Precision data-driven modeling of cortical dynamics reveals idiosyncratic mechanisms underlying canonical oscillations

Matthew F. Singh[a,b,c], Todd S. Braver[b], Michael W. Cole[c], and ShiNung Ching[a]

[a]*Electrical and Systems Engineering, Washington University in St. Louis, St. Louis, 63130, MO, USA*
[b]*Psychological and Brain Science, Washington University in St. Louis, St. Louis, 63130, MO, USA*
[c]*Center for Molecular and Behavioral Neuroscience, Rutgers University, Newark, 07102, NJ, USA*

Abstract

Task-free brain activity affords unique insight into the functional structure of brain network dynamics and is a strong marker of individual differences, but has proven difficult to model at fast time-scales. Previous approaches have either used normative models, in which only a small number of parameters need to be estimated, or relied upon statistical characterizations rather than true generative (predictive) modeling. By contrast, in this work, we present a novel optimization framework that makes it possible to directly invert brain-wide models from single-subject recordings. This technique provides a powerful neurocomputational tool for interrogating mechanisms underlying individual brain dynamics ("precision brain models") and make quantitative predictions. We extensively validate the models' performance in forecasting future brain activity and predicting individual variability in key M/EEG markers. Lastly, we demonstrate the power of our technique in resolving individual differences in the generation of alpha and beta-frequency oscillations. We reveal a powerful characterization of subjects based upon models' attractor topology and a dynamical-systems mechanism whereby these topologies generate individual variation in the expression of alpha vs. beta waves.

Keywords: Resting-State, Dynamical Systems, Neural Mass Model, Individual Differences, MEG, EEG

1. Introduction

A key goal of human neuroscience is to decipher how individual differences in brain signaling and dynamics relate to individual differences in cognition and behavior. Developing mechanistic models of individual human brains is one part of this endeavor. While considerable efforts have been directed at identifying individual differences at macroscopic spatial scales, e.g., by means of functional neuroimaging, individual-differences in fast dynamical interactions at the scale of M/EEG are less understood. Such dynamics reveal neural computation at a timescale commensurate with sub-second cognitive operations and are strongly nonstationary. There is a long and rich history of analysis of fast neural electrophysiology, yet the mechanisms and functional salience of many commonly observed phenomena, such as EEG oscillations, remains debated. Generative mechanisms of the alpha wave, for instance, are not yet resolved with competing accounts of either a thalamic ([1, 2]) or cortical origin ([3]). That these phenomena tend to vary across individuals underscores the need for reliable, biologically-plausible dynamical systems models with sufficient expressiveness so as to reveal individual mechanistic differences.

1.1. Challenges in Data-Driven Individualized Modeling

Dynamical systems models are premised upon describing how components of a system interact to shape its future. The classical example of such models in neuroscience is, of course, the Hodgkin-Huxley ([4]) model of neuronal spiking. The power of these models is that they are simultaneously descriptive and mechanistic. That is, they not only describe features of the observed phenomena but also provide an underlying generative process that produces those phenomena. This property means that dynamical systems models can potentially predict how the system will respond under novel perturbations. This ability is, of course, based on the accuracy of the underlying model, which in turn depends upon how the model is constructed.

Data-driven approaches to dynamical systems modeling attempt to use measured brain activity to parameterize (i.e., 'fit') a model. At the scale of large-scale, whole-brain activity, there have been significant efforts directed towards this problem in the context of functional neuroimaging. These previous approaches to individualized brain modeling include methods that directly estimate parameters from fMRI recordings, such as Dynamic Causal Modeling ([5]) and methods in which the primary model parameters (e.g., connectivity) are adopted from structural imaging (e.g. [6, 7, 8]). In the former, the prime difficulty has been the development of methods to estimate large, nonlinear models from noisy, indirect timeseries. In recent work, we developed a new algorithm termed Mesoscale Individualized NeuroDynamic modeling (MINDy, [9, 10]) to estimate individualized brain models from fMRI timeseries. The general method consisted of a novel optimization framework to simultaneously estimate brain network parameters and, in an extended algorithm ([10]), local hemodynamic responses. However, the temporal resolution of fMRI greatly limits its

ability to inform models of the fast, transient interactions thought to dominate neural computation. In the current work we aim to reveal these mechanisms at the individual level and are therefore concerned with high-temporal resolution modalities. As our approach consists of whole-cortex modeling, we emphasize the use of MEG or EEG (M/EEG) as functional data-sources, as opposed to lower-coverage invasive techniques (e.g., ECoG, LFP, and SEEG), although the proposed method is general. In either case, the fast timescale context produces a new set of challenges for individualized modeling that evade current optimization methods, including the original MINDy framework ([9]).

There are theoretical, biophysical, and computational barriers to this endeavor. At the theoretical level, fast electrophysiological activity such as oscillations are hypothesized to arise from the interplay of excitatory and inhibitory neurons ([11, 12]), meaning that any biologically interpretable model of brain oscillations must consider the interactions between specific neuron types including long-distance projections onto either cell-type. In addition, asymmetric patterns of connectivity (feed-forward vs. feed-back) are also believed critical to generating lower-frequency oscillations ([13]). However, neither of these two features is accessible using structural (diffusion) imaging. Functional data is similarly limited in directly assessing these differences using conventional analysis of M/EEG signals. Both modalities are thought to be driven by cortical pyramidal cell activity as interneuron geometry is not conducive to dipole generation ([14]). From an optimization (model-fitting) standpoint, these limitations mean that neither the model states (excitatory and inhibitory neural activity) nor model parameters are directly accessible, posing a challenging dual-estimation problem. To address these difficulties, we present a new framework to directly estimate detailed neural-mass style models (Fig. 1A) from fast functional data (M/EEG). To be clear, by 'directly estimate' we mean optimizing every component of the neural model to predict observed activity (timeseries measurements), within subject. The net result of our framework will encompass: (i) the estimation of latent activity in neural populations across the cortex, (ii) separate brain-wide 'connectomes' for excitatory and inhibitory targets (Fig. 1A), and (iii) direct estimation from single-subject recordings.

We will proceed to introduce the framework and validate it on the human connectome project dataset. Furthermore, we will highlight the mechanistic explanatory power of the method in the context of cortical oscillations, by studying individual variability in generative processes underlying EEG oscillations. Such oscillations are among the most frequent features extracted from EEG, yet their underpinning and cognitive significance remains an open question, in part due to their variable expression across individuals. Here, we show that this variation may reflect low-dimensional dynamics that we link with a greater ratio of excitation-to-inhibition. We suggest that individual variation in alpha-band frequencies are reflected in protracted, global rhythms, whereas variation in the beta-band is linked to transient dynamics.

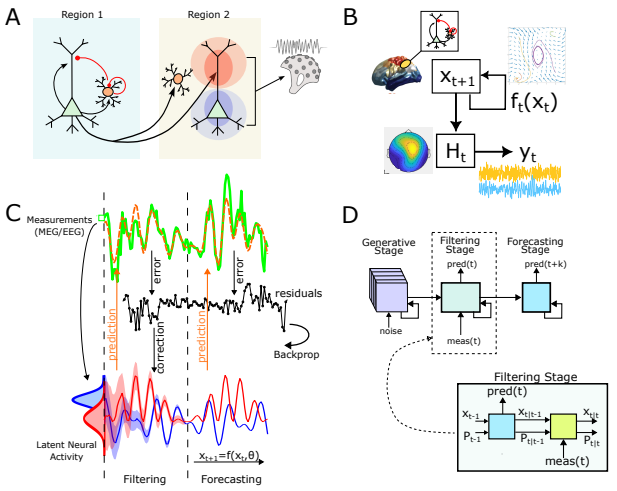

Figure 1: Schematic description of the proposed framework. A) Expanded neural-mass type model with long-distance connections onto both excitatory and inhibitory targets originating from distal pyramidal cells. Local EI circuits are fully connected. Arrows/dots indicate (directed) termination site. Dipoles are modeled as proportional to excitatory (pyramidal) depolarization at the soma. B) Relationship between components of the combined state-space and measurement models. Neural activity ($x(t)$) evolves according to the dynamics $f(x)$. Measurements (y_t) are produced by multiplying neural activity with the measurement matrix H_t. C) The gBPKF algorithm consists of three stages: generating baseline distributions, a Kalman-Filtering stage to estimate latent states (red and blue), and a Forecasting stage which predicts future brain activity measurements (green). Distributions (bottom left) indicate the posteriors at t_0. Note that the uncertainty (shading) decreases over time as the (nonlinear) Kalman filter corrects state-estimates. D) The algorithm instantiated as a nonlinear recurrent network. The generative (noise-driven) and forecasting (deterministic) layers evolve as conventional recurrent networks, whereas the filtering stage uses the Kalman filter to evolve both activity/states and uncertainty/covariance. Steady-state distributions from the generative stage are used to estimate the initial state/uncertainty for t_0.

2. Results

2.1. Solving the Dual Estimation Problem

The problem we are faced with consists of estimating both the states and parameters of an unknown dynamical system based upon two pieces of information: the measurement/forward model and the functional form of the dynamical system (neural-mass equations, Sec 6.1). Such problems is referred to as "dual-estimation" and, in general, are quite challenging in any circumstance. The application to brain-network modeling further challenges existing dual-estimation approaches, which become computationally-intractable due to the large number of unknown parameters, primarily in terms of network connections. Conventional source-reconstruction approaches (state-estimation) are also limited for this application as they are not sensitive to the activity of inhibitory neurons, as these cells do not generate strong dipoles. Rather, inhibitory neuron activity is reflected through its influence on M/EEG dynamics. In this work, we approach data-driven brain modeling using our Generalized Backpropagated Kalman Filter algorithm (gBPKF [15]; also see related earlier work by [16]).

2.1.1. Conceptual Motivation

We first motivate our algorithm conceptually, with the general solution saved for Sec. 6.4. Formally, we are faced with a system of the form:

$$x_{t+1} = f(x_t, \theta) + \epsilon_t \tag{1}$$

with latent neural activity x evolving according to model f parameterized by θ and ϵ denoting internal, physiological stochasticity (process-noise). We assume a functional form for f (a neural-mass model) but none of the variables in this equation are known or directly accessible. Rather, we must rely upon measurements y_t acquired from a noisy transformation H of the underlying neural activity (Fig. 1B):

$$y_t = Hx_t + noise \tag{2}$$

In the M/EEG context, the measurement matrix H is not invertible, even with source-localization, because x contains the activity of both excitatory and inhibitory neurons, the latter of which do not contribute to the M/EEG signal. Hence H takes the form: $[H_{Exc} \;\; 0_{k\times n}]$ for k measurement channels and n populations. The transformation of excitatory activity H_{Exc} could be direct (using the lead-field matrix for H_{Exc}) or, if data is already source-localized, $H_{Exc} = I_{k=n}$. However, in either case, the unknown latent activity of all populations x_t is not directly recoverable from y_t. Thus the dual-estimation results from the combination of two-problems: estimating states x_t and parameters θ. Our approach derives from the observation that both halves of this problem are individually well-studied and tractable but cannot be applied in isolation (solving for x_t requires θ and vice-versa). Instead we remove the problem of estimating latent brain-activity by replacing x with its pseudo-optimal estimate given θ and previous measurements: $\hat{x}_t := \mathcal{K}_t(\theta)$ which produces a conventional parameter-estimation problem (solve for θ, Fig. 1C). In other words, rather than treating states and parameters as unknown variables, we first define the best estimate of state, given parameter, and then solve for parameters which optimize this function. In practice, we use nonlinear-variants of the Kalman filter ([49, 50]) for the state estimate and attempt to minimize the prediction-error with respect to future measurements y_{t+k}. In short, we solve for parameters that generate the most-accurate Kalman filter. To retrieve these parameters, we treat the Kalman-filter recursions like a recurrent network (Fig. 1D) and analytically backpropagate error gradients through the entire algorithm (Sec. 6.4, Fig. 1C) which we combine with gradient/Hessian optimization. The accuracy and efficiency of this algorithm have already been established ([19]).

2.2. State-of-the-Art Model Identification with gBPKF

We first validated our optimization technique for solving dual-identification problems in which neither the system's states nor parameters are directly available. Recently, we validated this approach to the generic case of estimating an arbitrary recurrent neural-network model ([19]). The generic case established the general validity of our approach as an optimization method that could be used to fit a wide-variety of neural circuit models. Arbitrary circuit models are easily implementable

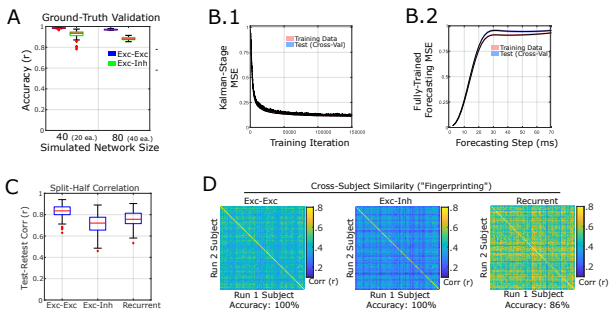

Figure 2: Validation of the current framework. A) Our method accurately recovers the excitatory-excitatory and excitatory-inhibitory connection strengths in realistic ground-truth simulations containing either 40 or 80 total populations (20 or 40 regions x 2 populations per region). B) Forecasting error converges and indicates similar performance on training and left-out data, meaning that models are stable and generalizable to new data within-subject. B.1) Change in error during the Kalman Filtering phase across training iterations during model estimation. B.2) Performance in forecasting future brain activity for various lags after 150k training iterations. Lines indicate the mean loss across subjects/runs and shading indicates standard error of measurements. C) Connectivity parameters are reliable across models trained on different data for the same subject. D) Model parameters are individual-specific, forming a unique "fingerprint" [20] that matches parameters fit to different data from the same subject. Accuracy indicates the percent of successful identifications (i.e., how often two models from the same subject are most similar, as opposed to another subject's model).

with the accompanying code. In the current work, we focus validation upon the specific biologically-plausible ground-truth models that we use for the remainder of this paper. This class is defined by three general rules:

- Each brain area contains an excitatory and inhibitory population
- Only excitatory populations generate long distance connections
- Measurements only reflect excitatory activity

The last condition, that measurements only reflect excitatory activity, makes the optimization particularly difficult. By contrast, in previous work ([19]), we bench-marked accuracy and efficiency of gBPKF against alternative methods for an unconstrained recurrent neural network model. Our results indicated that conventional approaches quickly became both untenable and inaccurate for large networks ([15]). The current set of analyses thus focused upon determining whether gBPKF remains accurate under the aforementioned biological constraints. The objective was to estimate the connectivity matrix W, containing both recurrent and long-distance connections, given the observed timeseries with all other parameters known. We generated simulated networks with either 20 or 40 regions each containing an excitatory and inhibitory population (40 or 80 total populations). The symmetric connection graph had a sparsity of 25% non-zero with the same graph used to define admissible EE and EI connections. The ratio of simulated channels to regions (75%) was based on the empirical rank of leadfield matrices (typically 65-80) relative the 100 parcels we later use.

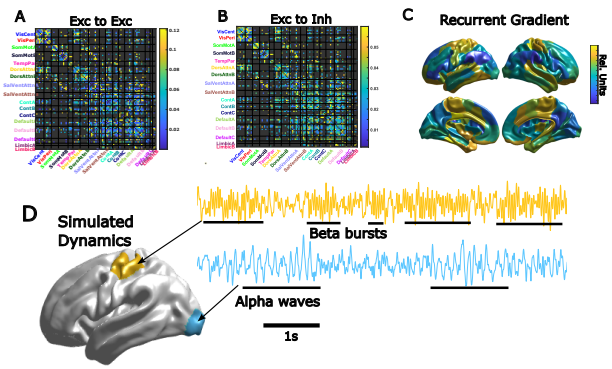

Figure 3: Group-average parameters and representative model dynamics. A) Group-average excitatory-to-excitatory connection matrix sorted by the Yeo-17 networks [21]. Dark-grey connections denote those which are not admissible as determined by the connectivity mask (see Sec. 6.6). B) Same as A) but for excitatory-to-inhibitory connections. C) Group-average spatial gradient of recurrent connections. Note that the separate recurrent connection types (EE,EI,II,II) are derived from affine transformations of this gradient (see Sec. 6.7). D)Model-simulated activity motor (top) and visual (bottom) parcels of a representative HCP MEG subject. We note that the simulated brain activity is highly non-stationary and spatially heterogeneous. We highlight the spontaneous generation of narrow-band bursts (beta and alpha, respectively) interspersed with wider-band oscillations.

Results demonstrate high performance in recovering ground-truth connectivity parameters in biologically-plausible simulations (Fig. 2A). We observed high-performance in recovering the true connectivity in both the 40 (EE : r = .983 ± .005, relative-MSE:.021 ± .006; EI : r = .919 ± .048, 60 sims) and 80 population conditions (EE : r = .970 ± .004, relative-MSE:.029 ± .003, EI : r = .884 ± .014, 30 sims). The performance for excitatory-excitatory connections was particularly strong (Fig. 2A). This advantage is expected, as the excitatory populations directly contribute to the simulated M/EEG signal whereas the effects of EI signaling are only indirectly observed through their delayed propagation along local IE coupling. However, despite this challenge, performance remained high. We conclude that our gBPKF algorithm is well-suited to estimate neural model parameters for all connectivity types (EE, EI, etc.).

2.3. *Learning Stable Estimates of Individual Mechanisms Underlying Brain Dynamics*

Empirical models were fit using the HCP MEG data which contains 15 total minutes of resting-state ([23]) which we divided in split-halves. We first analyzed the reliability of model parameters. For univariate parameters, we measured reliability in terms of the Intra-Class Correlation (ICC) which assesses reliability of individual differences. For multivariate parameters we present both the conventional test-retest correlations for overall similarity of the parameter and, for the reliability of individual differences, the Image Intra-Class Correlation (I2C2, [22]) which is a multivariate extension of ICC.

The primary parameters of interest, for reliability, are the connectivity parameters. Combined together, the connectivity matrix is highly reliable and individualized. Excitatory-excitatory long-distance connections had exceptional test-retest correlations (r = .83 ± .07, $I2C2$ = .72; Fig.2C). Long-distance excitatory-inhibitory connections also had good test-retest correlations (r = .71 ± .1) but were more modest for individual differences ($I2C2$ = .46). We also found good reliability for the spatial gradient of recurrent connections (r = .75 ± .10). Due to co-dependency (see Sec 6.7), the test-retest correlation, but not the I2C2 is the same for all recurrent connection types. Individual differences in recurrent connections were higher for inhibitory targets than excitatory: I2C2: EE = .58, IE = .56, EI = .85, II = .86. Fingerprinting accuracy was 100% for both EE and EI connectomes (Fig. 2D) and 84% for the spatial-gradient of recurrent connections.

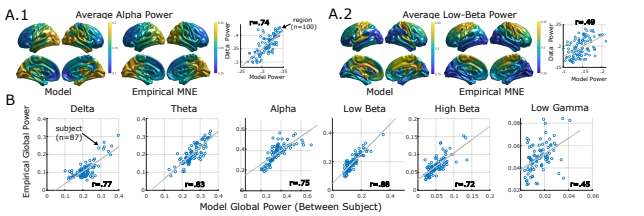

Figure 4: Precision brain models replicate empirical spatiospectral patterns. A) Group-level anatomical distributions of spectral power strongly correlate between model-predictions and source-localized MEG for two of the stereotyped M/EEG bands at rest: alpha (A.1) and low-beta (A.2). B) Individual differences in whole-brain spectral power (averaged over parcels) are reproduced across frequency bands by individiualized brain models.

Individual differences in decay rates had good reliability for excitatory populations (ICC = .70) and moderate for inhibitory populations (ICC = .64). The nonlinear connection gain (S^E) also had good reliability (ICC = .72). Estimated noise standard deviations ($\sqrt{Q}$) often saturated (lower-bound 0.1, upper-bound 0.3), and were, therefore, poor markers for individual differences ($\sqrt{Q^E}$: ICC = .34, $\sqrt{Q^I}$: ICC = .29). As expected (Sec. 6.7), reliability was low for the nonlinear threshold (v^E : $I2C2$ = .36) and baseline drive terms (c^E : $I2C2$ = .35, c^I : $I2C2$ = .12). While the inclusion of these parameters is important for reproducing the correct dynamics, they contribute via their values relative each-other (after a transformation) and, depending upon the forward model, may not be individually unique (see Sec. 6.7 for discussion). Group-average values for connectivity parameters and the spatial gradient of recurrent connections are displayed in Fig. 3A-C.

2.4. *Models are Explanatory with Respect to Brain Oscillations*

We next analyzed the model dynamics (see Fig. 3D for an example timeseries). We first tested our models' ability to correctly replicate the anatomical distributions of spectral power within the data (we refer to these as "spatiospectral features" for brevity). We divided the spectrum as follows: delta (1.5-4 Hz), theta (4-8 Hz), alpha (8-15 Hz), low beta (15-26 Hz), high beta (26-35 Hz), and gamma (>35 Hz) in accordance with the HCP MEG pipelines ([23]). We normalized spectral power to have a sum of one across bands, within subject, for both data and models. In all spatial comparisons we use the

precalculated source-level empirical estimates provided with the HCP ICA-MNE pipeline. We note that these analyses are not direct ground-truth tests, since the source-level empirical estimates are themselves limited in spatial resolution and likely overestimate smoothness. Hence, we do not expect exact agreement on parcel-level values, particularly near the boundaries between brain networks. At these boundaries, models produce much sharper spatial divisions than the empirical source estimates. We pay special attention to the alpha and beta spectral bands as these are most prominent in resting-state.

We first analyzed results at group level by comparing the group-average anatomical-profile of spectral power between model simulations and empirical estimates. Results demonstrate good spatial agreement for the alpha band ($r(98) = .74$; Fig. 4A.1) and moderate agreement for the beta bands ($r(98) = .49$, $r(98) = .42$, respectively; Fig. 4A.2). We note that the model-predicted low-beta is highly localized to the somatomotor network, compared to the blurrier source-estimates. The slow delta band also exhibited high similarity between data and models ($r(98) = .64$). By contrast, the theta band was moderately consistent: $r(98) = .52$.

We next examined the model's fidelity in replicating oscillatory dynamics at the individual level. At the coarsest level, we found that models strongly reproduce individual differences in global spectral power (whole-brain average) across spectral bands in the training data (delta through high-beta: $p's < 2.2E\text{-}15$, Fig. 4B). Interestingly, we also found a moderate correlation between individual differences in the low-gamma band ($r(85) = .45, p < 1.1E\text{-}5$, Fig. 4B) despite its suppression in training-data due to 30Hz low-pass filtering (excluding the full band). However, in contrast to other bands, the magnitude of model-predicted low-gamma power was significantly smaller than the HCP source-estimates and the average spatial profile was not consistent with data ($r(98) = .07$), so this result may simply reflect residual gamma-power retained in training data even after filtering.

We also found that models replicated individual differences in spectral power at the network-level for the main resting-state bands (alpha, low-beta). To compute network-level power, we averaged spectral power among parcels belonging to each of the 17 Yeo ([21]) resting-state networks as implemented in the Schaefer 100-17network parcellation ([24]). We correlated model-predicted and empirical power in a multilevel model with a fixed-effect of subject (global power) collapsed across all networks. We found the strongest similarity between model-predictions and data for the alpha ($r(1390) = .54$), low-beta ($r = .50$) and delta ($r = .43$) bands. While all statistical models were significant (due to the large number of data points), similarity was weak in the theta ($r = .35$), high-beta ($r = .32$), and gamma ($r = .10$) bands. These results suggest that, for the main resting-state bands (alpha, low-beta), models correctly predict network-level power at the single subject-level. However, despite high accuracy in predicting global power (see above), models are less accurate at predicting

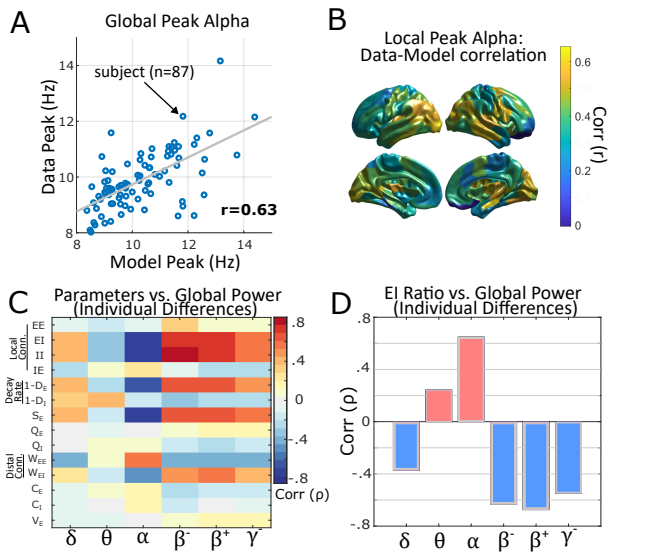

Figure 5: Models predict individual's peak frequency within the alpha band and explain individual differences in power. A) Correlation between predicted and empirical global-peak alpha (average over parcels). B) Model predictions of local peak-frequency are most accurate in posterior regions in which the alpha rhythm is dominant. C) Spearman correlation matrix between model parameters (see Sec 6.1 for definitions) and global power in each frequency band. D) Correlations between individual differences in the ratio of excitatory and inhibitory activity and global power by frequency band.

the anatomical/network sources of high-beta and theta-band power.

As a final validation, we examined whether models predict individual variation in the peak-frequency of the alpha band. Empirically, this characterization has proven a remarkably stable and predictive measure of individual differences in brain and behavior ([25, 26]). The functional significance of "peak-alpha" is not yet resolved, although many accounts posit that the width of an alpha-oscillation determines the temporal window over which phase-linked processes (e.g., information integration) occur ([27, 28]). We found that models were accurate at reproducing the global individual alpha-frequencies ($r(85) = .625, p < E\text{-}10$; Fig. 5A). At the parcel-level, we found that model accuracy was highest in predicting peak-alpha within posterior cortex which agrees with the anatomical expression of alpha power (Fig. 5B).

2.5. *Power, but not peak-frequency maps onto global patterns of recurrent connectivity*

Generative models, as we present here, can form predictions using either overt linkages to model-parameters or as emergent phenomena generated by the dynamics. We therefore tested whether individual differences in spectral power or peak-alpha are correlated with individual parameters embedded in the models (Fig. 5C). Between subjects, we found that connections onto inhibitory populations (local EI, local II, and distal EI) had a strong negative correlation with alpha-power ($\rho(85) = -.73, -.75, -.51$; p's<E-6, respectively), and a positive correlation with beta-power (low beta: $\rho = .74, .78, .48$; p's<3E-6, high-beta: $\rho = .72, .76, .67$; p's<2E-8). This relationship

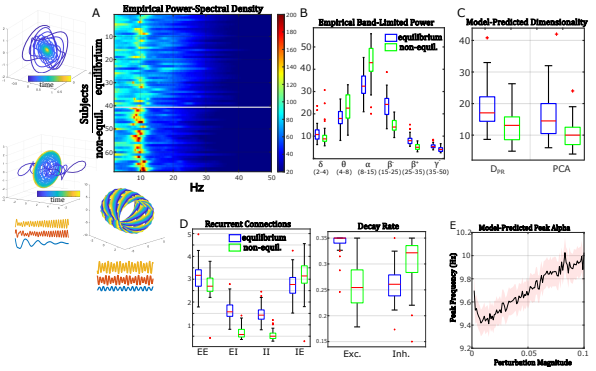

Figure 6: Models identify a taxonomy of subjects based upon attractor geometry. A) Global power-spectral density for all subjects in which models were consistent in producing equilibrium or non-equilibrium dynamics. Left side shows example attractors for three subjects, one with an equilibrium spiral-point attractor (top) and two with non-equilibrium attractors (a limit-cycle in bottom left and quasiperiodic torus in the bottom right). All attractors are projected onto 3-dimensional coordinates identified by PCA. B) Band limited power for each group of subjects. C) "Dimensionality" of model dynamics by attractor group as calculated using either Participation-Ratio Dimension or thresholded-PCA (Sec. 5.6). D) Models predict that attractor geometry is associated with alterations in local excitatory-inhibitory balance and timescales. E) Models predict that alpha oscillations become faster (higher-peak frequency) under perturbation.

also held for the excitatory decay-rate (alpha:$\rho = -.64$, low-beta: $\rho = .64$, high-beta: $\rho = .64$; p's<5E-11) and was reversed for distal EE connections (alpha, $\rho = .54$, low-beta: $\rho = -.43$, high-beta: $\rho = -.42$, p's<6E-4). Relationships were similar to alpha, but weaker, in the theta band and similar to beta in the low-gamma band (Fig. 5C). These parameter-level relationships suggest that the global excitation-inhibition ratio changes spectral power in the low vs. high frequency bands. Motivated by these parameter-level relationships, we quantified the model-predicted excitation-inhibition ratio for each parcel using the standard-deviation of simulated time-series: $\sigma(x^E)/\sigma(x^I)$. This estimate was highly reliable (for brain-wide average: $ICC = .78$) with much larger variation between-subject ($\sigma = 1.46$) than between-region ($\sigma = .10$), hence we only further investigated the brain-wide average due to small anatomical variation. In agreement with the aforementioned parameters, the predicted EI-ratio was positively correlated with global alpha-power ($\rho = .65, p \approx 0$, Fig. 5D), but negatively with beta power (low-beta: $\rho = -.63$, high-beta: $\rho = -.67$, p's≈ 0). Results thus indicate that individual differences in models' excitation-inhibition ratio predict empirical power in the higher-frequency bands. This result is complementary, but not identical, to theoretical models suggesting a relationship between excitation-inhibition ratio and the 1/f slope of power-spectral density ([29, 30]). Interestingly, however, these relationships only held at the global scale. Averaging across subjects, neither the model-predicted excitation-inhibition ratio, nor the strength of distal connections (EE, EI) predicted the anatomical distribution of spectral power for any band (max $\rho(98) = .24$, n.s.). The spatial gradient of local-recurrent connections was weakly correlated with gamma-band power $\rho(98) = .31, p = .040$ post-Bonferroni correction) and non-significant for all other bands. We did not find any significant relationships between model parameters/excitation-inhibition ratio and peak alpha frequency in terms of individual differences or anatomy. Thus, individual parameters strongly predict individual differences in global beta vs. alpha power, via the excitation-inhibition ratio but not the anatomical profile of spectral power or the peak alpha frequency. These features are, however, readily generated by the models (see above) and thus reflect emergent properties generated by model dynamics.

2.6. *Alpha Oscillations Branch-Off of Low-Dimensional, Intrinsic Rhythms*

Lastly, as a proof of concept, we used the models to interrogate the dynamical properties of two resting-state oscillations: alpha waves and beta waves (low-beta and high-beta). As previously mentioned, alpha rhythms (defined as 8-15Hz for HCP, but more commonly 8-12Hz) are high-amplitude posterior oscillations, often present at rest, which are amplified during eye closure and by tasks that require visual inhibition, broadly defined. While alpha constitutes one of the first discovered waking-EEG components, theories regarding its origin and mechanism have changed dramatically over the past two decades. Initial theories, based upon the eye-closure effect, posited that alpha represented a default "cortical idling" state to which the visual system would return, absent environmental stimuli (see [31, 32] for review), potentially driven by thalamic nuclei ([1, 2]). By contrast, alpha activity is now largely interpreted through the lens of preparatory attention ([33]) and active visual inhibition (e.g. [34, 35]). Several recent results have also indicated the potential for cortically-initiated alpha waves to propagate retrograde along the dorsal visual pathway, from anterior (higher-order) to posterior (lower-order) cortex ([3, 36]), in addition to a visually-evoked forward-propagating wave ([36, 37]).

Using individualized brain models, we tested dynamical mechanisms by which alpha and beta waves could be generated at rest. Several mechanisms can generate wave-like dynamics, including: limit cycles (stable, periodic behavior that the system will recover after small perturbations), quasiperiodic behavior (mixed oscillations at incommensurable frequencies), aperiodic waves (e.g. spectrally-concentrated chaos), spiral-points (transient damped-oscillations when the system is perturbed), and noise-driven "switching" between different equilibrium states. As an initial investigation we considered the attractor structure of the models, dividing models into those with equilibrium-style attractors and those with non-equilibrium attractors. The former class defines models which, in the absence of external perturbation, generate complex transient behavior, but will eventually settle into a steady-state. The latter class of dynamics, converge onto low-dimensional patterns of persistent activity, such as oscillations, even without perturbation. Conceptually, we appeal to the complex, but finite, motion of a satellite slowly falling out of orbit for the former vs. entering into a stable, predictable orbit for the latter. The dynamics embedded near an attractor are thus the system's "default" state, whereas transient patterns of activity require

some initial perturbation into that regime.

We found that these categories were consistent within-subject with 78% of subjects having the same category for separate models fit to each data half (Odds Ratio=14.4, independence $\chi^2(1) = 27.8$, $p < 1.4E-7$). Of the 87 subjects, 40 (46%) belonged to the equilibrium-group for both scans, 28 (32%) belonged to the non-equilibrium group, and 19 (22%) had one scan in each category. Surprisingly, these categories proved powerful markers of individual differences in empirical spectral power. We observed strong increases in alpha and theta power, but decreases in beta and low-gamma power for subjects with non-equilibrium dynamics (theta: $t(66) = 3.54$, alpha: $t = 6.1$, low-beta: $t = -8.57$, high-beta: $t = -6.80$, low-gamma: $t = -6.48$, p's<.0008). However, there was no difference between groups for the empirical peak-alpha frequency ($t(66) = -1.83$, n.s.). These results suggest that some aspect of the low-dimensional attractors promote the generation of high-amplitude oscillations in a band-selective manner as opposed to globally speeding/slowing dynamics (which would be reflected in peak frequency).

As an additional test, we considered how the induced dimension of model dynamics affect the alpha rhythm. By definition, attractors manifest the dynamics converging onto a reduced set of behaviors. While, in the limit (infinite time, no noise), nonequilibrium attractors have larger topological dimension than equilibrium, we are instead interested in the effective dimension of the stochastic (noisy) dynamics–i.e., how expressive are the models in terms of spatial activity patterns under physiological conditions. For this test, we used two different component-based measures of dimensionality for comparison: a hard-threshold dimension based upon the number of nontrivial PCA components (D_{PCA}, threshold = $.1\lambda_{max}$) and the Participation Ratio Dimension (D_{PR}, see 5.6) which is a graded measure. We did not use topological definitions (e.g. Hausdorff dimension) as we were interested in global dynamics embedded within the high-dimensional space, as opposed to only studying the attractors. Results indicated lower induced-dimension for non-equilibrium subjects ($D_{PR} = 13.2 \pm 5.3$, $D_{PCA} = 16.3 \pm 7.9$) compared with equilibrium subjects ($D_{PR} = 18.9 \pm 7.4$, $D_{PCA} = 10.4 \pm 5.0$) as assessed with two-sample t-tests corrected for unequal variance (D_{PR} : $t(65.9) = -3.7$, $p < .0005$, D_{PCA} : $t(65.5) = -3.8$, $p < .0004$; Fig. 6C). These findings agree with previous, empirical descriptions of lower-dimensional dynamics in the alpha band than beta ([38]). Thus, the global dynamics of non-equilibrium subjects are shaped by lower-dimensional structures.

To clarify the association of low-dimensional/ nonequilibrium dynamics with spectral power, we investigated how the models reacted to perturbations. This analysis is important for interrogating whether the alpha rhythm is, itself, an attractor (e.g., limit-cycle) or rather reflects transient behavior built upon the nearby dynamics. At the systems-level, the former case corresponds to a default-behavior which cannot be suppressed without external input (the historical alpha interpretation) whereas the latter represents a regime that leverages the intrinsic dynamics, but requires some perturbation to initialize. To differentiate these cases, we examined the model response under perturbations by simulating the models with various levels of intrinsic noise. We found that most non-equilibrium attractors had greatest power above 20Hz. When a spectral component in the alpha-range was present, the other frequencies were not harmonics of the alpha-component. These results suggest that the alpha oscillation builds off of low-dimensional intrinsic dynamics, but is not itself a self-sustaining behavior. This results indicate that, for eyes-open resting-state, active perturbations are required and sustain an alpha wave, although such perturbations need not be applied continuously.

2.7. Models Predict Increased Alpha Frequency under Perturbation

Lastly, we examined how models predict the strength of perturbation affects model dynamics. For this analysis, we focused upon changes within the alpha spectra, as opposed to between spectra. We also emphasize high-level properties, as opposed to simulating a specific task, so we modeled environmental perturbations as a random process independently delivered to each brain area. The perturbation magnitude is thus equivalent to the noise-level which we applied as a scaling-factor to simulated physiological/process noise (with baseline covariance Q as estimated by individual models). As before, process noise denotes stochasticity which drives a system, as opposed to artifact which only appears in measurements. We scaled noise-perturbations from 2% to 100% the variance used in simulations, with a resolution of 1%. Averaging over subjects, we found that the peak alpha frequency within the visual network linearly increased as a function of perturbation strength ($r(97) = .96$, $p \approx 0$; Fig. 6E). This effect agrees with empirical findings of the peak alpha frequency changing within-subject, depending upon the level of engagement (increases during task, ([25, 28]) and has been explored as a generic property in previous theoretical models ([39, 40]). We also note the observed change in peak-frequency indicates inherently nonlinear dynamics, so the existence of an equilibrium attractor should not be confused with approximately-linear dynamics. Together, these demonstrations indicate the potential of our framework to inform high-level systems neuroscience and implicate mechanisms of person-driven and context-driven variation.

3. Discussion

3.1. Application to Other Modalities

In the present work, we demonstrated our technique using magnetoencephalography (MEG) data from the Human Connectome Project (HCP; [23]). However, we expect that the approach will be similarly useful to estimating brain models from other fast-timescale modalities, particularly electroencephalography (EEG). These approaches also measure the effect of cortical dipoles at a distance and can be similarly modeled in a general state-space framework. One difference, however, is

that voltage-based modalities such as EEG, ECoG etc., are inherently referential i.e., they measure the electrical potential between points. Fortunately, this aspect is fully compatible with our algorithm and simply corresponds to a different forward model/measurement matrix (H). On a technical note, care should be taken so that the resultant data is full-rank (e.g. by factoring H; see Sec 6.8) as the leadfield is guaranteed to be rank-deficient (i.e. HH^T is not invertible) when mean-referencing is applied. The referential nature of voltage also generates a new shift invariance (nonuniqueness of C and V values), although these parametersF. However, we believe that the general enterprise of characterizing individual brain dynamics and estimating models will prove similarly applicable to EEG as MEG.

3.2. Limitations

We view our present model as having two primary limitations: 1) assumptions regarding how the M/EEG signal is generated and 2) the omission of subcortical sources. We stress that these limitations are features of our current model (two nodes per cortical parcel) rather than the approach *per se*. We have designed our publicly-available code so that it is easy to implement arbitrary models of cortical circuitry and M/EEG signal weighting (with the usual trade-offs between run-time/robustness and model complexity). We therefore examine some of our model's limitations with the caveat that these limitations are not inherent to our general estimation approach (gBPKF).

Our current results leverage a simplified "point-neuron" reduction of the excitatory pyramidal cells. This assumption was chosen as it is common throughout neural-mass modeling frameworks. However, it also belies the nuance of how different inputs are distributed across the pyramidal dendrites. For instance, inhibitory inputs may synapse closer to the soma, while inputs to the basal vs. apical dendrites have opposite contributions to the equivalent dipole ([14, 41]). Such considerations are important when parsing the biophysical origins of M/EEG signal and might be further included integrated into the model by subdividing the pyramidal cell into apical and somatobasal compartments.

A second limitation arises from the neglect of subcortical influences. Indeed, interactions with the brainstem, thalamus, and hippocampus are known to generate many slow EEG components. It is for this reason that our frequency-domain validations focused upon the faster alpha and beta bands, although the cortical vs. subcortical origins of either phenomena are not yet settled (although see [3]). In any case, it is clear that subcortical influences certainly guide cortical dynamics and that their inclusion could improve model performance. Several Dynamic Causal Modeling approaches, for instance, have included subcortical regions in modeling M/EEG ([42, 43]). This approach usually requires strong priors as the optimization may otherwise become ill-posed. By contrast, we have first prioritized validating our framework within an empirically-verifiable model space, as opposed to including as many features as possible. It is therefore possible that some of the would-be explanatory variance associated with cortico-thalamo-cortical pathways is mismodeled in our approach as direct cortico-cortical connections in the absence of a subcortical component. However, we have designed our modeling approach such that subcortical regions can be easily added with or without various priors as another latent-variable, treated analogously to the interneurons ($H_{subcort} = 0$). We encourage (verifiable) expansion in this direction.

4. Conclusion

We have presented a novel framework for estimating precision brain models from single-subject M/EEG data. Our gBPKF algorithm is shown highly capable of solving the dual-estimation problems inherent in direct brain modeling and reliably estimates latent brain model parameters directly from M/EEG timeseries. Importantly, we stress that our approach directly applies nonlinear system-identification to macroscale cortical activity, seeking to solve for the system's vector-field (i.e., the moment-to-moment variation) as opposed to replicating a specific phenomenon. Nonetheless, our approach reproduces the spatial distributions and individual differences in band-limited power across the main M/EEG bands with high fidelity. The models also link individual differences in alpha power to the geometry of underlying dynamics as shaped by excitatory-inhibitory balance. These applications, while not definitive in themselves, demonstrate the inferential power of individualized (precision) brain-modeling. We hope that these innovations were enable new insight into individual variation and, eventually, facilitate patient-specific interventions.

5. Data and Statistical Methods

In this section we briefly describe the resting-state magnetoencephalography (MEG) data used for model training and validation. An introduction to the Kalman Filter and details of the gBPKF algorithm are presented in Sec. 6. Technical descriptions of ground-truth model generation are also provided in Sec. 6

5.1. HCP MEG Data Processing

Models were fit to magnetoencephalography (MEG) data provided by the Human Connectome Project (HCP; [23]). We used the minimally-preprocessed HCP pipeline for MEG which centers on using Independent Component Analysis (ICA) to identify and remove artifact. Thus we used the HCP MEG data as-provided up to the ICA-removal step. Whereas the HCP pipeline proceeds using only the "good" independent components (IC's), we instead projected-out the bad ICs from the sensor-level timeseries. While both approaches remove the "bad" IC's as identified in the HCP MEG release, our approach still retains the original sensor space and the associated low-variance IC's which the HCP considered neglible. The rationale for this deviation is so that the those dimensions of measurement are retained as the absence of (significant) signal

along those dimensions is itself informative.

We then projected data according to the left singular-vectors of the leadfield matrix (derived from the pre-calculated Boundary Element Method headmodels). This step corresponds to reducing dimensionality based upon what spatial patterns MEG *can* measure as opposed to those that were observed (in practice these approaches overlap). Our criterion was to only retain leadfield dimensions (singular vectors whose singular values were $\geq$ 1% of the maximum singular value. These reductions were done on a per-scan basis which meant that, due to the removal of bad channels, projections (hence measurement models) often differed between scans of the same subject.

5.2. Frequency-Domain Filtering

Data was filtered between the delta and high-beta bands (1.3-30Hz). The 30Hz upper limit defines the high-beta band (typically considered 20-30Hz) while the 1.3Hz lower limit was adopted from the HCP ICA processing pipeline and does not include the full delta-band (i.e., is above the slowest delta waves). This range also includes the full alpha and theta-bands. As part of the HCP MEG pipeline, high-artifact data segments are automatically removed, resulting in variable length segments of good-quality data. To calculate power spectral density (PSD) at the subject-level we first discarded segments lasting less than 20s. The PSD was then calculated for each timeseries and discretized with resolution 0.25 Hz. The resultant PSDs were averaged across segments, weighted according to segment length.

For band-limited power, we used the pre-calculated HCP estimates in the ICA-MNE pipeline. This pipeline solves for a full dipole-vector timeseries at each of the 8004 vertices (3 dipole coordinates per vertex). Band-limited power is estimated by filtering the timeseries for each dipole coordinate and then calculating the squared magnitude of the filtered dipole vectors. HCP defines 8 specific bands: delta, theta, alpha, low-beta, high-beta, and low/mid/high gamma. In the presented results, we normalized the average band-limited power at each dipole by the whole-spectrum (unfiltered) power and then averaged within parcel to get parcel-level power.

5.3. ICA Artifact Detection

We used the HCP MEG-ICA pipeline to identify artifactual Independent Components (ICs). The HCP MEG2 and later releases sorts large ICs as having a brain or non-brain origin. However, whereas the original ICA pipeline retained only "brain" IC's (removing both artifactual and low-variance IC's) we simply removed the non-brain IC's by orthogonal projection, thereby retaining the original channel dimensions. This step retains information that the low-variance IC's are measurable and small, whereas removing them would imply that those dimensions are unmeasurable (potentially leading to model overfit). Denoting the ICA mixing matrix for the bad ICs as M_{bad} and the post-ICA corrected measurements y^{ICA}:

$$y^{ICA} = (I - M_{bad}M_{bad}^{+})y \tag{3}$$

5.4. Source Localization

We used the standard HCP anatomical pipeline to compute forward head models. Comparisons with empirical source-level power all used the HCP pre-calculated power-distributions which allow full 3d dipole configurations ([23]). However, model-training requires a fixed mapping between activity patterns and measurements, and thus a single, constant, dipole orientation per vertex (up to sign reversal). We reduced the 3d forward model to a single dipole direction per vertex by assuming that dipoles are oriented normal to the cortical surface (as calculated in FieldTrip using the vertex cross-product method). In order to transform MEG magnetometer/gradiometer measurements onto cortical dipoles we used Minimum Norm Estimation (MNE; [44]). Like other linear source-localizers, The MNE inverse matrix $\mathcal{M}$ maps sensor-level data onto the brain:

$$\hat{x}_t^{MNE} = \mathcal{M}y_t^{Chan}. \tag{4}$$

Conceptually, MNE minimizes the expected mean-square-error (like the Kalman Filter) and therefore takes the regression form:

$$\mathcal{M} = cov[x]L^T(Lcov[x]L^T + R^{MNE})^{-1} \tag{5}$$

Following the HCP ICA-MNE pipeline, we use the simplified assumption that $cov[x]$ and R^{MNE} are each independent and identically-distributed (iid) hence, using the noise-signal-ratio (NSR) coefficient λ:

$$\mathcal{M} = L^T(LL^T + \lambda I)^{-1} \tag{6}$$

We calculated λ analogous to the HCP minimal pipeline, but with a single λ per run, applied in channel-space, whereas the HCP rescaled λ's for each IC. This difference is because we retained the original channel space, instead of reducing dimensionality to the IC space.

$$\lambda = Tr[LL^T]\frac{\epsilon}{n_{Chan}} \tag{7}$$

with noise-value factor $\epsilon = 8$. The inverse solution $\mathcal{M}$ was calculated separately for each resting-state run and potentially differed (e.g. do to noisy channel removal). We further rescaled each row of $\mathcal{M}$ to produce unit variance in $\hat{x}^{MNE}$ which was done separately for each run.

5.5. Statistical Analysis

Our analyses generally fall into three categories: 1) validating/benchmarking the approach with a known (simulated) ground-truth, 2) assessing reliability with real-world data, and 3) exploring dynamical predictions made by the models. For these analyses we used three similarity measures depending upon the variable's dimensionality. We used simple correlation (collapsing over non-masked connections) to gauge the overall similarity of two matrices/vectors. To assess the reliability of individual differences, we used Intra-Class Correlation ([45]) for scalar-valued parameters (e.g., time constants) and Image Intra-Class Correlation (I2C2; [22]) for multivariate parameters (e.g. connectivity).

5.6. Quantifying Dimensionality

We quantified dimensionality of stochastic dynamics in two ways, both based off of the covariance eigenspectrum, with convergent results. First, we used the PCA-threshold method with a hard-boundary defined by 10% of the maximal component weight (eigenvalue). For this method, we quantified dimensionality as the number of components passing this threshold. For comparison, we used the participation ratio dimension (D_{PR}, [46, 47]) which is calculated from the covariance eigenvalues (λ):

$$\frac{(\sum_i \lambda_i)^2}{\sum_i \lambda_i^2}. \tag{8}$$

We used D_{PR} for comparison as it provides a soft dimensionality metric in terms of the variance spread as opposed to being premised upon a single, low-dimensional surface. It is also sensitive to dynamics which are nontrivial in the stochastic case, but eventually converge in the absence of noise. We present D_{PR} calculated using z-scored simulated data (i.e., using correlation instead of covariance) to control for the different scaling of excitatory and inhibitory neurons. However, statistical inferences led to the same conclusion with/without normalization.

6. Supplementary Methods

6.1. Mesoscale Individualized NeuroDynamic Modeling at Fast Timescales

We first describe the generative model to be estimated by each subject's data. Our model formulation is similar in motivation to conventional neural mass models (e.g., [48]). Each brain region contains two neural populations: an excitatory and an inhibitory population. A sigmoidal nonlinearity (ψ) converts population-average activation (an abstraction of depolarization) into a normalized output (analogous to firing rate). The shape of the nonlinear function is parameterized by a set of unknown variables (α). Macroscale electromagnetic fields (MEG, EEG, LFP, etc.) generated by the brain derive primarily from post-synaptic and dendritic potentials as opposed to action-potentials, hence, for our data-driven models, we use activation/depolarization (as opposed to firing-rate) as the state variable, denoted (x).

Each population receives some baseline level of drive c and it returns to baseline at a rate (1-D) with D a diagonal matrix of autoregressive coefficients. We refer to the quantity (1-D) as the "decay" rate. Both excitatory and inhibitory cells connect locally and excitatory cells also connect to distal brain areas via connection matrices (W). A major difference in our model, from previous approaches, is that we consider two types of inter-regional connections from excitatory cells: excitatory-excitatory and excitatory-inhibitory, whereas previous approaches have been constrained, by the nature of diffusion data, to a single, undirected form of connectivity (e.g., [7]). The noise terms ε correspond to unmodeled physiological processes and are assumed to be Gaussian with zero mean and covariance Q_t. Together these equations are:

$$x^E(t+1) = W^E\psi_E(x^E(t)) - \beta^E\psi_I(x^I(t)) + D^E x^E(t) + c_E + \varepsilon^E(t) \tag{9}$$

$$x^I(t+1) = W^I\psi_E(x^E(t)) - \beta^I\psi_I(x^I(t)) + D^I x^I(t) + c_I + \varepsilon^I(t) \tag{10}$$

The nonlinear function ψ is specified analogous a 2-parameter logistic function with gains (s^E,s^I) and bias terms v^E, v^I. Here, the gain s is scalar-valued (independent of parcel), whereas the biases v vary by parcel. For simplicity, we used tanh for the nonlinear function and note that such models can be directly rewritten with nonnegative activation (logistic sigmoid), if desired. Without loss of generality (see Sec. 6.7) we fix $s^I = 1, v^I = 0$, so these parameters are only solved for the excitatory population. We condense the separate population equations into the more general form with excitatory and inhibitory activity concatenated as x_t:

$$x_{t+1} = W\psi(s \circ x_t + v) + Dx_t + c + \varepsilon_t \tag{11}$$

This is referred to as the state equation as it defines how the state-variable "x" evolves in time. Coupled with the state-equation is an associated measurement-equation ("forward model") which defines how patterns of brain activation (x_t) are reflected in sensor readings (y_t). As we are interested in electromagnetic fields (which add), this transformation constitutes a linear mixing defined by the matrix H and sensor-level noise η (which evolves as a Gaussian process with zero-mean and covariance R_t).

$$y_t = H_t x_t + \eta_t \tag{12}$$

Previous research strongly indicates that brain potentials measured from the scalp (MEG and EEG) are primarily generated by cortical pyramidal cells, whose asymmetric geometry supports the formation of dipoles ([14]). By contrast, the symmetric geometry of inhibitory neurons (e.g., stellate cells) leads to the microscopic (subcellular) dipoles largely canceling when measured from a distance. Thus, for present purposes, we model the signal mixing matrix H_t as consisting of a forward matrix $\tilde{H}_t$ for excitatory populations (described in detail later) and zero for inhibitory populations. Similarly, we assume that dipoles are oriented normal to the cortical surface reflecting the underlying orientation of pyramidal cells. However, our methodology is relevant to any specification of measurement model and can therefore be adapted to alternative models of EEG signal generation (i.e., by defining a different $\tilde{H}_t$ matrix). The measurement matrix H is allowed to be time-varying (e.g., for dropping a channel during periods of artifact).

$$H_t = [\tilde{H}_t \;\; 0_{k\times n}] \tag{13}$$

We assume that the following pieces of information are known (or well-approximated):

1. A forward measurement model (H_t)
2. The sensor-noise covariance (R_t)
3. Some reasonable restrictions on the connectivity graph

The stringency of the last requirement depends upon the dimensionality of the measurements (e.g., channel count) relative the number of brain areas, particularly those with low-SNR. In our case, we directly inferred all these properties from data by using MRI data to calculate forward models (boundary element method) and empty-room recordings to estimate R. We used the

group-level distribution of fMRI MINDy models ([9]) to generate a binary "connectivity mask" of plausible connections. We used the same connectivity mask for excitatory-excitatory and excitatory-inhibitory connectomes. The remaining challenge consists of solving for the latent brain activity (x_t) and brain model parameters (Q, W, D, c, s, v) given measurements y_t. This endeavor is nontrivial as it involves a high-dimensional nonlinear optimization problem in which the system-states (brain activity for each population) are not directly accessible.

6.2. Kalman Filtering

The Kalman Filter is a recursive Bayesian algorithm for estimating *unknown* states of a *known* dynamical system. Given model parameters and measurements, the Kalman Filter and its nonlinear extensions, seek to minimize the expected difference (sum-of-squares) between true and estimated system states. The Kalman Filter differs from static approaches, however, in that its estimates also incorporate previous measurements. At each time-step the Kalman Filter passes the estimated distributions through a noisy dynamical-systems model. Using the measurement model, the Kalman Filter updates this distribution based upon the new data. By tracking the distribution over time, the Kalman Filter detects latent state variables (e.g., inhibitory neuron activity) through their predicted influence on measured variables.

6.2.1. The Prediction Step

The Kalman Filter consists of alternating prediction and correction steps. During the *Prediction step*, the dynamical systems model is used to propagate the current state distribution ($x_t \sim \mathcal{N}(\hat{x}_t, P_t)$) into an a-priori distribution for the next time step

$$\hat{x}_{t+1|t} = E[x|\hat{x}_t, P_t] \tag{14}$$

$$P_{t+1|t} = Cov[x_{t+1} - \hat{x}_{t+1|t}|\hat{x}_t, P_t] \tag{15}$$

It is important to note that the covariance matrix here, (P_t), is the error-covariance of estimating the true-state x_t using $\hat{x}_t$ and not the covariance of a sample (i.e., P_t tracks uncertainty in estimating the vector x_t not the distribution of multiple states). The original Kalman Filter dealt with the linear case: $x_{t+1} = A_t x_t + \omega_t$ in which the solutions are:

$$\hat{x}_{t+1|t} = A_t x_t \quad (\textit{linear case}) \tag{16}$$

$$P_{t+1|t} = A_t P_t A_t^T + Q_t \quad (\textit{linear case}) \tag{17}$$

Different methods have been developed to extend this prediction step to the nonlinear case ([49, 50]), which we discuss later (Sec. 6.5).

6.2.2. The Correction Step

The second step of the Kalman Filter uses new measurements to correct the a-priori predictions. This stage involves computing the Kalman Innovation (a prediction error) and adjusting state estimates proportionally to this error. Given prediction $\hat{x}_{t+1|t}$ and new measurement y_{t+1}, the prediction error ("Kalman Innovation") is:

$$r_{t+1} := y_{t+1} - H_{t+1}\hat{x}_{t+1|t} \tag{18}$$

The Kalman correction occurs by multiplying this error by the Kalman-gain matrix (K_{t+1}) to form an updated state estimate:

$$\hat{x}_{t+1} = \hat{x}_{t+1|t} + K_{t+1} r_{t+1} \tag{19}$$

The gain, K_{t+1} is selected to minimize uncertainty (variance) in the corrected estimate:

$$K_{t+1} := \arg\min_M E[\|x_{t+1} - (\hat{x}_{t+1|t} + M r_{t+1})\|^2]. \tag{20}$$

This problem constitutes least-squares regression and is therefore solved by:

$$K_{t+1} = Cov[x_{t+1}, r_{t+1}] Cov[r_{t+1}]^{-1} \tag{21}$$

In our case, the model is linear in terms of measurements and measurement-noise (i.e. the M/EEG signal is a linear function of dipole strength), so the analytic solution is given by:

$$K_{t+1} = P_{t+1|t} H_{t+1}^T (H_{t+1} P_{t+1|t} H_{t+1}^T + R_{t+1})^{-1} \tag{22}$$

The posterior estimates, incorporating measurement y_{t+1} are thus:

$$\hat{x}_{t+1} = \hat{x}_{t+1|t} + K_{t+1} r_{t+1} \tag{23}$$

$$\begin{aligned} P_{t+1} &= (I - K_{t+1}H_{t+1}) P_{t+1|t} (I - K_{t+1}H_{t+1})^T + K_{t+1} R_{t+1} K_{t+1}^T \\ &= (I - K_{t+1}H_{t+1}) P_{t+1|t} \end{aligned} \tag{24}$$

Following this correction, the Kalman Filter proceeds to the next time-step. Predictions, measurements and corrections are then assessed for $t + 2$. The major drawback of Kalman Filtering, however, is that it requires a known dynamical-systems model and initial distributions. In our framework, the dynamical-systems model is defined by the unknown parameters, so we express the Kalman Filter predictions as a function of these parameters which we optimize using gradient-based methods.

6.3. Optimization Objective

Our optimization framework in the current work solely seeks to minimize the error in predicting future sensor-level measurements, although associated code also enables the use of parameter-regularization and penalties based upon long-term model statistics (e.g. matching the observed covariance). For the present purpose, however, the model error corresponds to prediction error in the Kalman Filtering stage and in the free-running (Kalman-free) phase. For k Kalman-Filtering steps and n free-running prediction steps, we denote the error over start-times t_0, we use $\tilde{y}$ and $\tilde{H}$ to indicate that comparison measurements used to evaluate error may be in a different space than those used to estimate states with the Kalman Filter:

$$\frac{1}{n+k}\mathbb{E}_{\{t_0\}}\Big[\sum_{p=t_0+1}^{t_0+k} \mathbb{H}_{p-t_0}(\tilde{y}_p - \tilde{H}_p \hat{x}_{p|p-1}) + \sum_{q=t_0+k+1}^{t_0+k+n} \mathbb{H}_{q-t_0}(\tilde{y}_q - \tilde{H}_q \hat{x}_{q|t_0+k})\Big]. \tag{25}$$

Thus, our cost function seeks to minimize the combined prediction error over both the Kalman-Filtering (left hand side) and free-prediction (right hand side) steps. These errors are also averaged over all of the initial time points t_0 (minibatch seed) which are randomly selected during each training iteration. Within our cost function, the term $\mathbb{H}$ denotes the smooth Huber-loss transformation ([51, 52]):

$$\mathbb{H}_k(z_t) := 2\alpha_k\Big(-1 + \sqrt{1 + \frac{\|z_t\|^2_{M_t}}{\alpha_k}}\Big) \tag{26}$$

We applied the Huber-loss on-top of a conventional quadratic loss function with cost matrix M_t:

$$\|z_t\|^2_{M_t} := z_t^T M_t z_t \tag{27}$$

The smoothed Huber-loss function ($\mathbb{H}$) allows small values of $\|z\|^2$ to pass through unaltered, whereas large values of $\|z\|^2$ are quashed and limit to $\sqrt{\alpha_k}\|z\|$. This smooth loss-function approximates the original Huber-loss which penalizes squared errors below a threshold and absolute errors above it. The Huber-loss is thus robust to outliers and improves model training in the presence of unusual events (e.g., artifact). The parameter α_t sets the soft-threshold for values to be quashed under Huber-loss. For each filtering/prediction step we set α_k equal to twice the expected median error for that step (i.e., α_1 used the (estimated) median of 1-step errors etc.) based upon previous iterations. This value was updated autoregressively after each training-batch to track the evolving error distribution (i.e., α became smaller as the model became more accurate). Denoting median by *med* and training iteration m, α was updated each training iteration according to:

$$\alpha_k(m+1) = .99\alpha_k(m) + .02med(\|z_k(m)\|^2_M) \tag{28}$$

with the median being taken over initial conditions within a training minibatch. Note that this value is specific to the iteration within the prediction sequence (i.e., the kth-step prediction), not the recording time, hence $\mathbb{H}$ is indexed by $p - t_0$ and $q - t_0$ in the cost function.

6.4. The generalized Back-Propagated Kalman Filter (gBPKF)

We first motivate our algorithm conceptually, by considering the dual-estimation of states and parameters as a min-min optimization. Namely, suppose that we had some cost-function J indicating the goodness-of-fit between model predictions ($\hat{y}$) and recorded data (y) by penalizing their difference. Because model predictions are a function of the system's current state (x) the cost-function takes the form:

$$J(y_t - \hat{y}_t) = J(y_t - H_t f(\theta, x_{t-1})). \tag{29}$$

In the traditional model-fitting case, one solves for θ using x and y. However, M/EEG do not directly measure cellular activity, so both x and θ are unknown. A naive approach would be to directly optimize over both $\hat{x}$ and θ; yet such an approach quickly explodes in the number of unknown variables. However, an alternative is to replace x_{t-1} with its best possible estimator, given the data. This estimate is a function of both the previous measurements ($y_{\{k<t\}}$) and the system's dynamics (as modeled by parameters θ). Since the true values are unknown, we denote this probabilistically:

$$\hat{x}_t = \mathcal{K}_t(\theta, y_{\{k\leq t\}}) := \arg\min_z E[\bar{J}(x_t - z_t)|y_{\{k\leq t\}}, \theta]. \tag{30}$$

Here, $\bar{J}$ denotes whatever criterion is used to define a "good" state-estimate. Substituting for $\hat{x}$ in the previous equation, gives an equation solely in terms of the unknown parameters (θ) and the measurements recorded up to time t:

$$\theta = \arg\min_{\hat{\theta}} J(y_t - H_t f(\hat{\theta}, \mathcal{K}_{t-1}(\hat{\theta}, y_{\{k<t\}}))) \tag{31}$$

In our case, $\bar{J}$ corresponds to sum-of-squares differences, hence the state estimate $\mathcal{K}$ becomes the Kalman Filter (see Sec. 6.2,[53]). However, two elements of the Kalman Filter remain unknown: the initial state mean and covariance. To estimate these aspects we simulate the current model at the start of each iteration to derive baseline state (x_t) distributions. These distributions parameterize a static linear filter (total-least squares) which forms the initial state expectation and error-covariance. Unlike source reconstruction, these initial estimates include both excitatory and inhibitory cells. This difference is because the simulations provide the steady-state covariance between all populations (excitatory and inhibitory).

Denoting the concatenated free-running simulation data X_{sim} and initial measurement y_0 the initial state estimate and covariance is given by:

$$K_0 = Cov[X_{Sim}]H_0^T(H_0 Cov[X_{Sim}]H_0^T + R_0)^{-1} \tag{32}$$

$$\hat{x}_0 = E[X_{Sim}] + K_0(y_0 - E[y]) \tag{33}$$

$$P_0 = (I - K_0 H_0)Cov[X_{Sim}] \tag{34}$$

Note that these equations are identical to the Kalman update step (Eqs. 19,22) and, under trivializing assumptions, reduce to Minimum Norm Estimation (MNE; [44]) as is used in conventional source-localization for M/EEG. The initial state distribution parameters (expectation and covariance) are fed into a nonlinear Kalman Filtering algorithm. The Kalman Filter integrates a sequence of "k" measurement timesteps to refined estimates of latent neural activity. We then use the last state-estimate provided by the Kalman Filter to forecast the next "m" measurements. During the Kalman-Filtering interval, error corresponds to the difference between a-priori predicted measurements and true measurements at each time-step. Likewise, error during the final stage corresponds to the difference between true and forecasted measurements. We have previously derived and simplified the full analytical gradients of this process ([19]) which are fed into a stochastic gradient/Hessian optimization algorithm to update parameter estimates. In the present paper, we used Nesterov-Accelerated Adaptive Moment Estimation (NADAM; [54]) for the gradient updates, however a wide-variety of gradient-based algorithms

are included with the publicly available code.

For brevity, we present the gBPKF equations ([19]) in their most general form with all parameters contained within the (high-dimensional) variable θ. We also focus upon regressing gradients through the Kalman-Filtering stage, as opposed to the full filtering-forecasting sequence as the forecasting stage is simply a special case of the Kalman Filter with the probabilistic (P) and update (K) terms removed. Our objective is to minimize a function $\mathcal{L}$ of prediction error ($\mathcal{L}(y - H\hat{x})$). For an initialization time t_0 and filter-length m, we denote the accumulation of error from step t up to m as:

$$\overleftarrow{\mathcal{E}}_t^{(t_0)} := \sum_{k=t_0+t}^{t_0+m} \mathcal{L}[y_k - H\hat{x}_{k|k-1}^{\{\theta,\mathcal{F}[y]\}}] \tag{35}$$

The full objective is to solve for θ which minimizes the total error over all initialization times $\mathbb{E}_{t_0}[\overleftarrow{\mathcal{E}}_1^{(t_0)}]$. For brevity's sake, we omit the notation indicating dependencies upon parameters (θ), previous measurements ($\mathcal{F}[y]$) and start-times. Unlike a conventional dynamical systems model, the Kalman Filter evolves both the state-expectation ($\hat{x}$) and error-covariance P. To eliminate redundant equations, we use ω to denote the combined set $\{\hat{x}, P, \theta\}$ with time indices on $\hat{x}, P$ being all previous time-steps.

$$\nabla\overleftarrow{\mathcal{E}}_t = 2(\partial_\omega z_t)H^T\mathcal{L}' + \frac{\partial\overleftarrow{\mathcal{E}}_{t+1}}{\partial\hat{x}_t}(\partial_\omega\hat{x}_t) + \frac{\overleftarrow{\mathcal{E}}_{t+1}}{\partial P_t}(\partial_\omega P_t) \tag{36}$$

Evaluating the first term: $\partial_\omega z_t = -H\partial_\omega\hat{x}_{t|t-1}$

$$\partial_\omega\hat{x}_{t+1} = G_{t+1}(\partial_\omega\hat{x}_{t+1|t}) + (\partial_\omega K)z_{t+1} \tag{37}$$

$$\partial_\omega P_{t+1} = G_{t+1}(\partial_\omega P_{t+1|t}) - (\partial_\omega K)HP_{t+1|t} \tag{38}$$

Since H, Q, R are fixed, the Kalman gain is a direct function of $P_{t|t-1}$ (and its influence on S). Using that K_t minimizes $Tr[P_t]$ and applying the implicit function theorem:

$$-\partial P_{t|t-1}H^T + \partial K_t S + K_t H\partial P_{t|t-1}H^T = 0 \tag{39}$$

$$\partial_\omega K_t = G_t(\partial_\omega P_{t|t-1})H^T S^{-1} \tag{40}$$

$$\partial_\omega\hat{x}_t = G_t[\partial_\omega\hat{x}_{t|t-1} + (\partial_\omega P_{t|t-1})H^T S^{-1}z_t] \tag{41}$$

$$\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_{t|t-1}} = G_t^T\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_t} - 2H^T M_t z_t \tag{42}$$

$$\partial_\omega P_t = G_t(\partial_\omega P_{t|t-1})G_t^T \tag{43}$$

$$\mathcal{Z}_t := H^T S^{-1}z_t\left[G_t\frac{\partial\overleftarrow{\mathcal{E}}_{t+1}}{\partial\hat{x}_t}\right]^T \tag{44}$$

$$\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t|t-1}} = \frac{1}{2}(\mathcal{Z}_t + \mathcal{Z}_t^T) + G_t^T\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_t}G_t \tag{45}$$

Using that all Kalman-Filter approximations of $\mathbb{E}$ preserve linearity and cov preserve bilinearity, we have, for $\phi := \{\hat{x}_{t-1}, \theta_{t-1}\}$:

$$\partial_\phi\overleftarrow{\mathcal{E}}_t = \frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_{t|t-1}}\mathbb{E}[\partial_\phi f] + \frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t|t-1}}[cov[f, \partial_\phi f] + cov[\partial_\phi f, f]] \tag{46}$$

In particular, for the EKF ([49]), we have:

$$\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t-1}} = F_t'^T\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t|t-1}}F_t' \tag{47}$$

$$\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_{t-1}} = F'^T\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_{t|t-1}} + 2vec\left[\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t|t-1}}F'P_t\right]^T\frac{\partial vec[F'])}{\partial\hat{x}_t} \tag{48}$$

The final gradient of total error with respect to parameter is:

$$\frac{\partial\mathcal{E}}{\partial\theta} = \sum_t\frac{\partial f_t^T}{\partial\theta}\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial\hat{x}_{t|t-1}} + 2vec\left[\frac{\partial\overleftarrow{\mathcal{E}}_t}{\partial P_{t|t-1}}F'P_t\right]^T\frac{\partial vec[F']}{\partial\theta} \tag{49}$$

Gradients with respect to sample-based filters (e.g., Unscented Kalman Filter, [50]) can be found in [19].

6.5. Estimating Nonlinear Posteriors

In general, our approach (gBPKF) is compatible with a wide variety of techniques for estimating posterior means and covariances following a nonlinearity (see [19]). In the current work, however, we use a simple batch formulation of the Extended Kalman Filter (EKF; [49]) which allows the Kalman Filter to run on many data segments in parallel. We use $\hat{X}$ to denote the concatenated state estimates for all initial conditions within a minibatch, with a common error-covariance ($\hat{P}$) used for all members of the minibatch. We use the set-valued index index $\{\mathbf{q}\}$ to denote the current time step for all data segments within the minibatch. The prior distribution and posterior statistics are:

$$X \sim \mathcal{N}(\hat{X}_{\{\mathbf{q}\}}, P_{\{\mathbf{q}\}}) \tag{50}$$

$$\hat{X}_{(\{\mathbf{q}\}+1)|\{\mathbf{q}\}} = f_{\{\mathbf{q}\}}(\hat{X}_{\{\mathbf{q}\}}) \tag{51}$$

$$P_{(\{\mathbf{q}\}+1)|\{\mathbf{q}\}} = \mathbb{E}_{\{\mathbf{q}\}}\left[\frac{\partial f}{\partial x}\right]^T P_{\{\mathbf{q}\}}\mathbb{E}_{\{\mathbf{q}\}}\left[\frac{\partial f}{\partial x}\right] + \mathbb{E}_{\{\mathbf{q}\}}[Q_{\{\mathbf{q}\}}] \tag{52}$$

Hence we alter the EKF to use the average Jacobian over the minibatch. This step allows many data segments to be run in parallel by sharing a common covariance $P_{\{\mathbf{q}\}}$ at the cost of having less sensitivity to time-variation in the nonlinearity. To counteract this drawback, we chose minibatches as temporal chunks (80 initial timepoints with 5-step=10ms spacing) so as to track nonstationary statistics. In all of our applications, the process noise $Q_{\{\mathbf{q}\}}$ did not depend upon time, so the rightmost expectation simplifies to Q.

6.6. Generating the Connectivity Mask

We constrained eligible inter-area connections using a liberally-defined connectivity mask which was generated based upon previous modeling with fMRI data ([9]) as described below. While this step is not an inherent requirement of our approach, it is important for ensuring that fits are robust with M/EEG data. Specifically, this constraint can counteract modest deviations in the forward model (i.e., head position relative sensors) and promotes well-posedness when multiple brain areas have weak contributions to the M/EEG signal. We formed our connectivity-mask based upon group-level consensus from

the fMRI MINDy models ([9]) using the union of several criteria to avoid reliance upon any single measure of what constitutes a non-trivial connection. This procedure retains sensitivity to small, but consistent connections as well as connections that may not show up in every subject but are (on average) large. We rescaled each fMRI connection matrix (two runs per subject) to have a root-mean-square value of 1, excluding self-connections. For thresholding, we used an absolute magnitude threshold of 0.5 applied to each matrix. We also note that, unlike the M/EEG models presented in the current work, the fMRI MINDy models use a single connection per region-pair which can be positive or negative as opposed to both EE and EI projections. We then generated the connectivity mask through the union of three criteria:

1. Average magnitude: We admitted the top 15% of connections in terms of magnitude for the group-average.
2. Consensus: We admitted connections with an average post-threshold sign greater than 0.8 or less than -0.6 across subjects/runs.
3. Minimal Count: For each parcel we admitted the (up-to) three largest positive and negative connections in terms of input, output, and symmetrized-strength (average of input and output) with overlap between these categories.

We then symmetrized the resultant matrix so that if $W_{i,j}$ is an admissible connection, so is $W_{j,i}$. The final mask had 2,522 admissible long-distance connections out of the possible 9,900 (25.2%). We used the same connectivity mask to constrain long-distance excitatory-excitatory connections and long-distance excitatory-inhibitory connections for a total of 5,044 inter-region conncetions.

6.7. Promoting Unique Solutions

A key trap in data-driven modeling with latent-variables is that several model solutions may be observationally-equivalent, meaning that they behave identically in terms of the measured variables even if they are nonidentical in terms of the non-measured variables. In our case, this ambiguity largely corresponds to models in which the predicted excitatory activity is the same, but not the inhibitory activity which lacks a predefined unit (scale) since it is not directly measured. For many use-cases this ambiguity is irrelevant as it only affects interpretation of the absolute scale for inhibitory activity/parameters. However, for good form, we reduce the parameter-space to ensure unique solutions. To be clear, the model is not overparameterized in the usual sense–for a fixed set of latent states $\{x\}$ the optimal parameter choices are well-posed. Rather, the difficulty arrives with the possibility of transformed systems behaving identically in terms of measurements. This relationship is called observational equivalence and, in the deterministic case, can be stated as follows: the systems $x_{t+1} = f_t(x_t)$ and $\hat{x}_{t+1} = \hat{f}_t(\hat{x}_t)$ are observationally-equivalent with respect to the measurement-process $y(x) = Hx$ if for every x_0 there exists $\hat{x}_0$ s.t. $H_t x_t = H_t \hat{x}_t$ with x and $\hat{x}$ evolving according to f and $\hat{f}$, respectively. As a specific case, consider $f(x) = W\psi(S\,x + v) + Dx + c$ with S and D diagonal. If the vectors b, q satisfy $H_t diag(b) = H_t$ and $H_t q = 0, \quad \forall t$, then the transformed system $\tilde{x} = b \circ x + q$ is observationally-equivalent to x and has parameters:

$$\tilde{S} = b^{-1} \circ S, \quad \tilde{W} = diag(b)W \tag{53}$$

$$\tilde{v} = v - b^{-1} \circ q, \quad \tilde{c} = b \circ c + (I - D)q \tag{54}$$

Here and in all later cases we use $\circ$ to denote element-wise multiplication (Hadamard product). The inverse-notation b^{-1} is understood to be applied element-wise for vectors. In the above case, we note that b, q may be chosen arbitrarily for inhibitory indices since the corresponding measurement gains are zero (j inhibitory implies $H_{i,j} = 0$). Thus there are at least two arbitrary degrees of freedom per-region in the unrestricted model: parameter choices can be altered such that they uniformally shift or rescale estimates of the inhibitory population activity. To remove these arbitrary degrees of freedom, we fixed parameters of the inhibitory nonlinearity to be $S_I = 1, V_I = 0$, thereby removing the shift and scale symmetries, respectively.

However, some invariances may remain due to volume conduction and referencing. These invariances do not affect temporal dynamics in the measurement space, and instead reflect translating the system in a direction to which the M/EEG sensors are blind (i.e., a pattern of brain activity in which electric fields cancel at the sensor-level). In mean-referenced EEG, for instance, the relativistic nature of voltages mean that the result of shifting each region's activation by a constant (i.e. $\hat{x}_i(t) = x_i(t) + c, \forall i, t$) will be observationally-equivalent to the original system.

To further constrain the problem, we reduced the space of recurrent connectivity patterns by assuming that they shared a common non-negative spatial gradient (b) within-subject and that each recurrent connectivity vector ($u \in \mathbb{R}^{n_{parc}}$) can be expressed as an affine transformation of this gradient:

$$u_w = \hat{a}_w b^{rec} + \hat{d}_w \tag{55}$$

with separate scalar values of a_w, d_w for $w \in \{EE, EI, IE, II\}$. In experimentation, we found that this restriction can be further eased with separate excitatory and inhibitory spatial gradients, but do not recommend fully unrestricting the recurrent connections due to the aforementioned possibility of parameter symmetries. The associated software enables users to define arbitrary constraints of this form between parameters and use multiple-component bases (matrix-valued b^{rec}, vector-valued a_w).

6.8. Defining the MEG Measurement Model

For this initial validation with MEG, we built a measurement model in which the timeseries had already been source-localized with Minimum Norm Estimation (MNE), while the optimization objective was calculated in a rank-reduced subspace of this projection (described below). The noise-covariance matrix was adapted from empty-room recordings (R^{chan}). Separately for each scan, we rank-reduced the data according to the singular-values of the post-ICA leadfield: $\hat{L} = L(I - M_{bad}M_{bad}^T)$. We rank reduced L based upon a

singular-value threshold of 1% the maximal singular value and denote the reduced left-singular vector matrix U_L. We projected the leadfield onto this space (premultiplied by U_L^T) which, inherently removes the ICA-censored dimensions from the leadfield (reflecting the fact that those dimensions were discarded). Source-estimation was then performed using MNE on this new leadfield to produce the source-inversion matrix $\mathcal{M}_0$. We rescaled rows of the MNE inverse solution ($\mathcal{M}_0$) so that $\mathcal{M}_0 U_L^T y^{chan}$ had unit variance for each channel. Thus, the final inverse transformation was:

$$\mathcal{M} = sd[\mathcal{M}_0 U^T y^{chan}]^{-1} \circ \mathcal{M}_0 U_L^T \tag{56}$$

We correspondingly transformed variables such that the Kalman-Filter receives source-estimates as the measurement variable:

$$y^{MNE} = \mathcal{M} y^{chan} \tag{57}$$

$$R^{MNE} = \mathcal{M} R^{chan} \mathcal{M}^T \tag{58}$$

Within this space, the measurement matrix was:

$$H = \begin{bmatrix} I_{n_{parc}} & 0_{n_{parc}} \end{bmatrix} \tag{59}$$

reflecting a direct, but noisy, source-estimate for excitatory activity and no sensitivity to inhibitory activity. Thus, standard source-estimates were fed into the Kalman Filter during state-estimation. However, the model error was assessed over the more parsimonious space span($\mathcal{M}$) which removes artificial dimensions created during source-estimation. Thus, models are not held to fully obey the source estimates in dimensions of ambiguity. Writing the singular-value decomposition of $\mathcal{M}$:

$$\mathcal{M} = \begin{bmatrix} U_1 & U_0 \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} V^T \tag{60}$$

We used the cost matrix: $M = U_1 U_1^T$ (see Eq. 27).

6.9. gBPKF Training Parameters for MEG Data

The gBPKF algorithm ultimately consists of training a (biological) recurrent neural network with the Kalman-Filter as an additional, interconnected unit. Thus, the hyperparameter classes are conceptually analogous to any deep learning scenario although the actual training equations are quite different. For the HCP MEG data, we used a filtering period of 26-steps with the Kalman correction only applied every fifth step starting at one (i.e., 6 updates total) followed by a forecasting stage of 35 prediction steps. On each batch, initial distributions (mean and covariance) were estimated by simultaneously simulating 50 initial conditions with process noise for 50 time-steps. Initial values of x for these simulations were drawn from a moving record of predictions during previous forecasting stages. This step reduces the time to reach steady-state ($t \to \infty$) distributions, thereby enabling shorter simulation periods. The covariance of this simulation was smoothed according to an autoregressive average with coefficient 0.05, i.e. $cov(k) = .95cov(k-1) + .05cov[sim(k)]$. Gradients were fully propagated through each simulation's influence on $cov[sim(k)]$, but not to previous training iterations (which had different parameter estimates). Minibatches consisted of 80 time segments and there were 4 minibatches per training iteration. For efficiency, minibatches began at the filtering stage, so state distributions were estimated once per training iteration instead of once per minibatch. Training used a fixed budget of 150,000 iterations (determined based upon examination of convergence rate). Gradients were clipped ([55]) with a moving threshold of 3 times the average norm over the past 200 iterations. Parameter updates were performed using the NADAM algorithm ([54]) with hyperparameters $\beta_1 = .98, \beta_2 = .99, \nu = .0001$ and rate .0001.

6.10. Ground-Truth Simulations

In previous work, we validated and benchmarked the gBPKF algorithm in estimating parameters for unstructured, randomly generated recurrent neural networks ([19]). However, these previous analyses did not consider whether our approach would perform equally well with networks obeying an excitatory-inhibitory structure. We therefore performed a new set of ground-truth simulations which directly mirrored our model setup. Each ground-truth model consisted of interconnected "regions" which each contained an excitatory and an inhibitory population. Only excitatory populations made long-distance connections which targeted both excitatory and inhibitory populations. Measurements were simulated by multiplying excitatory activity with a randomly generated "leadfield" and adding temporally-independent Gaussian noise (with covariance R).

To parameterize the ground-truth models, we first randomly generated a symmetric connectivity mask W_{mask} with 25% density on the off-diagonals and zero on the diagonals. The same connection mask is used for long-distance EE and EI connections. The EE and EI connection strengths are also identically generated from the similarly constructed distributions, hence we omit the EE vs. EI distinction and refer to a single connectivity matrix for now. Other than symmetry ($W_{mask} = W_{mask}^T$), the entries of W_{mask} are uncorrelated. As with data applications, the values of W_{mask} simply indicate if a connection is plausible while the actual value could be effectively zero. Connection strengths were not symmetric and were generated as the sum of a sparse matrix and a low-rank matrix. The sparse matrix W_s was distributed:

$$W_s \sim \mathcal{N}(0,1)^3 \tag{61}$$

The low rank component (W_L) was the product of two rectangular matrices ($W_L = W_1 W_2^T$) generated according to:

$$W_1, W_2 \sim \mathcal{N}(0,1)^3 + \mathcal{N}(0, 1/25) \tag{62}$$

These terms are combined according to

$$\hat{W} = W_{mask} \circ \left(\frac{8 W_1 W_2^T}{c_0 n \;\; max(W_1 W_2^T)} + \frac{16}{n} W_S \right) \tag{63}$$

Note that the individual connection strengths are inversely proportional to the number of nodes (n) so that the total input

strength is similar across simulation sizes (akin to subdividing the brain into smaller pieces). This process was used to generate long-distance EE connections ($\hat{W}_E$) and EI connections ($\hat{W}_I$). For the EE connections $c_0 = 2$, whereas for the EI connections $c_0 = 1.5$, hence EE projections were generally stronger than EI. Recurrent connections were generated from a narrower distribution designed to sustain nontrivial dynamics (oscillations etc.). The base values for each local connections were: EE: .5, EI: 1.25, IE: -1.25, II: -.25. The spatial gradation in local connectivity (b) was generated as $b \sim .85 + .3\mathcal{N}(0, I_{n\times n})^2$. The final local connectivity strength was generated by multiplying the base value for each connection type by b. As simulations were randomly-parameterized, a small number produced pathological steady-state behavior in which model inversion was theoretically impossible. Specifically, in these cases the dynamics generated attractors so far outside the dynamic range of ψ that the network was stuck in a saturated state. These dynamics are unambiguous and easy to detect. We removed simulations in which the median moving-variance of ψ was less than .01 for at least half of the populations/nodes (normal values are around .9 for all nodes).

References

[1] F. L. Da Silva, T. Van Lierop, C. Schrijer, W. S. Van Leeuwen, Organization of thalamic and cortical alpha rhythms: spectra and coherences, Electroencephalography and clinical neurophysiology 35 (6) (1973) 627–639.

[2] M. Schreckenberger, C. Lange-Asschenfeld, M. Lochmann, K. Mann, T. Siessmeier, H.-G. Buchholz, P. Bartenstein, G. Gründer, The thalamus as the generator and modulator of eeg alpha rhythm: a combined pet/eeg study with lorazepam challenge in humans, Neuroimage 22 (2) (2004) 637–644.

[3] M. Halgren, I. Ulbert, H. Bastuji, D. Fabó, L. Erőss, M. Rey, O. Devinsky, W. K. Doyle, R. Mak-McCully, E. Halgren, L. Wittner, P. Chauvel, G. Heit, E. Eskandar, A. Mandell, S. S. Cash, The generation and propagation of the human alpha rhythm, Proceedings of the National Academy of Sciences 116 (47) (2019) 23772–23782. doi:10.1073/pnas.1913092116.

[4] A. L. Hodgkin, A. F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, The Journal of physiology 117 (4) (1952) 500.

[5] K. Friston, L. Harrison, W. Penny, Dynamic causal modelling, NeuroImage 19 (4) (2003) 1273 – 1302. doi:10.1016/S1053-8119(03)00202-7.

[6] C. J. Honey, O. Sporns, L. Cammoun, X. Gigandet, J. P. Thiran, R. Meuli, P. Hagmann, Predicting human resting-state functional connectivity from structural connectivity, Proceedings of the National Academy of Sciences 106 (6) (2009) 2035–2040. doi:10.1073/pnas.0811168106.

[7] P. Wang, R. Kong, X. Kong, R. Liégeois, C. Orban, G. Deco, M. P. van den Heuvel, B. Thomas Yeo, Inversion of a large-scale circuit model reveals a cortical hierarchy in the dynamic resting human brain, Science Advances 5 (1) (2019). doi:10.1126/sciadv.aat7854.

[8] M. Demirtaş, J. B. Burt, M. Helmer, J. L. Ji, B. D. Adkinson, M. F. Glasser, D. C. Van Essen, S. N. Sotiropoulos, A. Anticevic, J. D. Murray, Hierarchical heterogeneity across human cortex shapes large-scale neural dynamics, Neuron 101 (6) (2019) 1181–1194. doi:10.1016/j.neuron.2019.01.017.

[9] M. F. Singh, T. S. Braver, M. W. Cole, S. Ching, Estimation and validation of individualized dynamic brain models with resting state fmri, NeuroImage 221 (2020) 117046. doi:https://doi.org/10.1016/j.neuroimage.2020.117046.

[10] M. Singh, A. Wang, T. Braver, S. Ching, Scalable surrogate deconvolution for identification of partially-observable systems and brain modeling, Journal of Neural Engineering (2020). URL http://iopscience.iop.org/10.1088/1741-2552/aba07d

[11] N. Brunel, X.-J. Wang, What determines the frequency of fast network oscillations with irregular neural discharges? i. synaptic dynamics and excitation-inhibition balance, Journal of Neurophysiology 90 (1) (2003) 415–430. doi:10.1152/jn.01095.2002.

[12] B. V. Atallah, M. Scanziani, Instantaneous modulation of gamma oscillation frequency by balancing excitation with inhibition, Neuron 62 (4) (2009) 566–577. doi:10.1016/j.neuron.2009.04.027.

[13] T. van Kerkoerle, M. W. Self, B. Dagnino, M.-A. Gariel-Mathis, J. Poort, C. van der Togt, P. R. Roelfsema, Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex, Proceedings of the National Academy of Sciences 111 (40) (2014) 14332–14341. doi:10.1073/pnas.1402773111.

[14] G. Buzsáki, C. A. Anastassiou, C. Koch, The origin of extracellular fields and currents—eeg, ecog, lfp and spikes, Nature reviews neuroscience 13 (6) (2012) 407–420.

[15] M. F. Singh, A. Wang, M. Cole, S. Ching, T. S. Braver, Enhancing task fmri preprocessing via individualized model-based filtering of intrinsic activity dynamics, NeuroImage (2021) 118836.

[16] T. Haarnoja, A. Ajay, S. Levine, P. Abbeel, Backprop kf: Learning discriminative deterministic state estimators, Advances in neural information processing systems 29 (2016).

[17] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the royal statistical society: series B (methodological) 39 (1) (1977) 1–22.

[18] J. Kokkala, A. Solin, S. Särkkä, Expectation maximization based parameter estimation by sigma-point and particle smoothing, in: 17th International Conference on Information Fusion (FUSION), IEEE, 2014, pp. 1–8.

[19] M. F. Singh, M. Wang, M. W. Cole, S. Ching, Efficient identification for modeling high-dimensional brain dynamics, in: 2022 American Control Conference (ACC), IEEE, 2022, pp. 1353–1358.

[20] E. S. Finn, X. Shen, D. Scheinost, M. D. Rosenberg, J. Huang, M. M. Chun, X. Papademetris, R. T. Constable, Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity, Nature neuroscience 18 (11) (2015) 1664–1671.

[21] B. T. Thomas Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, B. Fischl, H. Liu, R. L. Buckner, The organization of the human cerebral cortex estimated by intrinsic functional connectivity, Journal of Neurophysiology 106 (3) (2011) 1125–1165. doi:10.1152/jn.00338.2011.

[22] H. Shou, A. Eloyan, S. Lee, V. Zipunnikov, A. Crainiceanu, M. Nebel, B. Caffo, M. Lindquist, C. M. Crainiceanu, Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (i2c2), Cognitive, Affective, & Behavioral Neuroscience 13 (2013) 714–724.

[23] L. J. Larson-Prior, R. Oostenveld, S. Della Penna, G. Michalareas, F. Prior, A. Babajani-Feremi, J.-M. Schoffelen, L. Marzetti, F. de Pasquale, F. Di Pompeo, et al., Adding dynamics to the human connectome project with meg, Neuroimage 80 (2013) 190–201.

[24] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, B. T. Yeo, Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri, Cerebral Cortex (2017) 1–20doi:10.1093/cercor/bhx179.

[25] S. Haegens, H. Cousijn, G. Wallis, P. J. Harrison, A. C. Nobre, Inter- and intra-individual variability in alpha peak frequency, NeuroImage 92 (2014) 46–55. doi:https://doi.org/10.1016/j.neuroimage.2014.01.049.

[26] T. H. Grandy, M. Werkle-Bergner, C. Chicherio, F. Schmiedek, M. Lövdén, U. Lindenberger, Peak individual alpha frequency qualifies as a stable neurophysiological trait marker in healthy younger and older adults, Psychophysiology 50 (6) (2013) 570–582. doi:https://doi.org/10.1111/psyp.12043.

[27] W. Klimesch, Evoked alpha and early access to the knowledge system: The p1 inhibition timing hypothesis, Brain Research 1408 (2011) 52–71. doi:https://doi.org/10.1016/j.brainres.2011.06.003.

[28] R. Cecere, G. Rees, V. Romei, Individual differences in alpha frequency drive crossmodal illusory perception, Current Biology 25 (2) (2015) 231–235. doi:https://doi.org/10.1016/j.cub.2014.11.034.

[29] R. Gao, E. J. Peterson, B. Voytek, Inferring synaptic excitation/inhibition balance from field potentials, Neuroimage 158 (2017) 70–78.

[30] F. Lombardi, H. J. Herrmann, L. de Arcangelis, Balance of excitation and inhibition determines 1/f power spectrum in neuronal networks, Chaos:

An Interdisciplinary Journal of Nonlinear Science 27 (4) (2017).
[31] G. Pfurtscheller, A. Stancak Jr, C. Neuper, Event-related synchronization (ers) in the alpha band—an electrophysiological correlate of cortical idling: a review, International journal of psychophysiology 24 (1-2) (1996) 39–46.
[32] N. R. Cooper, R. J. Croft, S. J. Dominey, A. P. Burgess, J. H. Gruzelier, Paradox lost? exploring the role of alpha oscillations during externally vs. internally directed attention and the implications for idling and inhibition hypotheses, International journal of psychophysiology 47 (1) (2003) 65–74.
[33] J. J. Foxe, G. V. Simpson, S. P. Ahlfors, Parieto-occipital 1 0hz activity reflects anticipatory state of visual attention mechanisms, Neuroreport 9 (17) (1998) 3929–3933.
[34] W. Klimesch, P. Sauseng, S. Hanslmayr, Eeg alpha oscillations: the inhibition–timing hypothesis, Brain research reviews 53 (1) (2007) 63–88.
[35] W. Klimesch, Evoked alpha and early access to the knowledge system: the p1 inhibition timing hypothesis, Brain research 1408 (2011) 52–71.
[36] A. Alamia, L. Terral, M. R. D'ambra, R. VanRullen, Distinct roles of forward and backward alpha-band waves in spatial visual attention, eLife 12 (2023) e85035. doi:10.7554/eLife.85035.
[37] D. Lozano-Soldevilla, R. VanRullen, The hidden spatial dimension of alpha: 10-hz perceptual echoes propagate as periodic traveling waves in the human brain, Cell Reports 26 (2) (2019) 374–380.e4. doi:https://doi.org/10.1016/j.celrep.2018.12.058.
[38] L. Aftanas, S. Golocheikine, Non-linear dynamic complexity of the human eeg during meditation, Neuroscience Letters 330 (2) (2002) 143–146. doi:https://doi.org/10.1016/S0304-3940(02)00745-0.
[39] A. Hutt, A. Mierau, J. Lefebvre, Dynamic control of synchronous activity in networks of spiking neurons, PloS one 11 (9) (2016) e0161488.
[40] A. Mierau, W. Klimesch, J. Lefebvre, State-dependent alpha peak frequency shifts: Experimental evidence, potential mechanisms and functional implications, Neuroscience 360 (2017) 146–154.
[41] Biophysically detailed forward modeling of the neural origin of eeg and meg signals, NeuroImage 225 (2021) 117467. doi:https://doi.org/10.1016/j.neuroimage.2020.117467.
[42] S. J. Kiebel, O. David, K. J. Friston, Dynamic causal modelling of evoked responses in eeg/meg with lead field parameterization, NeuroImage 30 (4) (2006) 1273–1284.
[43] S. J. Kiebel, M. I. Garrido, R. J. Moran, K. J. Friston, Dynamic causal modelling for eeg and meg, Cognitive neurodynamics 2 (2) (2008) 121–136.
[44] M. S. Hämäläinen, R. J. Ilmoniemi, Interpreting magnetic fields of the brain: minimum norm estimates, Medical & biological engineering & computing 32 (1994) 35–42.
[45] P. E. Shrout, J. L. Fleiss, Intraclass correlations: uses in assessing rater reliability., Psychological bulletin 86 (2) (1979) 420.
[46] B. Kramer, A. MacKinnon, Localization: theory and experiment, Reports on Progress in Physics 56 (12) (1993) 1469.
[47] E. Altan, S. A. Solla, L. E. Miller, E. J. Perreault, Estimating the dimensionality of the manifold underlying multi-electrode neural recordings, PLoS computational biology 17 (11) (2021) e1008591.
[48] H. R. Wilson, J. D. Cowan, Excitatory and inhibitory interactions in localized populations of model neurons, Biophysical Journal 12 (1972) 1–24.
[49] S. J. Julier, J. K. Uhlmann, New extension of the kalman filter to nonlinear systems, in: Signal processing, sensor fusion, and target recognition VI, Vol. 3068, 1997, pp. 182–193.
[50] S. Julier, J. Uhlmann, H. F. Durrant-Whyte, A new method for the nonlinear transformation of means and covariances in filters and estimators, IEEE Transactions on automatic control 45 (3) (2000) 477–482.
[51] P. J. Huber, Robust estimation of a location parameter, in: Breakthroughs in statistics: Methodology and distribution, 1992, pp. 492–518.
[52] P. Charbonnier, L. Blanc-Féraud, G. Aubert, M. Barlaud, Deterministic edge-preserving regularization in computed imaging, IEEE Transactions on image processing 6 (2) (1997) 298–311.
[53] R. E. Kalman, A new approach to linear filtering and prediction problems, Transactions of the ASME–Journal of Basic Engineering 82 (1960) 34–45.
[54] T. Dozat, Incorporating nesterov momentum into adam, Proceedings of 4th International Conference on Learning Representations, Workshop Track, 2016 (2016).
[55] R. Pascanu, T. Mikolov, Y. Bengio, On the difficulty of training recurrent neural networks (2013). arXiv:1211.5063.