (how fast the MN pattern switches from one level to another, see Methods and Supplementary Table 3). A larger ratio $\frac{T_{sw}}{T_s}$ leads to an MN spectrum that focuses more on slower time-scales because the MN pattern makes switches less frequently and thus changes slower[18,139] (also see Supplementary Fig. 14). In our IO datasets, we tested MN patterns with $\frac{T_{sw}}{T_s} = 1, 4, 6$. By increasing $\frac{T_{sw}}{T_s}$, we gradually suppressed the fast time-scale component of the MN pattern. For MN patterns with $\frac{T_{sw}}{T_s} \neq 1$, we define the cut-off time-scale as the time-scale faster than which (or equivalently, inverse time-scale f above which) the PSD of MN decreased by over 3 dB compared with the flat part of the PSD at the slow time-scales (Supplementary Fig. 14). For IO datasets with $\frac{T_{sw}}{T_s} = 4 \text{ or } 6$, the cut-off time-scale was 8.88 s and 13.33 s, respectively, meaning we focused the MN spectrum on time-scales slower than these values and suppressed the faster time-scales (larger f's) (Supplementary Fig. 14). In 10 IO datasets we used $\frac{T_{sw}}{T_s} = 1$ and for the remaining 6 IO datasets we used $\frac{T_{sw}}{T_s} = 4 \text{ or } 6$ to focus on slower time-scales. However, we found that the IO prediction accuracy corresponding to the three different $\frac{T_{sw}}{T_s}$'s were not significantly different (Kruskal-Wallis test, $P = 0.1361$). This result shows that suppressing the faster time-scale component of the MN amplitude and frequency does not degrade IO prediction accuracy, further confirming that that the LFP response to stimulation mainly happened at slow time-scales.

Finally, in 10 IO datasets, we had $\frac{T_{sw}}{T_s} = 1$, so the MN had flat PSD across all time-scales, including the fast time-scales. The effective time-scale of neural response across LFP power features in these 10 datasets was 21.73 ± 0.71 s for stimulation amplitude and 21.21 ± 0.70 s for stimulation frequency, which were both slower than the slowest cut-off time-scale of MN (13.33 s) in the other 6 IO datasets with $\frac{T_{sw}}{T_s} = 4 \text{ or } 6$. In addition to the IO prediction accuracy being similar for different $\frac{T_{sw}}{T_s}$'s, this result also again suggests that most of the effect of stimulation happened at time-scales slower than the cut-off time-scale of MN in the 6 IO datasets with $\frac{T_{sw}}{T_s} = 4 \text{ or } 6$.

Supplementary Note 15. Cross-correlation between the input stimulation and output LFP power features at different time-delays

To further confirm our findings about the time-scale of the effect of stimulation, we also examined the cross-correlation between the stimulation amplitude and frequency (input) and the ground-truth single-trial input-driven dynamics (output) at different time-delays. We found that the time-delay of significant cross-correlation was close to the time-scale of the effect of stimulation found using our IO model.

More specifically, for each predictable power feature, we computed the cross-correlation coefficient $\rho(\tau)$ (τ is the time-delay) between the time-delayed input stimulation amplitude $u_{k-\tau}^{\text{amp}}$

within a single trial and the ground-truth single-trial input-driven dynamics $\bar{y}_k$ obtained by averaging the measured brain activity across all MN trials (Supplementary Fig. 5d, note that the input was identical for all trials since the same MN input was repeated for each trial). This cross-correlation coefficient is given by

$$\rho(\tau) = \frac{\mathrm{Cov}(u_{k-\tau}^{\mathrm{amp}}, \bar{y}_k)}{\sqrt{\mathrm{Var}(u_{k-\tau}^{\mathrm{amp}})\mathrm{Var}(\bar{y}_k)}}, \qquad \text{(S11)}$$

where τ is the time-delay from the input stimulation to the output. We refer to this cross-correlation coefficient as the IO cross-correlation coefficient.

We then tested whether the absolute value of the IO cross-correlation coefficient, i.e., $|\rho(\tau)|$, was significantly larger than chance level at various time-delays. To do so, we first define the chance level as the absolute value of the cross-correlation coefficient between a randomly generated MN input stimulation amplitude with the same output as in the actual IO dataset. We randomly generated 500 MN input stimulation amplitude time-series and computed a distribution of the absolute value of the chance level cross-correlation coefficient. Then, we Z-scored the absolute value of the actual IO cross-correlation coefficient using the mean and standard deviation of the chance level distribution, which is denoted as $|\rho(\tau)|^{(Z)}$. Finally, for each time-delay $\tau = 1\ s, 2\ s, \ldots, 60\ s$ (where 60 s corresponds to the shortest trial length among the 16 IO datasets, see Supplementary Table 3), we tested if $|\rho(\tau)|^{(Z)}$ was significantly larger than chance (random-test). A significant $|\rho(\tau)|^{(Z)}$ shows that the input stimulation amplitude at time-delay τ has stronger correlation with the output than would be expected by chance.

As the computation of the cross-correlation is done without any IO modelling, the cross-correlation results can be compared with the time-scale results obtained using the IO model. We found that the time-delay in IO cross-correlation computed without a model and the effective time-scale of the stimulation effect computed from our IO model are similar (Supplementary Fig. 17). This result again confirms that the current output LFP power feature depends on the history of the input stimulation. This is intuitively sound as our IO model utilizes the IO cross-correlation to optimally predict the current output with the history of the input. Thus the effective time-scale of the IO model quantifies how long into the past the inputs are needed for the prediction of the output or equivalently how long into the past the inputs are correlated with the output. Thus, the effective time-scale is close to the time-delay of the significant IO cross-correlations.

Supplementary Note 16. Calculation of at-rest LFP power features

To obtain the at-rest LFP power features, first, we concatenated the first 250 s of the 5 minutes-long (300 s) pre-stimulation at-rest raw signals and the last 250 s of the 5 minute-long (300 s) post-stimulation at-rest raw signals. The reason we did not use the 50 s periods around the stimulation period (50 s before and 50 s after) was because turning the electronic stimulation device on and off can induce large stimulation artifacts that can last for tens of seconds. Thus, we avoided these unusable segments of data to ensure we only used at-rest LFP data without any electronic artifact in this analysis. Second, we repeated the same preprocessing procedure on these signals that we had applied on the raw signals recorded during stimulation (except that no artifact rejection was required now). Third, we calculated the LFP power features within the same 4 bands again using the multi-taper spectrogram analysis with a time step of 1 s.

Supplementary Note 17. Details of the derivation of at-rest functional controllability

For a stable LSSM in equation (18) in Methods, the infinite-horizon controllability Gramian

$$W_c = \sum_{j=0}^{+\infty} T^j D D' (T')^j, \qquad \text{(S12)}$$

provides an energy-related quantification of controllability and can be computed by solving the Lyapunov equation as in equation (19) in Methods[46]. Given an initial condition $z_0 = 0$, we denote a target state on the unit circle to be arrived at in N-steps by $z_N \in \{z \in \mathbb{R}^{N_y \times 1} : \|z\| = 1\}$ ($\|\cdot\|$ represents the vector 2-norm). Using standard linear systems theory, we have $z_N = Fv$, where $F = [D, TD, \dots, T^{N-1}D]$ and $v = [g_{N-1}, \dots, g_0]^T$ is the input sequence (with initial condition $z_0 = 0$). The optimal (minimum energy) input sequence to move from z_0 to z_N is given by the least norm solution of v, denoted by $v_{ln} = F'(FF')^{-1} z_N$, and the energy of this optimal input sequence is

$$\|v_{ln}\|^2 = z_N' (FF')^{-1} z_N = z_N' W_c(N)^{-1} z_N. \qquad \text{(S13)}$$

Now taking the average across all z_N over the unit hypersphere, we can find the average value of $\|v_{ln}\|^2$ as

$$\frac{\int_{\|z_N\|=1} z_N' W_c(N)^{-1} z_N dz_N}{\int_{\|z_N\|=1} dz_N} = \frac{1}{N_y} \mathrm{Tr}(W_c(N)^{-1}). \qquad \text{(S14)}$$

Since we are interested in quantifying the input energy that is needed to eventually reach the unit hypersphere from z_0, we allow the time to get there to be as long as needed as is common in the control literature[46,138]. We thus take $N \to \infty$, and have $\mathrm{Tr}(W_c(N)^{-1}) \to \mathrm{Tr}(W_c^{-1})$. Therefore, $\mathrm{Tr}(W_c^{-1})$ is proportional to the energy needed on average to move the state z_k around the state-space.

Now we consider moving the *i*th component in z_k, i.e., $z_k^{(i)} = Q^{(i)} z_k$, with $Q^{(i)} = [0,0,..,0,1,0,\dots,0]$ in which only the *i*th component is non-zero and equal to 1. Then the optimal energy related to this component is

$$\|v_{ln}\|^2 = \left\| \left(Q^{(i)}F\right)' \left(Q^{(i)} F F' Q^{(i)\prime}\right)^{-1} z_N^{(i)} \right\|^2 = z_N^{(i)} \widetilde{W}_c(N)^{-1} z_N^{(i)}, \qquad \text{(S15)}$$

where $\widetilde{W}_c(N) = Q^{(i)} W_c(N) Q^{(i)\prime} = W_c(N)(i,i)$, where $W_c(N)(i,i)$ represents the (i,i)th component in $W_c(N)$. Similarly, by taking $N \to \infty$, $\widetilde{W}_c^{-1} = \frac{1}{W_c(i,i)}$ is then proportional to the energy needed to move the ith component around the state space within infinite time steps, or equivalently $W_c(i,i)$ is inverse proportional to the needed energy. We finally take the logarithm of $W_c(i,i)$ as the functional controllability for the ith component.

Supplementary Note 18. Stimulation-induced changes/adaptation in functional brain network topology after stimulation ends

As one analysis on stimulation-induced adaptation in the main text, we investigate the change in functional brain network topology after stimulation ends by comparing functional brain network topologies computed using post stimulation at-rest data vs. using pre-stimulation at-rest data.

To do this, for each MN stimulation and constant stimulation dataset, we assess the functional brain network topology by defining four frequency-specific coherence networks[23,51,110] computed at the same four frequency bands used in the main analyses, i.e., 1-7 Hz, 8-12 Hz, 12-30 Hz and 30-50 Hz. In each frequency-specific coherence network, the spectral coherence is computed at one of the four bands in non-overlapping windows of 8 seconds each also using the multitaper method[105] (Supplementary Fig. 21a). We then study how each frequency-specific coherence network changes after stimulation ends.

To quantify the change of each coherence network, we compute three standard network topology measures in graph theory. (i) Change in mean node strength: Node strength quantifies how strongly one node is connected to the rest of the network and is computed by the average value of the edges that are connected to that node[56,140,141] (average value of the coherences with that node). Mean node strength is defined as the mean of node strength across all nodes in the network; (ii) Change in the variance of node strength, which is defined as the variance of node strength across all nodes in the network and quantifies the heterogeneity of node strengths in the network[41]. (iii) Network similarity. Network similarity is computed as the linear correlation coefficient between the pair-wise coherence values of one network to those of another network, also termed network covariance in prior studies[57,142]. Network similarity quantifies for two coherence networks, how similar the configurational patterns of the network edges are[41,57,142].

To compute the change in network topology measures, we do the following: (i) We compute the coherence network in the first 8 s window of pre-stimulation period and term it the reference coherence network. (ii) We compute the change from this reference coherence network to each coherence network computed in 8 s windows during the post-stimulation period (Supplementary Fig. 21a). (iii) We take the stimulation-induced change as the mean change in (ii). Since the computation of spectral coherence is influenced by noise[51,143] and since coherence networks may even have inherent changes without stimulation[23], we need to control for non-stimulation induced changes in the coherence network. To do so, we compute baseline values for the change in network topology measures as follows: (I) We compute the change from the reference coherence network to each coherence network computed in later 8 s windows during the pre-stimulation period (Supplementary Fig. 21a); (II) We take the baseline change as the mean change in (I).

Across all MN stimulation datasets, we found that after stimulation ended, there existed a significant increase in mean node strength compared with baseline for the low-frequency coherence network (Supplementary Fig. 21b, 1-7 Hz, two-sided Wilcoxon signed-rank test, P = 0.0026 after FDR correction for multiple frequency bands and network topology measures). Also, there existed a significant increase in variance of node strength compared with baseline for the low-frequency coherence network (Supplementary Fig. 21b, 1-7 Hz, two-sided Wilcoxon signed-rank test, P = 0.0123 after FDR correction). Moreover, network similarity between the post-stimulation coherence networks and the pre-stimulation reference coherence network significantly decreased compared with baseline for all four bands (Supplementary Fig. 21b, two-sided Wilcoxon signed-rank test, $P < 0.02$ after FDR correction). While the above changes were significant, the absolute values of the changes were relatively small. For example, the increase in mean node strength from baseline was on average less than 0.005 for all bands (Supplementary Fig. 21b, first panel) in the unit of coherence value (maximum range between 0 and 1). The absolute stimulation-induced percentage increase in mean node strength from the baseline was 9.49% ± 0.91% (mean ± s.e.m.). As another example, the decrease in network similarity from baseline was on average less than 0.02 for all bands (Supplementary Fig. 21b, last panel) in the

unit of linear correlation coefficient (maximum range between 0 and 1). The absolute stimulation-induced percentage decrease in network similarity from the baseline was 3.28% ± 0.33%.

Similar patterns of change in coherence network were observed for constant stimulation datasets (Supplementary Fig. 21c). For the significant changes we observe in MN stimulation datasets, we also observed them in constant stimulation datasets (Supplementary Fig. 21c). The amount of these significant changes was similar between MN and constant stimulation datasets (two-sided Wilcoxon sum-rank test, $P > 0.1886$ for all comparisons). Only in two frequency bands, the change in mean and variance of node strengths was significant for the constant stimulation datasets but not the MN stimulation datasets (12-30 Hz and 30-50 Hz, Supplementary Fig. 21c). However, the absolute values of the changes in these two frequency bands for the constant stimulation datasets were similar to the changes for the MN stimulation datasets. Also, for these two frequency bands, the changes in both the constant stimulation datasets and the MN stimulation datasets were very small. For example, absolute value of change in mean node strength in 12-30 Hz was $5.92\times10^{-4} \pm 4.39\times10^{-4}$ (in the unit of coherence) for the constant stimulation datasets and had a similarly small value of $2.21\times10^{-4} \pm 3.71\times10^{-4}$ for the MN stimulation datasets (Supplementary Fig. 21c).

These results indicate that MN and constant stimulation induce largely similar changes in functional brain networks after stimulation ends. The addition of the constant stimulation datasets thus confirms that there exists significant but relatively small adaptation in functional brain network topology after stimulation ends. Finally, our results are consistent with the findings of change in electrocorticography (ECoG) coherence networks after stimulation ends in human subjects due to constant stimulation, and the absolute values of the changes are comparable[41].

Supplementary Note 19. Stimulation-induced changes/adaptation in LFP power feature dynamics from one trial to another during stimulation

While stimulation only has relatively small effects on functional brain network topology (Supplementary Note 18 and Supplementary Fig. 21), such relatively small effects may still introduce changes in the LFP power feature dynamics during stimulation from one trial to another. As another analysis to investigate stimulation-induced changes/adaptation, we explicitly examine the changes of LFP power feature dynamics during stimulation across trials. To do so, we examine stimulation-induced, across-trial changes in terms of the auto-correlations of power features at delays longer than the duration of a single trial. Stimulation-induced, across-trial changes happen if compared with auto-correlations of pre-stimulation at-rest LFP power features, the auto-correlations of during-stimulation LFP power features are larger in absolute value. Therefore, we examine if there exist any stimulation-induced changes in the auto-correlations of LFP power features that happen slower than the duration of one trial.

To do this, for each individual LFP power feature, we compute its auto-correlation coefficient during stimulation (Supplementary Fig. 22a) and compute the average absolute value of the auto-correlation coefficients at delays longer than the duration of one trial. This average absolute value of the auto-correlation coefficients quantifies how correlated the LFP power feature dynamics are beyond the duration of one trial. To evaluate if these auto-correlations are induced by stimulation, we compare this average absolute value of the auto-correlation coefficient during stimulation with that computed the same way for the same LFP power feature but for the pre-stimulation at-rest data. We take the difference as the stimulation-induced change in auto-correlation coefficient. For each LFP power feature, we then evaluate whether there exists a significant change in auto-

correlation coefficient during MN stimulation. We repeat the same procedure for constant stimulation datasets.

For the MN stimulation datasets, 28% of predictable power features showed a significant stimulation-induced change in auto-correlation coefficient during stimulation (two-sided Wilcoxon sum-rank test, $P < 0.05$ after FDR correction, Supplementary Fig. 22b). The absolute value of these significant stimulation-induced changes in auto-correlation coefficient was relatively small, on average less than 0.05 (0.039 ± 0.002, mean ± s.e.m., Supplementary Fig. 22b). For the constant stimulation datasets, we had similar results. Here 31% of predictable power features showed a significant stimulation-induced change in auto-correlation coefficient during stimulation (two-sided Wilcoxon sum-rank test, $P < 0.05$ after FDR correction, Supplementary Fig. 22b). The absolute value of these significant stimulation-induced changes in auto-correlation coefficient was also relatively small for constant stimulation datasets, again on average less than 0.05 (0.042±0.003, Supplementary Fig. 22b). Moreover, the values of stimulation-induced changes in auto-correlation coefficient were not significantly different between the MN and the constant stimulation datasets (0.039 ± 0.002 vs. 0.042 ± 0.003, two-sided Wilcoxon sum-rank test, $P = 0.78$, Supplementary Fig. 22b). These results show that although stimulation-induced adaptation in LFP power features during-stimulation indeed exists, the effect of such adaptation on the auto-correlation of LFP power features across trials is relatively small. The added new constant stimulation datasets show similar results and reinforce our conclusions.

The relatively small stimulation-induced change in auto-correlation coefficient suggests that our prediction accuracy is likely unaffected by the adaptation. To confirm this, in a control analysis, we high pass filter the LFP power features at a frequency equal to the inverse of the duration of an MN trial before any IO modelling to remove any changes beyond the duration of one trial. This filtering ensures that there is no auto-correlation beyond the duration of one trial. Then, we repeat the IO modelling procedure. We found that the IO prediction accuracy did not change significantly compared with our main analyses (two-sided Wilcoxon signed-rank test, $P = 0.41$, Supplementary Fig. 22c). This result shows the appropriateness of our experiment design (Supplementary Note 7).

Supplementary Note 20. Closed-loop control simulations using the fitted IO models

Summary

Here we describe the details of closed-loop control simulations and expand on how there are two possible sources for the forward-prediction error (defined as the error between the predicted single-trial input-driven dynamics and the ground-truth single-trial input-driven dynamics). We can thus run best-case and worst-case simulations to characterize the range of closed-loop performance expected for the IO models fitted in our experiments. We show that in both cases, we can achieve control of a simulated internal brain state and that the IO model is key in enabling such closed-loop control.

Details of Simulation Setup

1. Sources of the forward-prediction error

Forward prediction quantifies the ability of IO models for predicting the input-driven dynamics (using only knowledge of the input). To obtain a measure of ground-truth for single-trial input-driven dynamics in a given cross-validation fold, we average the measured LFP power features

across trials in that cross-validation fold, i.e., across the test sets in that fold (Supplementary Fig. 5d). We then compare the predicted single-trial input-driven dynamics from forward prediction with these ground-truth single-trial input-driven dynamics. There are two sources for the error between the predicted single-trial input-driven dynamics and their ground-truth: (i) error in identifying the IO relationship: we term this IO identification error; (ii) model noise: this is not relevant to the input and is due to the noise in recordings or to intrinsic dynamics that are input-irrelevant. To see better why there are these two sources, we can decompose the overall brain network dynamics y_k (i.e., the measured LFP power feature time-series) into the sum of two parts (Fig. 7, also see Methods for details):

(i) Input-driven dynamics $y_{k,I}$ that are only driven by the stimulation input (equation (2) in Methods, also see Fig. 7). These dynamics are exactly repeated across the trials since the stochastic stimulation input is repeated across trials by design (Supplementary Fig. 5b).

(ii) Intrinsic dynamics $y_{k,N}$ that are that are not dependent on the stimulation input and thus in the model are only driven by the noise (equation (3) in Methods, also see Fig. 7). These dynamics change from trial to trial as they are input-irrelevant (driven by noise in the model).

Thus, a given forward-prediction error can originate from a combination of two possible sources:

(i) Source 1 – IO identification error: Our IO model identification can have error, leading to a mismatch between the fitted IO model and the actual IO relationship (this error could for example be due to a limited number of samples in fitting the IO model). This IO model identification error may be one source that contributes to the total forward-prediction error.

(ii) Source 2 – model noise: Our multi-trial experiment is designed to carefully assess the modelling of input-driven dynamics. As the input is repeated across trials, we average the measured overall brain network dynamics y_k across trials; this averaging keeps the input-driven dynamics $y_{k,I}$ intact given that they are the same for all trials, but reduces the amplitude of the intrinsic dynamics $y_{k,N}$ because they are independent from trial to trial (Supplementary Fig. 5d, also see Supplementary Note 6 for details). However, since we have a limited number of trials (10 to 30, see Supplementary Table 3) in our experiments, the intrinsic dynamics may not be completely removed by trial-averaging and thus may be another source that contributes to the total forward-prediction error. Thus, model noise is another source of forward-prediction error. Our IO models fit the noise that changes from trial to trial because these models are trained on concatenated single-trial data (Supplementary Fig. 5c). Thus, the noise in our models, i.e., the state noise term w_k and the observation noise term v_k, quantify this source of error within our simulations.

2. Worst-case and best-case scenarios for closed-loop control

A given value of forward-prediction error can be due to any combination of the above two sources. For the same value of forward-prediction error, closed-loop control performance is influenced by the source of this error as follows:

(i) The worst-case scenario for control is when the IO identification error (source 1) is as large as the entire amount of forward-prediction error, i.e., IO identification error = forward-prediction error. This is because in a closed-loop system, the feedback controller is

designed based on the identified IO relationship (note in our worst case closed-loop simulations, in addition to this worst-case IO identification error, we also include model noise on top of it, i.e. stochastic noises in the simulated brain as clarified below).

(ii) The best-case scenario for control is when there is no IO identification error and accordingly, the model noise is the only source of forward-prediction error. This is because our fitted IO models quantify the noise with the state noise term w_k and the observation noise term v_k and thus the controller can account for them.

In reality, the forward-prediction error likely comes from a combination of the two sources and thus the true control performance will lie somewhere between the worst-case and the best-case. Nevertheless, to show that control works even in the worst possible case for our fitted IO models, we run closed-loop simulations in which we both include the largest amount of IO identification error and include model noise on top of it as follows (Fig. 8a): (1) We make the IO identification error (source 1) as large as the entire amount of forward-prediction error; this is done by introducing mismatch between the fitted IO model and the simulated brain model; (2) We also keep the model noise (source 2) as large as what was fitted in the fitted IO model; this means the simulated brain also exhibits stochastic noises. In addition to the worst case, we also evaluate the best-case scenario by setting the IO identification error (source 1) to zero while keeping the simulated model noise (source 2) as large as what was fitted in the fitted IO model. This best-case is thus achieved by simulating a brain model that is equivalent to the fitted IO model but, at the same time, exhibits stochastic noises as large as fitted in our IO data.

3. Worst-case: simulated brain using model mismatches

Here we assume that IO identification error (source 1) is as large as the entire forward-prediction error, and that on top of the IO identification error we also have stochastic noises as large as fitted in our IO model (Fig. 8a). To simulate the worst-case, we use the fitted IO model from actual IO data to design the feedback controller but assume the brain model is actually different from this fitted model due to IO identification error, i.e., we introduce model mismatches. We make this mismatch large enough to get the same model explained variance (EV) as observed in our experiments even in the noiseless case (see section “Sources of the forward-prediction error” above). On top of this, we then also include as much noise in the simulated brain as was fitted based on the actual IO data; this means that the noise covariance in the simulated brain is set to be the same as the noise covariance in the fitted IO model.

In particular, we design the feedback controller using the fitted IO models from our data, which are given in equation (1) in Methods as (copied below for convenience):

$$\begin{cases} x_{k+1} = Ax_k + Bu_k + w_k \\ y_k = Cx_k + v_k \end{cases}.$$

To simulate the IO identification error, we use the following mismatched model to generate simulated brain activity from a mismatched simulated brain (Fig. 8a):

$$\begin{cases} x_{k+1} = Ax_k + (B + \delta B)u_k + w_k \\ y_k = Cx_k + v_k \end{cases}, \qquad \text{(S16)}$$

where A, B, C and the noise covariance of w_k, v_k are the same as those fitted from our data, and δB is a random Gaussian matrix representing the model mismatch term that leads to the IO

identification error. We include the mismatch in B because it is the input matrix, which directly dictates the effect of stimulation input on the brain activity.

To match the IO identification error with the forward-prediction error we see in our experimental data, we compute the noiseless ground-truth of input-driven dynamics for the mismatched simulated brain as

$$\begin{cases} x_{k+1,S} = Ax_{k,S} + (B + \delta B)u_k \\ y_{k,S} = Cx_{k,S} \end{cases}, \qquad \text{(S17)}$$

where $x_{k,S}$ is the input-driven latent state in the simulated brain and $y_{k,S}$ is the simulated input-driven dynamics. We then compute the forward prediction of the fitted IO model as in equation (4) in Methods (copied below for convenience)

$$\begin{cases} s_{k+1} = As_k + Bu_k \\ \hat{y}_k = Cs_k \end{cases},$$

where s_k is the forward-predicted latent state using the fitted IO model and $\hat{y}_k$ is the forward prediction of input-driven dynamics using the fitted IO model.

To simulate the worst-case scenario, we pick the amplitude of δB such that the IO identification error equals the forward-prediction error we see in real IO data (i.e. source 1 equals the entire forward-prediction error). More specifically, the variance of the simulated ground-truth input-driven dynamics, $y_{k,S}$, explained by the fitted IO model is given by

$$\mathrm{EV}(\delta B) = \left(1 - \frac{\frac{1}{L}\sum_{k=1}^{L}\left(y_{k,S} - \hat{y}_k\right)^2}{\mathrm{Var}(y_{k,S})}\right) \times 100\%, \qquad \text{(S18)}$$

which is a function of the model mismatch term δB since the simulated ground-truth input-driven dynamics $y_{k,S}$ is a function of δB. We can thus tune the amplitude of δB such that the simulated explained variance $\mathrm{EV}(\delta B)$ matches the explained variance in forward prediction we see in experimental IO data. Note that in addition to this mismatch that makes the IO identification error as large as the entire forward-prediction error, the simulated brain model in equation (S16) still also includes the noise terms w_k, v_k whose covariances are the same as those fitted from our actual IO data.

4. Simulated internal brain state to be controlled

In a given application, the IO model will be used in closed-loop control of a desired internal brain state, such as mood or pain, rather than the brain activity itself. Thus, to run the simulations, we need to simulate an internal brain state. The representation of an internal brain state in brain activity is likely contaminated by a large amount of noise, which likely leads to the observation that relatively subtle changes in brain activity induced by stimulation can lead to significant changes in behaviour/brain state (as we expand on in Supplementary Note 22 below). Motivated by our prior work on neural decoding of mood state in human subjects[22], we simulate an internal brain state m_k, such as a mood state, as a linear function of the latent state x_k in our IO model:

$$m_k = Tx_k, \qquad \text{(S19)}$$

where $T \in \mathbb{R}^{1\times N_x}$.

We define mood state as the score in validated self-reports that measure depression and anxiety symptoms[13,22,23]. In particular, the mood is simulated to have the same unit as a validated self-report mood questionnaire termed Immediate Mood Scalar (IMS)[144] that has been used in prior studies[13,22,23] and measures depression and anxiety symptoms. To simulate a diseased brain that has negative mood when there is no stimulation (e.g. in depression), we add a constant disturbance $d \in \mathbb{R}^{2\times 1}$ (in the same unit as the input stimulation u_k) to the simulated brain as in our prior simulation work[18]. Thus, the mismatched brain network activity is simulated as:

$$\begin{cases} x_{k+1} = Ax_k + (B+\delta B)u_k + w_k + (B+\delta B)d \\ y_k = Cx_k + v_k \\ m_k = Tx_k \end{cases}. \qquad \text{(S20)}$$

With this model structure and based on our prior work on mood state decoding[22], we construct a plausible simulated brain model that satisfies the following conditions:

(a) When there is no stimulation $u_k = 0$, the mood has a negative mean value of -40, which represents a depressed mood state. The mood is simulated to have the same unit as IMS[144].

(b) The mean value of stimulation (denoted as $\bar{u}$) that is needed to take the depressed mean value of mood to a therapeutic level (a mood value of 0) should be within a clinically reasonable range. We thus take this range to be within our MN parameter range: $[20, 40]\ \mu A$ for amplitude and $[50{,}100]\ Hz$ for frequency. This indeed corresponds to the range that has been shown to improve mood states in patients with depression traits[22].

(c) The simulated mood should be sensitive to the change of stimulation input. This is motivated by prior work showing 1) stimulation with constant parameters (i.e., amplitude and frequency) can significantly improve mood in patients with depression traits[13], and 2) the amount of improvement in mood is dose-dependent, i.e., depends on the value of stimulation parameters[13].

(d) The simulated mood should not have large natural/intrinsic variations around its mean since mood often refers to a pervasive and sustained emotional state[145,146], thus should not have large natural/intrinsic variations over time. Indeed, in general, the goal of a closed-loop controller of internal brain states is to control large abnormal pathological variations of these states rather than their natural/intrinsic variations that are relatively smaller in general.

The free design parameters that we can select to achieve the above plausible simulated brain model are T and d in equation (S20). Other parameters are either obtained from the fitted IO model in the actual experimental data (A, B, C and the noise covariance of w_k, v_k) or are dictated by the forward-prediction error in the actual experimental data (IO identification error determines the model mismatch δB as described above). We translate the above four requirements to the requirements on T and d as follows:

(1) When there is no stimulation $u_k = 0$, the mean value of mood $\bar{m}$ is related to the disturbance level d via[18]: $\bar{m} = T(I-A)^{-1}(B+\delta B)d$. Thus we require $T(I-A)^{-1}(B+\delta B)d = -40$ (see our prior simulation work for details[18]). This thus satisfies requirement (a) above.

(2) The mean value of stimulation to take the mean mood state to a target value of zero can be easily found as: $\bar{u} = -d$. Thus, we require the elements of $-d$ to be within clinically reasonable ranges of $[20, 40]\ \mu A$ and $[50{,}100]\ Hz$ (also our MN parameter ranges). This thus satisfies requirement (b) above.

(3) The sensitivity of the mean value of mood $\bar{m}$ to a constant input stimulation $\bar{u}$ is given by $\frac{\partial \bar{m}}{\partial \bar{u}}$. With a constant input stimulation $\bar{u}$, the mean value of mood can be computed as (see our prior simulation work for details[18]): $\bar{m} = T(I - A)^{-1}(B + \delta B)\bar{u} + T(I - A)^{-1}(B + \delta B)d$. Thus, $\frac{\partial \bar{m}}{\partial \bar{u}} = T(I - A)^{-1}(B + \delta B)$, and we want the 2-norm of $\frac{\partial \bar{m}}{\partial \bar{u}}$ to be such that we have enough sensitivity to the input. This means that we want the sensitivity defined as:

$$\left\| \frac{\partial \bar{m}}{\partial \bar{u}} \right\|_2 = \sqrt{T(I - A)^{-1}(B + \delta B)(B + \delta B)'(I - A)'^{-}T'}, \qquad \text{(S21)}$$

to be large enough ($\cdot'$ represents matrix transpose). We quantify this requirement in a cost function as shown below. This thus satisfies requirement (c) above.

(4) The variance of x_k (denoted by X) can be computed as a function of A and the noise covariance of w_k (denoted by W) and as the solution to the Lyapunov equation: $X = AXA' + W$[113]. Then the variance of mood $m_k = Tx_k$ is given by:

$$\mathrm{Var}(m_k) = T'XT. \qquad \text{(S22)}$$

We want the above variance of mood to not be large given that natural/intrinsic mood variations over time are not large. We quantify this requirement in a cost function as shown below. This thus satisfies requirement (d) above.

We next use a constrained optimization problem to find T and d in the simulated brain model by formulating a cost function and its constraints based on (1)-(4) above as:

$$\min_{T,d} \frac{1}{\text{sensitivity (i.e., equation (S21))}} + \text{natural mood variations (i.e., equation (S22))}$$

$$\text{subject to}$$

$$T(I - A)^{-1}(B + \delta B)d = -40,$$

$$-d^{(1)} \in [20, 40], -d^{(2)} \in [50{,}100],$$

$$T'XT < 40^2$$

The two terms in the cost function quantify the goals in (3) and (4) above to have sensitivity to stimulation input while not having large natural/intrinsic variations of mood, respectively. The first two constraints quantify the goals in (1) and (2) above, respectively. The last constraint is only introduced to prevent unreasonably large variations in mood, where the upper limit in the mood variance is set to be the square of the negative mood value without control. The actual standard deviation of natural/intrinsic mood variations resulting from the optimization was 16.6 ± 0.67 (mean ± s.e.m.), which rarely approached the upper bound of 40 and was close to the standard deviation of intrinsic mood variations observed in human data in our prior work[22] (15.3 ± 1.55).

5. Controller design using the mismatched IO model

The goal of the closed-loop controller is to take the simulated brain state $m_k = Tx_k$ from a negative value of -40 to a therapeutic level of 0. We design a linear-quadratic Gaussian (LQG) closed-loop controller based on the fitted IO model, which is mismatched with respect to the simulated brain model (see model mismatch term δB in equation (S20)). Note that the model mismatch term δB is not known in the controller design.

The details of the LQG controller is given elsewhere[137]. Briefly, the LQG controller consists of a Kalman state estimator and an optimal feedback controller in the form of a linear-quadratic regulator (LQR). The state estimator recursively estimates the latent state:

$$\hat{x}_{k+1} = A\hat{x}_k + Bu_k + K(y_k - C\hat{x}_k), \qquad \text{(S23)}$$

where K is the Kalman gain, which is a function of the fitted A, C and fitted covariance of w_k, v_k[137]. The LQR controller is a linear feedback controller:

$$u_k = -G(\hat{x}_k - x^*) + u^*, \qquad \text{(S24)}$$

where G is the LQR feedback gain matrix computed from the fitted IO model parameters A and B[137], $x^* = 0$ is the target value of the latent state to get a target value of 0 for mood (i.e., $Tx^* = 0$) and u^* is the mean value of input that is needed to reach the target value of mood. Based on the fitted IO model, the controller computes u^*, with the goal of taking the depressed mood from -40 to 0 and with the constraint that u^* should also be within the stimulation parameter range used in our experiments. The controller thus computes the mean value of input u^* with the following optimization problem:

$$\min_{u^*} \ (T(I - A)^{-1}Bu^* - 40)^2,$$

$$\text{subject to}$$

$$u^{*(1)} \in [20, 40], u^{*(2)} \in [50,100]$$

The key point here is that to solve for u^*, the controller only knows the fitted B, but not the model mismatch term δB in the simulated brain.

Finally, as the standard LQR solution does not bound u_k, we design the controller to keep u_k within clinically realistic ranges of $[0,100]$ μA and $[0,185]$ Hz used in prior clinical work with DBS[8,96]. For example, if the LQR amplitude or frequency is smaller than their range, the lower bound of the range is used and if they are larger than their range, the upper bound of the range is used.

6. Best-case closed-loop control simulation

Beyond the above worst-case scenario, we also consider the best-case scenario where we set $\delta B = 0$ in the simulated brain and assume the forward-prediction error is due to noise in the models (source 2; see section "Sources of the forward-prediction error"). Therefore, we repeat the above simulations by just assuming $\delta B = 0$.

7. Dissociating the effect of IO model for closed-loop control

Closed-loop control performance is dictated by two factors: IO model and feedback of brain activity. To show that the IO model is key in enabling closed-loop control, we compare the model-based feedback controller above with a model-free feedback controller that turns stimulation on and off based on feedback and does not need an IO model for its design. The on-off control law is

$$u_k = \begin{cases} u_{\text{on}}, & \text{if } T\hat{x}_k < 0 \\ 0, & \text{otherwise} \end{cases}, \qquad \text{(S25)}$$

Since there is no IO model, the stimulation amplitude and frequency for the on-state u_{on} is chosen randomly in each closed-loop control trial from the same range used for LQR control. Note that each trial of on-off control uses a different independent realization for u_{on}.

8. Closed-loop performance measure

We quantify the closed-loop control performance as the root mean square error (RMSE) between the controlled internal brain state m_k and the target level 0 as:

$$\text{RMSE} = \sqrt{\frac{1}{L}\sum_{k=1}^{L} m_k^2}, \qquad \text{(S26)}$$

where L is the duration of control. We compute the baseline performance as the performance when there is no stimulation, i.e., when $u_k = 0$. The baseline RMSE is denoted by $\text{RMSE}_{\text{no-stimulation}}$. We then compute a normalized closed-loop control error, which is the RMSE relative to the RMSE with no stimulation:

$$\epsilon = \frac{\text{RMSE}}{\text{RMSE}_{\text{no-stimulation}}} \qquad \text{(S27)}$$

Note that the baseline normalized error is thus 1.

9. Studying how closed-loop control performance changes with a change in the simulated forward-prediction error

We further study how the performance of the closed-loop controller changes if we change the IO model's forward-prediction error or EV/CC in simulations. This analysis further characterizes the effect of the IO model on closed-loop control performance. We perform this analysis again for both the worst-case scenario and the best-case scenarios as follows.

For the worst-case scenario, we sweep the amplitude of model mismatch δB to change the IO identification error such that the simulated forward-prediction EV changes from 5% (large IO identification error) to 100% (no IO identification error). At the same time, regardless of the δB value, we always keep the same amount of noise in the simulated brain as in the actual fitted IO model (i.e., the same noise covariances). This is similar to what we did for the worst-case simulation above except that now we sweep the model mismatch δB (see section "Worst-case: simulated brain using model mismatches").

For the best-case scenario, the IO identification error is always zero so $\delta B = 0$ and all the forward-prediction error comes from the model noise (source 2). Given that $\delta B = 0$ (no model

mismatch), the predicted single-trial input-driven dynamics is always the same as the ground-truth single-trial input-driven dynamics because forward prediction is unaffected by noise (see equation (4) in Methods). However, in our model evaluation, the ground-truth of single-trial input-driven dynamics need to be estimated by trial-averaging the measured overall brain network dynamics (i.e., the measured LFP power feature time-series). Thus, the noise in the measured overall brain network dynamics introduces error in the estimate of the ground-truth single-trial input-driven dynamics. Since the EV is computed by comparing this estimate of the ground-truth single-trial input-driven dynamics with the predicted single-trial input-driven dynamics, EV is a function of the noise in the measured overall brain network dynamics. The noise in the measured overall brain network dynamics is quantified by the covariance of the model noise w_k and v_k in the simulated brain in equation (S20). Therefore, to simulate different levels of EV in the best-case scenario, we scale the covariance of w_k and v_k such that the EV changes from 5% to 100%.

Details of Simulation Results

1. The fitted IO models achieve closed-loop control of the internal brain state

We found that we could achieve closed-loop control using our fitted IO models with the exact same forward prediction performance (CC/EV) as in real neural data. The normalized closed-loop control error in the best-case scenario was 0.18 ± 0.01, which was markedly smaller than the baseline error of 1 with no stimulation (two-sided Wilcoxon signed-rank test, $P = 6.7 \times 10^{-35}$, Figs. 8b and 8c). Even for the worst-case scenario, the normalized closed-loop control error was 0.42 ± 0.01, which was again markedly smaller than the baseline error of 1 with no stimulation (two-sided Wilcoxon signed-rank test, $P = 8.6\times10^{-34}$, Figs. 8b and 8c). Some example control traces are included in Fig. 8b. These results indicate that (1) the actual average normalized closed-loop control error will be between [0.18, 0.42], showing the ability of the fitted IO models to enable closed-loop control, (2) even in the worst-case scenario, the fitted IO models have the accuracy to do reasonably good control.

2. The fitted IO models are critical for closed-loop control

We show that IO models are critical for closed-loop control using three analyses:

(1) We found that model-based closed-loop control even in the worst-case scenario markedly outperformed model-free closed-loop on-off control that used feedback of brain activity (0.42 ± 0.01 vs. 0.80 ± 0.05, two-sided Wilcoxon signed-rank test, $P = 4.1\times10^{-24}$, Figs. 8b and 8c). The model-free closed-loop control performed on average nearly twice worse than the worst-case model-based closed-loop control. Some example control traces are included in Fig. 8b. This is intuitively sound: if we don't have an IO model, we don't know how stimulation would affect brain activity to even decide on the on-level of an on-off controller. This result shows that the IO models are critically needed to achieve closed-loop control.

(2) We found that closed-loop control performance improved with better IO identification in the worst-case scenario. The normalized closed-loop control error monotonically decreased as the IO model's forward-prediction error decreased, or equivalently as the EV increased (Spearman's rank correlation, Spearman's $P < 10^{-30}$, Fig. 8d). This result again confirms the importance of the IO model in enabling closed-loop control.

(3) As a control analysis and sanity check, in the best-case scenario, we also found that the normalized closed-loop control error monotonically decreased as the IO model's EV increased (Spearman's rank correlation, Spearman's $P < 10^{-30}$, Fig. 8d). As expected, for each simulated EV value, the normalized closed-loop control error in this best-case scenario was better than that in the worst-case scenario in (2) above. Note at 100% EV, while the model mismatch was zero for the worst-case, the model still did not achieve zero closed-loop control error due to stochastic noises in the simulated brain; these noises are always included in the worst case regardless of the EV level (see "Worst-case and best-case scenarios for closed-loop control" above). In contrast, the best-case for 100% EV corresponded to the case with no model noise and no model mismatch and thus achieved zero closed-loop control error (since the only source of forward-prediction error in the best-case is model noise, this noise needs to be zero for 100% EV). Taken together, for each EV value, the actual normalized closed-loop control error is expected to be between the best-case and the worst-case scenarios shown in Fig. 8d by the shaded green area.

Supplementary Note 21. Invertibility of the fitted dynamic IO models

The closed-loop simulations in Supplementary Note 20 show that our fitted IO models can achieve closed-loop control by guiding how stimulation amplitude and frequency should be changed and thus are invertible. To further show this, here we present a theoretical derivation to show the invertibility of the fitted dynamic IO models. The invertibility of the IO model is closely related to the notion of controllability of the IO model in control theory[138], which implies the following: If the IO model is controllable, we can control the latent state (and thus the internal brain state) to reach any value by determining a sequence of inputs as we expand on below.

As we fully expand on in Supplementary Note 20, the goal in closed-loop control is to control an internal brain state, which is a function of the latent state that describes the brain network activity, i.e. $x_k \in \mathbb{R}^{N_x \times 1}$. Therefore, the invertibility/controllability of the IO model is equivalent to saying that for any given initial latent state value x_0, the IO model can find an input sequence $\left[u_{K-1,} u_{K-2,\ldots,} u_0\right]$ that achieves any desired target latent state value x^* at time-step K, i.e., achieves $x_K = x^*$.

To show this, without loss of generality, assume that the initial state x_0 starts from 0. The case for non-zero initial state can be derived in the same way by simply shifting the origin[138]. We also consider the noise-free case as the noise only introduces variations around the target without affecting the ability to reach the target itself by using the input sequence (this is based on the separation principle for estimation and control in control theory[137]). Using the fitted input-driven IO model

$$x_{k+1} = Ax_k + Bu_{k,} \qquad \text{(S28)}$$

and by expanding the recursion in this model and setting $x_0 = 0$, we get

$$x_K = [B, AB, A^2B, \ldots, A^{K-1}B] \times \left[u_{K-1,} u_{K-2,\ldots,} u_0\right]'. \qquad \text{(S29)}$$

Therefore, as long as $M \stackrel{\text{def}}{=} [B, AB, A^2B, \ldots, A^{K-1}B]$ has full row rank, the needed input sequence $\left[u_{K-1,} u_{K-2,\ldots,} u_0\right]'$ can be found by inverting the model as

$$\left[u_{K-1,} u_{K-2,\ldots,} u_0\right]' = M^{\dagger} x_K. \qquad \text{(S30)}$$

Here $M^{\dagger}$ represents pseudoinverse (Moore–Penrose inverse) of matrix M. If M has full row rank, MM' is invertible and $M^{\dagger} = M'(MM')^{-1}$. We can easily verify that $M\left[u_{K-1,}u_{K-2,\ldots,}u_0\right]' = MM^{\dagger}x_K = MM'(MM')^{-1}x_K = x_K$. Thus, to achieve the target at time K, i.e., achieve $x_K = x^*$, the needed input sequence is

$$\left[u_{K-1,}u_{K-2,\ldots,}u_0\right]' = [B, AB, A^2B, \ldots, A^{K-1}B]^{\dagger}x^*. \quad \text{(S31)}$$

From the above theoretical derivations, we see that the invertibility of the IO model is the same as ensuring the full row rank of $[B, AB, A^2B, \ldots, A^{K-1}B]$. By the theory of controllability of linear systems[138], this is equivalent to ensuring the full rank of the following controllability matrix:

$$[B, AB, A^2B, \ldots, A^{N_x-1}B] \in \mathbb{R}^{N_x \times 2N_x}, \quad \text{(S32)}$$

i.e., the rank should be equal to the latent state dimension N_x. We computed the rank of the controllability matrix of all IO models we fitted, and they all had full rank. This result shows that all the fitted IO models are controllable and can indeed be inverted to compute the input sequence needed to move the latent state to any desired target value.

Supplementary Note 22. On the fidelity achieved by our IO modelling method

The closed-loop simulations in Supplementary Note 20 and the theoretical derivations in Supplementary 21 suggest that our fitted IO models have the accuracy to have utility for control of internal brain states. Here, we expand more on the fidelity achieved by our IO models. Further, we discuss the amount of stimulation-induced changes in brain network activity measured in prior experiments.

Model goal and fidelity

Our work presents the first demonstration of predicting large-scale multiregional brain network dynamics in response to ongoing stimulation. This large-scale prediction is needed especially for developing closed-loop stimulation systems for a broad range of neuropsychiatric disorders that involve multiple brain regions[2,24,25,86]. Prior models in other studies are designed for different goals. Prior biophysical models do not aim to predict the response within an individual or across general large-scale brain networks. Prior models fitted with experimental data study the static response after stimulation ends[41,42] or describe the columnar responses[38] or single-neuron response[39] within a particular brain region rather than large-scale multiregional brain networks. As these models are for different purposes, their results are not directly comparable to our large-scale dynamic prediction during ongoing stimulation. That is why we compare our dynamic IO model with four other models that we construct in this study and within our own experiments: (i) the static (non-dynamic) regression model, (ii) the smoothing model, (iii) the non-oscillatory model and (iv) a general nonlinear model (Figs. 4 and 6). We show that the dynamic IO model enables significantly better forward prediction of LFP power features compared with these four alternative models. Finally, even these alternative models of large-scale multiregional brain network response to ongoing stimulation were not studied before as the goals in prior studies were different.

Measured amount of stimulation-induced change in brain activity in prior experiments

As mentioned above, prior studies do not aim to perform model-based prediction of large-scale multiregional brain network dynamics in response to stimulation. Nevertheless, to provide some context, we look at the change in brain activity that some studies measure in response to a stimulation with constant amplitude and frequency.

These prior studies have shown that in many cases the change of neural features due to stimulation—whether after stimulation ends or during stimulation—can be relatively subtle yet still change the internal brain state and behavior. For example, De Hemptinne et. al.[44] show that the energy of the stimulation-induced change in human ECoG phase-amplitude-coupling after stimulation ends is ~40% of the energy of natural variations in ECoG phase-amplitude-coupling pre-stimulation, yet this relatively small change in brain activity can significantly alleviate symptoms of Parkinson's Disease in human patients. Ezzyat et al.[147] show that the energy of the stimulation-induced change in human ECoG gamma power after stimulation ends is ~40% of the energy of natural variations in ECoG gamma power pre-stimulation, yet can significantly modulate memory performance. Our prior work on the effect of stimulation on improving mood in epilepsy patients[13] examines the change of power features during stimulation and finds that the energy of the stimulation-induced change in human ECoG powers during stimulation is ~16% of the energy of natural variations in ECoG powers pre-stimulation, yet can significantly improve mood in epilepsy patients with depression traits. O'Doherty et al.[148] show that the energy of the stimulation-induced change in neuron firing rate is ~30% of the energy of natural variations in neuron firing rate pre-stimulation in non-human-primates, yet can induce significant tactile sensation that can be used as feedback in a real-time brain-machine interface.

Similarly, some prior studies without behavioral contexts have also shown subtle changes of brain activity or connectivity due to stimulation with constant amplitude and frequency. For example, Solomon et. al.[149] show that the energy of the stimulation-induced change in human ECoG powers after stimulation ends is ~25% of the energy of natural variations in ECoG powers pre-stimulation. Khambhati et. al.[41] show that the stimulation-induced change in the topology of human ECoG coherence network after stimulation ends is ~0.5% (~0.005 in the unit of coherence value between 0 and 1). Stiso et. al.[42] use a structural model using diffusion weighted imaging (DWI) data and find that the maximum correlation coefficient between their structural model and the change in human ECoG power features after stimulation ends is on the order of ~0.03 (in the unit of correlation coefficient between 0 and 1).

Finally, it is interesting discuss why subtle changes measured in brain activity due to stimulation can still coincide with a significant change in behavior due to stimulation. This is because the measured brain activity can have large input-irrelevant intrinsic dynamics as well as large dynamics that may not be relevant to the internal brain state/behavior of interest (e.g. mood) but to other brain functions. Also, the internal brain state of interest (e.g., mood) can be encoded in brain activity that has large noises on a given recording channel. Thus, even when stimulation significantly changes the internal brain state, the corresponding change in the recorded brain activity can be relatively small compared with the large noises and the behavior-irrelevant or input-irrelevant intrinsic dynamics on a given recording channel. Motivated by these observations, we have conducted a comprehensive closed-loop simulation to show that with the accuracy achieved by our IO model, we can achieve successful control of a simulated internal brain state, as we expand on in Supplementary Note 20.

Supplementary References

120. Calabrese, E. *et al.* A diffusion tensor MRI atlas of the postmortem rhesus macaque brain. *Neuroimage* 117, 408–416 (2015).
121. Haegens, S., Händel, B. F. & Jensen, O. Top-down controlled alpha band activity in somatosensory areas determines behavioral performance in a discrimination task. *J. Neurosci.* 31, 5197–5204 (2011).
122. Lin, J.-J. *et al.* Theta band power increases in the posterior hippocampus predict successful episodic memory encoding in humans. *Hippocampus* 27, 1040–1053 (2017).
123. Gola, M., Magnuski, M., Szumska, I. & Wróbel, A. EEG beta band activity is related to attention and attentional deficits in the visual performance of elderly subjects. *Int. J. Psychophysiol.* 89, 334–341 (2013).
124. Benchenane, K. *et al.* Coherent theta oscillations and reorganization of spike timing in the hippocampal-prefrontal network upon learning. *Neuron* 66, 921–936 (2010).
125. Marceglia, S. *et al.* Basal ganglia local field potentials: applications in the development of new deep brain stimulation devices for movement disorders. *Expert Rev. Med. Devices* 4, 605–614 (2007).
126. Brown, P. Oscillatory nature of human basal ganglia activity: relationship to the pathophysiology of Parkinson's disease. *Mov. Disord.* 18, 357–363 (2003).
127. Priori, A. *et al.* Deep brain electrophysiological recordings provide clues to the pathophysiology of Tourette syndrome. *Neurosci. Biobehav. Rev.* 37, 1063–1068 (2013).
128. Jerbi, K. *et al.* Task-related gamma-band dynamics from an intracerebral perspective: Review and implications for surface EEG and MEG. *Hum. Brain Mapp.* 30, 1758–1771 (2009).
129. Gründemann, J. *et al.* Amygdala ensembles encode behavioral states. *Science* 364, eaav8736 (2019).
130. Allen, W. E. *et al.* Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. *Science* 364, 253–253 (2019).
131. Okun, M., Steinmetz, N. A., Lak, A., Dervinis, M. & Harris, K. D. Distinct structure of cortical population activity on fast and infraslow timescales. *Cereb. Cortex* 29, 2196–2210 (2019).
132. Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science* 364, 255–255 (2019).
133. Qian, S. *Introduction to time-frequency and wavelet transforms*. 68, (Prentice Hall PTR Upper Saddle River, NJ: 2002).
134. Grossmann, A., Kronland-Martinet, R. & Morlet, J. Reading and understanding continuous wavelet transforms. *Wavelets* 2–20 (1990).
135. Elder, C. M., Hashimoto, T., Zhang, J. & Vitek, J. L. Chronic implantation of deep brain stimulation leads in animal models of neurological disorders. *J. Neurosci. Methods* 142, 11–16 (2005).
136. Knijnenburg, T. A., Wessels, L. F., Reinders, M. J. & Shmulevich, I. Fewer permutations, more accurate P-values. *Bioinformatics* 25, i161–i168 (2009).
137. Bertsekas, D. P. *Dynamic programming and optimal control*. 1, (Athena scientific Belmont, MA: 1995).
138. Chen, C.-T. *Linear system theory and design*. (Oxford University Press, Inc.: 1998).
139. Tulleken, H. J. Generalized binary noise test-signal concept for improved identification-experiment design. *Automatica* 26, 37–49 (1990).
140. Opsahl, T., Agneessens, F. & Skvoretz, J. Node centrality in weighted networks: Generalizing degree and shortest paths. *Social networks* 32, 245–251 (2010).

141. Newman, M. E. Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality. *Phys. Rev. E* 64, 016132 (2001).
142. Donnat, C., Holmes, S. & others Tracking network dynamics: A survey using graph distances. *Ann. Appl. Stat.* 12, 971–1012 (2018).
143. Bastos, A. M. & Schoffelen, J.-M. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front. Syst. Neurosci.* 9, 175 (2016).
144. Nahum, M. *et al.* Immediate Mood Scaler: tracking symptoms of depression and anxiety using a novel mobile mood scale. *JMIR mHealth and uHealth* 5, e44 (2017).
145. *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. (American Psychiatric Publishing, Arlington: 2013).
146. Russell, J. A. Core affect and the psychological construction of emotion. *Psychol. Rev.* 110, 145 (2003).
147. Ezzyat, Y. *et al.* Direct brain stimulation modulates encoding states and memory performance in humans. *Curr. Biol.* 27, 1251–1258 (2017).
148. O'Doherty, J. E. *et al.* Active tactile exploration using a brain–machine–brain interface. *Nature* 479, 228 (2011).
149. Solomon, E. *et al.* Medial temporal lobe functional connectivity predicts stimulation-induced theta power. *Nat. Commun.* 9, 4437 (2018).