Hybrid Modeling of Whole-Brain Dynamics balances Data Fidelity and Interpretability

Daniel López-Montero[1], Lorenzo Liverani[1], Enrique Zuazua[1,4,5], Xenia Kobeleva[2,3]

[1]Chair for Dynamics, Control, Machine Learning, and Numerics (Alexander von Humboldt Professorship), Department of Mathematics, Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058, Erlangen, Germany.
[2]Department of Neurostimulation, Medical Faculty, Ruhr University Bochum, Universitätsstraße 150, 44801, Bochum, Germany.
[3]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany.
[4]Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049, Madrid, Spain.
[5]Chair of Computational Mathematics, Fundación Deusto, Av. de las Universidades 24, 48007, Bilbao, Basque Country, Spain.

Contributing authors: mail@xenia-kobeleva.com;

Abstract

Whole-brain models (WBMs) are crucial for understanding large-scale neural dynamics. Two main approaches exist: (i) physics-based WBMs built on coupled oscillators such as the Hopf oscillator, which rely on rigid parametrizations with few parameters and computationally expensive model fitting; and (ii) data-driven models such as Neural SDEs, which offer scalable learning but are sample-inefficient and lack interpretable structure. We propose a hybrid modeling framework combining both models and using gradient descent-based fitting to simultaneously optimize multiple metrics representing neural activity, such as network structure and its dynamics. We demonstrate that hybrid models with a data-driven connectome component achieve higher data fidelity than both physics-based WBMs and data-driven models across static and dynamic network descriptors, while preserving interpretable oscillator-based structure and replacing the need for exhaustive parameter searches. These results highlight the flexibility and scalability of hybrid WBMs and support their utility for personalized modeling and future digital brain twin applications.

Keywords: whole-brain modeling, Neural SDE, hybrid modeling, model fitting, gradient-based optimization

Mathematical models of whole-brain dynamics (whole-brain models, WBMs) describe brain activity and interactions between brain regions through systems of differential equations. WBMs are capable of simulating human neural activity estimated by functional MRI (fMRI) and electroencephalography (EEG). Recent research has recognized their potential to be incorporated in clinical trials as add-on digital biomarkers for more precise diagnoses [1, 2] or to assist with matching treatments to patients based on simulations [3–6]. Further model development could thus pave the way for WBMs towards digital twins of the human brain for clinical purposes.

Currently, two opposite model architectures dominate the field of modeling with traditional physics-based WBMs from computational neuroscience and newer data-driven models from machine learning. Traditionally, physics-based WBMs have been derived from neuroscientific theories of how large-scale dynamics emerge from the interaction of distributed neural populations [7–10]. These early WBMs were designed primarily to understand general brain function rather than for clinical use. Their theory-focused application has led varying degrees of data fidelity of WBMs to individual data and overreliance on rigid assumptions about underlying brain dynamics. Furthermore, model fitting is usually performed through computationally expensive grid search over few parameters [11, 12]. As a consequence of focus on theory and computational limitations regarding fitting parameters, many WBM studies have not focused on generating more expressive models.

Contrasting physics-based WBMs, novel data-driven models such as neural stochastic differential equations (Neural SDEs), incorporate multitudes of data-driven parameters, whose combinations lead allow to discover complex nonlinear dynamics directly from data [13–16]. While resulting in enhanced expressivity, their main drawback is limited interpretability with reduced options for quality control, which can reduce trust and acceptance among neuroscientists and medical professionals. Moreover, owing to their complexity, data-driven models are data-inefficient and thus might miss the full range of relevant dynamical features when applied to small-sample studies common in neuroscience.

Hybrid modeling in scientific machine learning provides a balanced compromise between physics-based and data-driven WBMs. Rather than replacing neuroscientific domain theory and interpretable priors with black-box function approximators, hybrid approaches embed learnable components within differential-equation models or use scientific structure as a constraint on learning. Known mechanisms therefore remain explicit, while unknown closures, couplings, or residual dynamics are inferred from data [17–21]. In scientific and biological modeling, this strategy has improved data efficiency, extrapolation, and interpretability [22, 23].

As outlined above, inefficient fitting procedures and inappropriate fitting objectives often restrict researchers from enhancing WBMs to incorporate more data-driven parameters, as required for hybrid model architectures. For instance, pointwise signal matching alone often fails to reproduce the spatiotemporal brain dynamics of brain activity [24, 25], a limitation also observed in domains such as speech enhancement [26–28]. While less established for data-driven models, statistical and dynamical descriptors of brain dynamics are often used as evaluation metrics to fit physics-based WBMs [8].

However, for data-driven and hybrid models these metrics would need to be integrated into a more efficient fitting procedure.

To merge strengths of both physics-based and data-driven WBMs, we propose a differentiable hybrid modeling framework with two main methodological contributions, i.e., we establish efficient and accurate model fitting and generate a novel hybrid model architecture, as shortly outlined below.

First, we establish a hybrid fitting procedure merging fitting techniques which have been used in isolation for either physics-based or data-driven WBMs. Specifically, we fit continuous-time dynamical systems to neural recordings via gradient descent as used for data-driven models, but we go beyond point-wise trajectory matching or single evaluation metrics. Gradient-descent based optimization allows us to integrate neuroscience-aligned loss functions simultaneously targeting multiple key network and dynamical descriptors of brain activity such as functional connectivity and its dynamical counterparts. As gradient-based optimization requires differentiability of evaluation metrics and several standard metrics such as connectivity dynamics distributions are non-differentiable, we introduce their differentiable surrogates for optimization and preserve the original metrics for evaluation. The hybrid fitting procedure matches the fitting accuracy of grid search with heightened computational efficiency, making data-driven continuous-time models such as Neural SDEs practical for the WBM setting.

Second, building on this efficient and accurate fitting procedure, we introduce a Hybrid model architecture that retains the physics-based WBM oscillator while replacing fixed diffusive inter-regional coupling with learned nonlinear stochastic interactions over the connectome. This design preserves the interpretable oscillatory prior while adding the expressivity needed to capture brain dynamics that are unknown or too complex to specify as explicit functions [29–31].

For quantitative comparison, we evaluate all three model architectures located along a spectrum between interpretability and flexibility using several validation tests. We demonstrate that the hybrid variant achieves overall higher data fidelity than physics-based and data-driven WBMs balancing representation of static and dynamic brain network features. We further observe that hybrid models perform robustly on datasets with small sample sizes, and even on individual data as required for personalized modeling. Beyond data fidelity, we take first steps toward assessing interpretability, demonstrating that the hybrid model preserves relevant network structure and signal complexity. Together, these results highlight the scalability, robustness, and flexibility of hybrid models as needed for future personalized whole-brain modeling and digital brain twin applications in neuroscience.

Results

In this study, we generate hybrid models for computational neuroscience by enhancing data-driven expressivity while maintaining physics-based interpretability. Our entire modeling workflow is summarized in Figure 1. Our starting point is a physics-based WBM (i.e., Hopf model) based on a nonlinear Stuart–Landau oscillator to reproduce fMRI activity (Figure 1a) [11, 12, 32]. In this formulation, oscillatory brain regions are

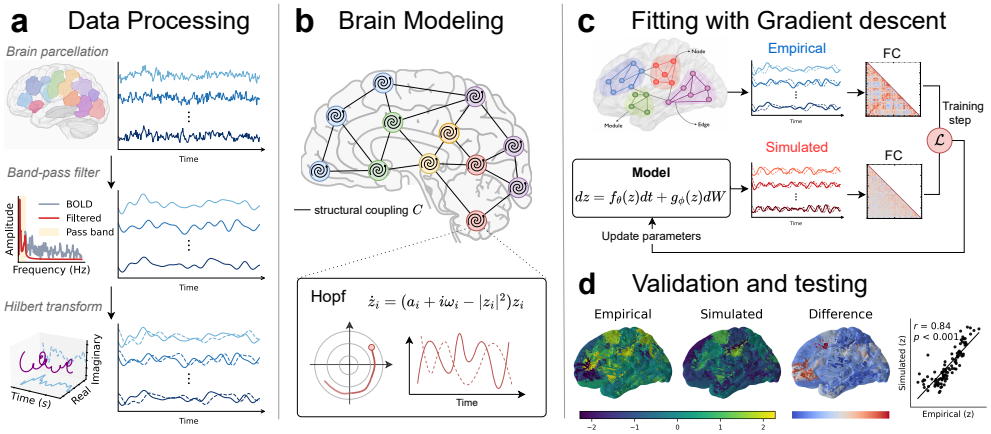

Fig. 1 Overview of the data-driven whole-brain model (WBM) fitting workflow. (a) *functional MRI (fMRI) data processing.* Raw blood-oxygen-level-dependent (BOLD) fMRI signals are parcellated into regions of interest (ROIs), filtered, standardized, and embedded in the complex plane via the Hilbert transform. **(b)** *Whole-brain modeling.* The brain is modeled as n regions, each a Hopf oscillator below the bifurcation, coupled through a connectivity matrix C that is fixed in the physics-based Coupled Hopf and learnable in the hybrid variants. **(c)** *Fitting with gradient-based optimization.* Empirical and simulated trajectories are reduced to functional connectivity (FC) matrices and additional evaluation metrics, whose discrepancy drives gradient-based parameter updates. **(d)** *Validation and testing.* The fitted models are evaluated on held-out data across a range of metrics for data fidelity, personalization, robustness to sample size, and interpretability.

coupled through a structural connectome estimated from diffusion MRI (Figure 1b) [33, 34]. Physics-based WBMs have been traditionally fit by grid search over few interpretable parameters [11, 12], but our fitting procedure uses gradient-based optimization to increase the scalability of model fitting, as explained in the next paragraph and in Figure 1c. Finally, the fitted models are validated on held-out data across a panel of metrics covering data fidelity, personalization, robustness to sample size, and interpretability (Figure 1d).

Scaling model fitting with neuroscientific target objectives: Multi-target fitting via gradient-based optimization

Although efficient model fitting is a prerequisite for data-driven modeling, it remains under-researched. Because grid search is still the most widely used model fitting technique for WBMs, we use it as the baseline method for comparison.

Figure 2a illustrates the grid search technique, which evaluates a lattice over a few parameters by statistically comparing evaluation metrics from empirical and simulated data. Three factors have made grid search the natural default: Parameter spaces are low-dimensional; many evaluation metrics describing stochastic neural activity are not directly differentiable (see Table 2); and stochastic forward simulations introduce gradient noise complicating first-order optimization. None of these factors precludes gradient-based optimization, but their cumulative impact explains why gradient-based optimization of stochastic WBMs has remained underused.

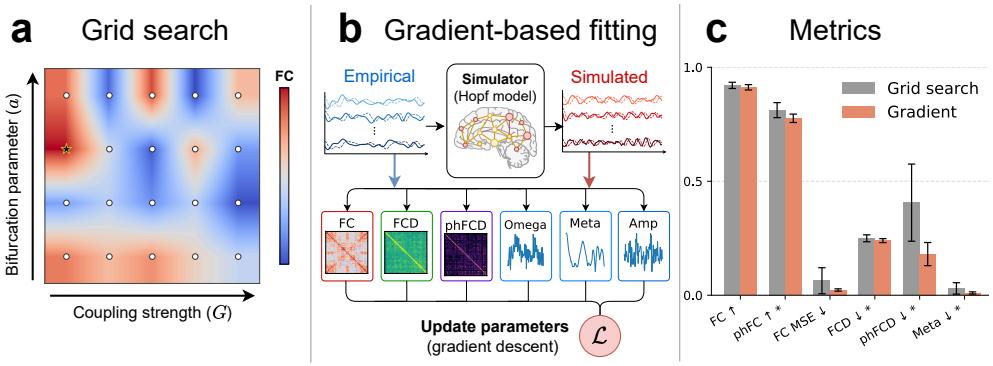

Fig. 2 Grid search vs. gradient-based optimization of the Hopf model. (a) Grid-search loss landscape over the bifurcation parameter a and global coupling strength G. **(b)** Gradient-based optimization pipeline through the differentiable simulator. Empirical and simulated trajectories are compared using functional connectivity (FC), functional connectivity dynamics (FCD), phase functional connectivity dynamics (phFCD), angular frequency ω (Omega), metastability (Meta), and amplitude (Amp), and their discrepancies are combined into the loss $\mathcal{L}$ for parameter updates. **(c)** Comparison of grid search and gradient-based optimization on the same Hopf equations. Metrics are FC correlation, phase functional connectivity (phFC) correlation, FC mean-squared error (MSE), FCD Kolmogorov–Smirnov (KS) distance, phFCD KS distance, and metastability error (Meta).

We acknowledge several alternatives between grid search and gradient-based optimization. Bayesian optimization methods [35–38] can return posterior parameter estimates. Still, posterior inference over the thousands of weights in a Neural SDE remains impractical, and neither Bayesian methods nor evolutionary search [39] utilize the derivative information. For our hybrid and Neural SDE model architectures, gradient-based optimization through the SDE solver [13–15] therefore becomes the most practical option (see Figure 2b).

Beyond efficiency, gradient-based optimization also enables multi-target optimization. This directly addresses the major shortcoming of single-metric objectives classically used in neuroscience, which lead to model degeneracy and fail to capture the full range of brain dynamics. Instead, our multi-target optimization combines complementary neuroscience-aligned metric families describing network structure and dynamics aggregated into a single composite loss $\mathcal{L}$, whose gradients update all model parameters simultaneously. Multi-target optimization matters for WBMs as simulators of stochastic brain activity, where network structure and dynamical reconfigurations reflecting global synchronization capture complementary aspects.

In addition to neuroscientific evaluation metrics, we include a finite-dimensional-matching (FDM) regularizer commonly used for fitting Neural SDEs, which stabilizes optimization by constraining the short-time transition structure of the trajectories.

Committing to gradient-based optimization is nontrivial, as it imposes a differentiability requirement on all loss function components. This is an obstacle for evaluation metrics focusing on dynamics of stochastic signals; e.g., the KS distance to compare FCD distributions is piecewise constant in the trajectory.

Our solution is to fit against a surrogate metric, i.e., the entrywise mean-squared error on matrices of brain dynamics (see Figure 2b for explanation of metrics), and to

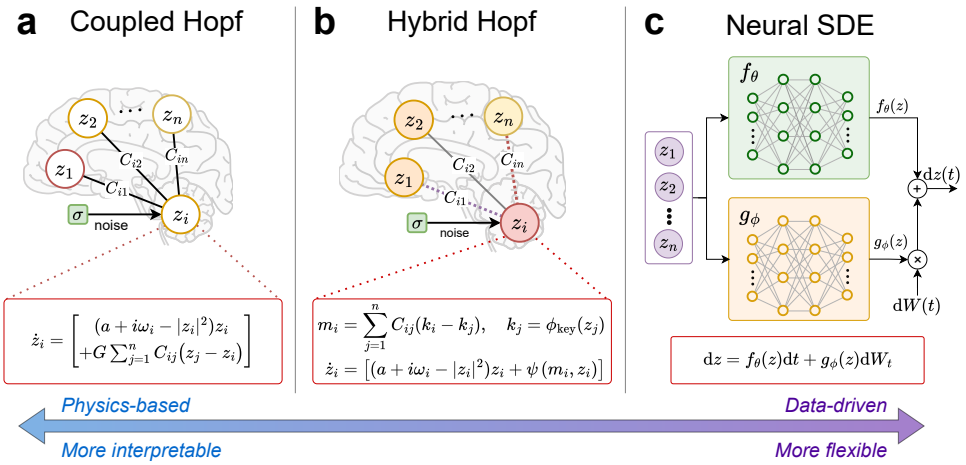

Fig. 3 Whole-brain model (WBM) architectures. The three model architectures span modeling paradigms from rigid but interpretable to flexible. **(a)** Coupled Hopf model, in which each region of interest (ROI) is a noisy Stuart–Landau oscillator coupled through the connectivity C. **(b)** Hybrid Hopf model, which keeps the local Hopf oscillator and replaces fixed diffusive coupling with a learned nonlinear interaction over C. **(c)** Neural stochastic differential equation (Neural SDE) model, which learns both the drift $f_\theta(z)$ and diffusion $g_\phi(z)$.

report the conventional distribution statistics (i.e., Kolmogorov-Smirnov (KS) distance) only at evaluation. We provide mathematical proof that the two are well aligned (see Supplementary Material).

Our conceptual considerations are supported by empirical results (Figure 2c and Supplementary Table 5). We compare grid search and gradient-based optimization and demonstrate that the two approaches reach similar performance for static metrics with an improved fit of dynamical metrics when using gradient-based optimization.

Our results justify gradient-based optimization for model fitting at a markedly lower computational cost (see Supplementary Table 6). Grid search entails a full simulation for each run, with cost growing exponentially with parameter count; gradient-based optimization cost scales *linearly* with parameter count. Our validated hybrid model fitting procedure now opens the possibility of using data-driven extensions for WBMs, as outlined next.

Whole-brain modeling from physics-based insights, data-driven solutions, and hybrid architectures

As shown in Figure 3, we utilize three models along a spectrum from rigid physics-based priors to fully data-driven. We place the Coupled Hopf model (Figure 3a) at the physics-based end, as it encodes established priors about oscillatory dynamics, connecting each local oscillator through a structural connectivity C. Due to its simplicity and interpretability, the Coupled Hopf model has been used in many neuroscientific applications [40, 41]. However, the same simplicity has resulted in reduced expressivity and worse fit to individual data.

At the data-driven end, we locate the Neural SDE (Figure 3c), which lets universal function approximators learn both the *drift* $f_\theta(z)$ and the *diffusion* $g_\phi(z)$ fully from data, resulting in less parameter interpretability.

Between these extremes, we propose a novel WBM architecture, i.e., the Hybrid Hopf model (Figure 3b), which incorporates relevant characteristics from both ends: It retains the interpretable local Hopf oscillator while replacing the fixed diffusive coupling with a learned nonlinear interaction over C, letting the interactions depart from the structural connectome, consistent with recent studies [42, 43].

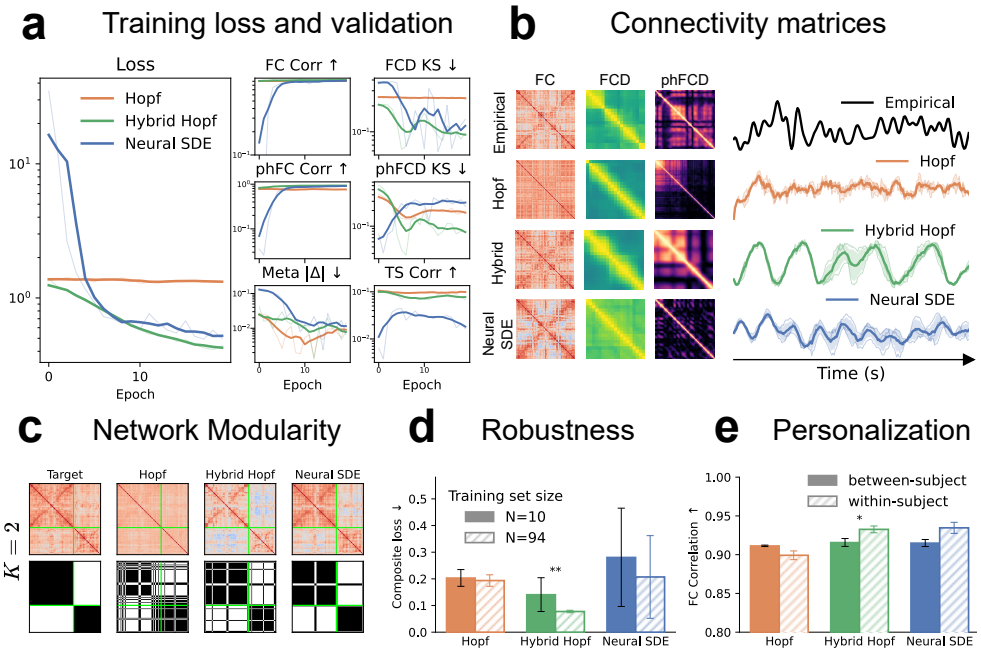

Fig. 4 Model validation overview. (a) Training loss and validation diagnostics across epochs for the three fitted model architectures. **(b)** Empirical functional connectivity (FC), functional connectivity dynamics (FCD), and phase FCD (phFCD) matrices for a held-out subject. Simulated samples for the three models compared against the empirical reference. **(c)** Brain clustering analysis over the empirical FC matrix and reordering using the obtained classes ($K = 2$). **(d)** Robustness analysis showing the effect of the dataset size on the global performance of the three model architectures. **(e)** Patient personalization analysis comparing within-subject and between-subject FC reconstruction for the three model architectures.

Within the hybrid family we evaluate three variants: the *Hybrid Hopf*, which routes the inter-regional coupling through a learned message-passing interaction over the connectome; the *GNN-Hopf*, which replaces the coupling with a graph neural network layer acting on a learnable connectome; and *Hopf+Neural*, which keeps the full physics-based drift and adds an unconstrained neural correction. As the Hybrid Hopf achieves strong overall performance while maintaining interpretability and the priors of the baseline connectome (see Supplementary Table 3), we use it as the representative hybrid model for comparison with non-hybrid architectures.

Table 1 Model architecture comparison on the Human Connectome Project (HCP) resting-state dataset. Functional connectivity (FC) correlation; phase functional connectivity (phFC) correlation; functional connectivity dynamics (FCD) Kolmogorov–Smirnov (KS) distance; phase FCD (phFCD) KS distance; metastability absolute error (Meta $|\Delta|$); FC adjusted Rand index (ARI). Arrows indicate the direction of improvement (↑ higher is better, ↓ lower is better). Best per column in bold.

Model	FC corr ↑	phFC corr ↑	FCD KS ↓	phFCD KS ↓	Meta $\lvert\Delta\rvert$ ↓	FC ARI ↑
Coupled Hopf	0.913 ± 0.011	0.783 ± 0.021	0.242 ± 0.009	0.208 ± 0.092	9.7×10^{-3} ± $5.4{\times}10^{-3}$	0.124 ± 0.123
Neural SDE	0.915 ± 0.005	0.911 ± 0.001	**0.049 ± 0.006**	0.305 ± 0.001	6.7×10^{-3} ± $\mathbf{1.3}{\times}10^{-3}$	**0.328 ± 0.273**
Hybrid Hopf	**0.926*** **± 0.009**	**0.923**** **± 0.013**	0.098 ± 0.019	**0.099 ± 0.040**	0.010 ± 0.005	0.210 ± 0.190

We evaluate all three model architectures on the high-quality standard Human Connectome Project (HCP) resting-state fMRI benchmark [44] (see Methods for details on the dataset) using four complementary evaluation criteria relevant for digital twins: data fidelity, personalization capability, robustness to small sample size, and interpretability.

The training dynamics in Figure 4a show that all metrics improve over training, but display differences between model architectures. The Coupled Hopf saturates after the first epoch and barely improves thereafter, consistent with its limited expressivity. The Neural SDE reduces the composite loss in early epochs, reflecting the flexibility of its fully learned drift and diffusion. The Hybrid Hopf fits most stably and reaches the lowest final loss, indicating that the oscillatory prior and the learned nonlinear coupling are complementary. Some metrics, such as phFCD KS, fluctuate more because we optimize a surrogate rather than this metric directly.

Hybrid models provide the most balanced data fidelity across several evaluation metrics

First, we assess data fidelity using neuroscience-aligned evaluation metrics. *Network-structure* metrics describe static (i.e., time-averaged) connectivity, capturing average statistical relationships between brain regions: static functional connectivity (FC), phase-based FC (phFC), and modularity, quantified by the adjusted Rand index (ARI) between empirical and simulated community partitions. *Network-dynamics* metrics capture how connectivity and global synchrony evolve over time, represented by functional connectivity dynamics (FCD), phase FCD (phFCD), and metastability. Additional metrics describing the time-series trajectories are reported in the supplementary results (see Supplementary Material).

A visual inspection of exemplary FC matrices and time series (Figure 4b) suggests that dynamic connectivity is least expressive in the Coupled Hopf model, whereas both the Neural SDE and the Hybrid model produce richer dynamics.

Quantification of evaluation metrics reflecting network structure (Table 1) displays high static FC correlation across all architectures, while phFC and modularity (ARI) reach a higher fit in the Neural SDE and Hybrid models. The Hybrid model attains the highest static and phase FC correlation overall.

Regarding metrics representing network dynamics, the Neural SDE and Hybrid models generally have a higher fit than the Coupled Hopf model. The Neural SDE leads on connectivity dynamics (FCD) and achieves the lowest metastability error,

while the Hybrid model achieves the lowest phFCD error. Overall, the Hybrid model combines balanced, high performance across metric types; the modest gap between the Neural SDE and Hybrid Hopf raises the question of whether a fully data-driven solution justifies its use for WBMs.

The evaluation metrics confirm that the Coupled Hopf model imposes a strong inductive bias through its fixed coupling derived from the structural connectome. This limits its expressivity, as reflected in the FC and FCD matrices. In turn, the learned nonlinear coupling in the Hybrid Hopf recovers dynamics that the fixed structural coupling cannot express.

Hybrid models display robustness, data efficiency, and suitability for personalization

After evaluating data fidelity, we turn to two properties that are often limiting in neuroscientific settings: *robustness*, i.e., the stability of model fit across repeated runs, measured by the standard deviation σ of the test error; and *data efficiency*, i.e., the ability to fit well from few recordings, measured by the test error ε in the small-sample regime ($N = 10$) and its degradation relative to the large-sample regime ($N = 94$).

We refit each architecture on a small ($N = 10$) and large ($N = 94$) sample size, aggregating four core metrics (FC and phFC correlations, FCD and phFCD KS) into a composite test error (Figure 4d). The Coupled Hopf is data efficient due to a small parameter count ($\varepsilon = 0.204$ at $N = 10$ vs. 0.194 at $N = 94$). However, while a strong inductive bias leads to robustness ($\sigma = \pm 0.031$ at $N = 10$ and ± 0.022 at $N = 94$), there is little room for additional data to improve the WBM with accordingly small variance. In terms of *data efficiency*, the Neural SDE degrades the most when data is limited ($\varepsilon = 0.281$ at $N = 10$ vs. 0.207 at $N = 94$). Its large variance ($\sigma = \pm 0.184$ at $N = 10$ and ± 0.155 at $N = 94$) reflects low *robustness*, consistent with an over-parameterized model at low sample sizes. The Hybrid Hopf shows the best fit with high data efficiency ($\varepsilon = 0.141$ at $N = 10$ vs. 0.078 at $N = 94$, paired Wilcoxon $p < 0.01$ across 10 seeds), accompanied by high robustness ($\sigma = \pm 0.063$ at $N = 10$ and ± 0.005 at $N = 94$). The Hybrid Hopf therefore combines the data efficiency of the physics-based baseline with robustness to different training-set sizes, which is the regime most relevant for clinical use cases with small sample sizes.

We now explore a use case for clinical purposes: personalization, i.e., fitting a model to $N = 1$. A model used as a personalized digital twin must be able to preserve subject-specific signatures. Thus, our personalization analysis (Figure 4e) tests whether each model captures subject-specific network structure measured by FC or only a population-level fit. We use a two-level split of the dataset: across subjects into training and test sets, and within each training subject into two temporal halves. Models are fit on the first half of the training data. The *within-subject* correlation then evaluates FC reconstruction on the *second* half of the same training subjects (seen subject, unseen time window), while the *between-subject* correlation evaluates FC reconstruction on held-out test subjects (unseen subject). The Neural SDE and the Hybrid Hopf reach the highest within-subject FC correlations while their between-subject correlations stay close to the population baseline, producing positive within$-$between gaps Δ of $+0.020$ and $+0.017$, respectively. The physics-based Coupled Hopf shows the opposite

pattern: its within-subject correlation falls slightly *below* the between-subject baseline, yielding a negative gap of $\Delta = -0.012$. The physics-based prior alone therefore collapses individual variability into a population-level fit, which is useful for group analyses but less suitable for digital twins. In contrast, data-driven and hybrid architectures preserve a personalized neural activity signature of their dynamics [6, 45].

Hybrid models reproduce modular network structure and complexity of brain network dynamics

Our results show enhanced data fidelity, robustness, data efficiency, and personalization capability in hybrid models. We proceed to evaluate selectedinterpretability metrics, aside from interpretable built-in model parameters. Because interpretability of neural networks remains an open research topic [46], we focus on interpreting static and dynamic descriptors of brain network activity, namely modular network organization and complexity markers of neural activity.

As a general connectivity fit used to assess data fidelity does not inform us *how* hybrid models can better reproduce brain network structure, we assess the modular organization of simulated FC and its similarity to the empirical one. Spectral clustering of the empirical FC is used to obtain a community partition ($K = 2$ in Figure 4c; $K = 3$ in Supplementary Figure 5a). Both the Neural SDE and the Hybrid Hopf recover the block-diagonal community structure of the empirical FC with high fidelity, with the Hybrid Hopf yielding the sharpest within-module contrast against the off-diagonal background. The Coupled Hopf produces a flatter reordering with barely visible community boundaries, consistent with the limitations of fixed coupling: learnable, dynamic interactions are required to absorb the modular structure that a structural connectome on its own cannot produce. Our observations are quantified by the FC ARI (Table 1).

To interpret the effect of hybrid models on brain dynamics, we probe the complexity of the simulated dynamics by computing the Shannon entropy and Lempel–Ziv–Welch complexity of the FCD and phFCD matrices (Supplementary Figure 5b). We demonstrate that the hybrid prior reproduces the information richness of temporal network reconfiguration better than physics-based and data-driven WBMs without overfitting, possibly reflecting state-dependent connectivity changes.

Discussion

In this paper we pioneer a framework for hybrid whole-brain modeling that combines physics-based Hopf oscillators with data-driven components, fitted with gradient-based optimization against a multi-target loss tailored to neuroscience-relevant metrics reflecting brain network structure and dynamics. Our enhanced modeling workflow yields higher data fidelity compared to physics-based WBMs, while preserving the interpretability of the physics-based component compared to the data-driven Neural SDE. The Hybrid Hopf additionally retains individual dynamical properties, which are relevant for personalized digital twins [6, 47].

Inspired by scientific machine learning, we show that gradient-based optimization through an SDE solver is a successful option for data-driven WBMs with thousands of

parameters [13–15]. Gradient-based optimization necessitates differentiable evaluation metrics, which are not compatible with commonly used metrics to compare distributions of stochastic signal dynamics. We address this limitation by fitting against a differentiable *surrogate* loss and reporting distribution statistics only for evaluation purposes. Our results demonstrate high accuracy using surrogacy metrics. However, multi-target may introduce potential risks, such as local minima and challenges in hyperparameter tuning due to loss-weight balancing. Future research could build on our approach by focusing on combining Bayesian techniques with efficient SDE solvers to provide uncertainty quantification of parameters [35–37].

Enabled by efficient fitting, we first consider two potential hybrid model architectures. *Additive residual learning*, denoted by *Hopf+Neural*, retains the full physics-based drift and adds a learned correction [48]. *Component replacement* substitutes a specific physics-based element, i.e., the coupling, with a learnable analogue while preserving local Hopf dynamics [49], mimicking state-dependent changes in connectivity, such as those induced by neurotransmitters. Within this family we consider two variants: (i) *Hybrid Hopf* which replaces the linear structural connectivity with a learned message-passing interaction over the connectome [50, 51]. (ii) the *GNN-Hopf*, which routes the coupling through a graph-convolutional layer with a learnable connectivity matrix unrelated to the structural connectome.

Among hybrid variants, the component replacement variants exhibits higher data fidelity than the additive-residual variant. We attribute this to the inductive bias imposed by learnable corrections on the connectome, which proves more effective for whole-brain dynamics thanan unconstrained additive drift. The component replacement variant offers interpretability, as: the correction stays inside the coupling term and can be interpreted as data-driven connectivity [52]. In contrast, the unconstrained additive residual may also influence both connectivity and local dynamics, making it difficult to understand its effects. Given the similar data fidelity of Hybrid Hopf and GNN-Hopf but higher interpretability of the Hybrid Hopf, we select the *Hybrid Hopf* for comparisons with physics-based and data-driven WBMs.

Comparison across model architectures shows that the Hybrid Hopf model displays the most balanced data fidelity across multiple metrics describing network structure and dynamics compared to its physics-based and data-driven model competitors. Further, we find that the Hybrid Hopf model exceeds the Coupled Hopf and Neural SDE in terms of robustness and data efficiency, showing added value for smaller sample sizes and particular suitability for neuroscientific studies. Hybrid and data-driven models perform well on single-subject datasets, as needed for personalized digital brain twins, exceeding the classical Hopf model, which collapses to a population mean [45]. These results stress the suitability of hybrid models to be used as personalized WBMs while maintaining computational efficiency.

Lastly, our study contributes first analyses towards *mechanistic* interpretability of hybrid and data-driven models, emphasizing the preservation of interpretable parameters and outputs like network structure and complexity. However, to ensure ongoing trust by neuroscientific end users, further research on interpretability and explainable AI is essential [46].

Methods

Figure 1 illustrates the workflow. All code, trained models, and evaluation scripts are publicly available (see Code availability).

Mathematical setting and notation

Let $T > 0$ be a fixed time horizon and let $n \geq 1$ denote the number of brain regions of interest (ROIs) under consideration. We represent brain activity as a vector-valued signal $z : [0, T] \to \mathcal{X}$ whose i-th component $z_i(t)$ records the state of region $i \in \{1, \ldots, n\}$ at time t. The state space $\mathcal{X}$ is $\mathbb{R}^n$ for real-valued signals and $\mathbb{C}^n$ when working with analytic signals (which represent of oscillatory activity).

We model brain state trajectories as solutions of differential equations. In the simplest deterministic case, the evolution takes the form of an ordinary differential equation (ODE),

$$\dot{z}(t) = f\big(z(t)\big), \quad z(0) = z_0. \tag{1}$$

The vector field f is in general unknown; estimating it from recorded brain activity is the central objective of this work. However, large-scale brain activity exhibits stochastic fluctuations, such as neural noise, physiological variability, and unmodeled inputs (e.g., effects of neurotransmitters inducing state changes). Accordingly, a deterministic model is often not expressive enough to capture the variability in the data. We therefore work with stochastic differential equations (SDEs),

$$dz(t) = f\big(z(t)\big)\, dt + \sigma\big(z(t)\big)\, dW_t. \tag{2}$$

Here W_t is a standard Wiener process and $\sigma : \mathcal{X} \to \mathbb{R}^{n \times n}$ is a diffusion coefficient. Thus, altogether the term $\sigma(z(t))\, dW_t$ introduces stochasticity in the dynamics.

This mechanism increases expressivity at the expense of structure. Indeed, because different noise realizations produce different sample paths, the fitness of a stochastic model such as (2) has to be assessed using evaluation metrics computed from the simulated and empirical signals rather than by pointwise trajectory comparison, as discussed below.

We model the available data as trajectories, recording at every time step t a state observation $x(t)$. More precisely, we work with a dataset consisting of N recordings; each recording $k \in \{1, \ldots, N\}$ provides an initial condition $x_0^{(k)}$ and a trajectory $\{x^{(k)}(t) : t \in [0, T]\}$. Throughout the document, time is measured in seconds, in agreement with the resolution of the empirical fMRI recordings (i.e., repetition time $\mathrm{TR} = 0.72\,\mathrm{s}$).

We denote the neural activity signal across n ROIs as $x : [0, T] \to \mathbb{R}^n$.

Data and preprocessing

The dataset consists of BOLD time series from the preprocessed high-quality resting-state HCP dataset [44], collected from 94 young healthy subjects during resting-state fMRI sessions. Each subject's brain is parcellated into $n = 100$ cortical regions of interest following the Schaefer atlas [53], and the recording yields $T_{\mathrm{obs}} = 1{,}200$ time points per subject at a repetition time (TR) of 0.72 s.

The signal recorded from each brain region is the BOLD signal, a hemodynamic proxy for local neural activity measured by fMRI [54]. This signal is a noisy, broadband time series, and before fitting any model we apply standard preprocessing to remove movement-related artefacts and to isolate the relevant frequency range of fMRI-related slow oscillations. Following standard practice in whole-brain modeling [40, 52], we apply the following preprocessing pipeline:

1. **Standardization.** We z-score each region's time series per recording to remove baseline offsets.
2. **Bandpass filtering.** We retain only frequencies in $[0.008, 0.08]$ Hz, the band associated with slow hemodynamic fluctuations.
3. **Complexification.** We apply the Hilbert transform to obtain a complex analytic signal $z(t) = (z_1(t), \dots, z_n(t)) \in \mathbb{C}^n$, whose i-th component $z_i(t)$ has modulus $|z_i(t)|$ capturing the envelope amplitude and argument $\phi_i(t) = \angle z_i(t)$ capturing the instantaneous phase of region i. The real part $\Re(z(t)) = x(t)$ recovers the filtered BOLD signal.

Metrics

Stochastic oscillatory models are inherently generative: different noise realizations produce different sample paths, so pointwise trajectory matching is ill-posed. The standard practice in whole-brain modeling is therefore to compare *evaluation metrics* describing key static characteristics of neural activity [8, 45]. These metrics are robust to stochastic variability, while still retaining the spatial and temporal correlation structure with neuroscientific meaning. In addition to these neuroscience-aligned metrics, we add a finite-dimensional-matching regularizer during training (FDM, see below), which is commonly used to stabilize the optimization of more flexible models such as the Neural SDE [55].

All metrics below are defined for a pair of trajectories, a simulated one $\hat{z}$ and an empirical one z, both observed on a common interval $[0, T]$ over n ROIs and taking values in $\mathbb{C}^n$ (when the analytic signal is needed) or in $\mathbb{R}^n$ (when only the real-valued BOLD signal is needed). For notational convenience we write $x_i(t) := \Re(z_i(t))$ for the real part, $a_i(t) := |z_i(t)|$ for the envelope amplitude, and $\phi_i(t) := \angle z_i(t)$ for the instantaneous phase of region i.

Differentiability constraint. Since the models are fitted by gradient-based optimization, every loss term must be differentiable with respect to the simulated trajectory $\hat{z}$. This is a nontrivial constraint for metrics that represent distributions of stochastic dynamics, i.e., the two-sample Kolmogorov–Smirnov (KS) distance [56, 57], which is used in our setting to compare the FCD and phFCD distributions: for two samples $\{u_i\}_{i=1}^{N_u}$ and $\{v_j\}_{j=1}^{N_v}$ with empirical cumulative distribution functions (CDFs) $\hat{F}_u$ and $\hat{F}_v$, it is defined as $D_{N_u,N_v} := \sup_{x\in\mathbb{R}} \left|\hat{F}_u(x) - \hat{F}_v(x)\right|$, which is piecewise constant in $\hat{z}$ and non-differentiable. For these metrics we therefore use a differentiable *surrogate* during training : an entrywise mean-squared error between the simulated and empirical FCD (resp. phFCD) matrices. The KS distance is reserved for evaluation. For the remaining training metrics, we can directly use differentiable comparisons. All metrics

and their use are summarized in Table 2and each metric is then explained in more detail in the following text.

Table 2 Overview of neuroscience-aligned metrics, grouped by the neural feature they describe. Functional connectivity (FC); functional connectivity dynamics (FCD); phase FC (phFC); phase FCD (phFCD); Kolmogorov–Smirnov (KS) distance; mean-squared error (MSE); region of interest (ROI); maximum mean discrepancy (MMD); finite-dimensional matching (FDM); metastability (Meta); adjusted Rand index (ARI). *Diff.* marks whether the evaluation comparison admits useful gradients in $\hat{z}$ (✓) or not (×). The *Role* badge encodes how each metric enters the pipeline: T training term, S differentiable surrogate, R regularizer, and E evaluation.

#	Metric	Comparison at evaluation	Diff.	Role
Static functional connectivity				
1	FC corr	Pearson corr. on $\text{vec}_\triangle\text{FC}$	✓	T E
2	FC MSE	MSE on $\text{vec}_\triangle\text{FC}$	✓	T
3	phFC corr	Pearson corr. on $\text{vec}_\triangle\text{FC}^\phi$	✓	T E
4	FC ARI	Adjusted Rand index between FC clusters	×	E
Functional connectivity dynamics				
5	FCD KS	two-sample KS on $\text{vec}_\triangle$FCD entries	×	E
6	FCD MSE	MSE on $\text{vec}_\triangle$FCD entries (surrogate for #5)	✓	S
7	phFCD KS	two-sample KS on $\text{vec}_\triangle$phFCD entries	×	E
8	phFCD MSE	MSE on $\text{vec}_\triangle$phFCD entries (surrogate for #7)	✓	S
Node-level oscillatory features				
9	Amplitude	Per-ROI MSE on $\bar{a}_i$	✓	T
10	Frequency	Per-ROI MSE on $\bar{\omega}_i$	✓	T
11	Meta $\lvert\Delta\rvert$	$\lvert\text{Meta}(\hat{z}) - \text{Meta}(z)\rvert$	✓	T E
Temporal fidelity				
12	FDM	MMD on two-time joints of z	✓	R
13	TS corr	Mean per-ROI Pearson corr. on $\Re(z)$	✓	E
14	TS PSD	MSE on normalized power spectra of $\Re(z)$	✓	E
15	Autocorr	MSE on per-ROI autocorrelations of $\Re(z)$	✓	E

Functional connectivity (FC).

The most widely adopted evaluation metric in whole-brain modeling is the matrix of pairwise Pearson correlations between ROI time series, averaged over the recording, capturing the full pairwise covariance pattern in a single $n \times n$ matrix [8, 58]. A loss on this matrix penalizes mismatches in the network structure of simulated and empirical activity. Define per-ROI temporal mean and standard deviation

$$\tilde{x}_i(t) := \frac{x_i(t) - \mu_i}{\sigma_i}, \quad \text{where} \quad \mu_i := \frac{1}{T}\int_0^T x_i(t)\,\mathrm{d}t, \quad \sigma_i^2 := \frac{1}{T}\int_0^T \big(x_i(t) - \mu_i\big)^2\,\mathrm{d}t.$$

The functional connectivity matrix is the Pearson correlation matrix

$$\mathrm{FC}_{ij}(z) := \frac{1}{T}\int_0^T \tilde{x}_i(t)\,\tilde{x}_j(t)\;\mathrm{d}t, \qquad i,j = 1,\dots,n.$$

By symmetry, all the information in $\mathrm{FC}(z)$ is contained in its strict upper triangle, which we vectorize into a vector $\mathrm{vec}_\triangle(\mathrm{FC}(z)) \in \mathbb{R}^q$ of length $q = n(n-1)/2$. Agreement between a simulated FC matrix $\mathrm{FC}(\hat{z})$ and an empirical one $\mathrm{FC}(z)$ is then quantified by two scalar comparisons on these vectorized triangles, namely their mean-squared error and their Pearson correlation.

Functional connectivity dynamics (FCD).

FCD captures this nonstationarity of the signal by computing FC in sliding windows and measuring the similarity between all pairs of windows; the resulting distribution describes how fluidly the brain transitions between correlation patterns, and penalizes models that reproduce the correct average correlation pattern but with incorrect temporal variability [24, 25].

We set the window length $h > 0$ to $h = 30\,\mathrm{s}$ with a window kernel $w : \mathbb{R} \to [0,\infty)$ with $\int w = 1$ and step size of $2\,\mathrm{s}$. The corresponding time-localized weight centered at τ is $w_\tau(t) := \frac{1}{h}\,w\big(\frac{t-\tau}{h}\big)$. The windowed first and second moments are

$$\mu_i(\tau) := \int_0^T w_\tau(t)\,x_i(t)\;\mathrm{d}t, \qquad \sigma_i^2(\tau) := \int_0^T w_\tau(t)\,\big(x_i(t)-\mu_i(\tau)\big)^2\;\mathrm{d}t,$$

and the windowed FC at time τ is

$$\mathrm{FC}_{ij}(\tau) := \int_0^T w_\tau(t)\,\tilde{x}_i(t;\tau)\,\tilde{x}_j(t;\tau)\;\mathrm{d}t, \quad \text{where} \quad \tilde{x}_i(t;\tau) := \frac{x_i(t)-\mu_i(\tau)}{\sigma_i(\tau)}.$$

Denoting by $v(\tau) := \mathrm{vec}_\triangle(\mathrm{FC}(\tau)) \in \mathbb{R}^q$ its vectorized strict upper triangle, the FCD kernel

$$\mathrm{FCD}(\tau,\tau') := \mathrm{corr}\big(v(\tau), v(\tau')\big)$$

measures how similar the windowed correlation patterns at times τ and τ' are. On a window grid $\mathcal{T}$, the strict upper-triangular entries form the finite sample $\mathcal{S}_{\mathrm{FCD}}(z) := \big(\mathrm{FCD}_z(\tau,\tau')\big)_{\tau,\tau'\in\mathcal{T},\,\tau<\tau'}$, whose empirical distribution captures the temporal variability of FC. At evaluation, the FCD distance is the two-sample KS distance $D\big(\mathcal{S}_{\mathrm{FCD}}(\hat{z}), \mathcal{S}_{\mathrm{FCD}}(z)\big)$, which probes mismatches across the full range of similarity values rather than at their mean alone. Since D is non-differentiable in $\hat{z}$, training instead minimizes the full-matrix MSE

$$\frac{1}{|\mathcal{T}|^2}\sum_{\tau,\tau'\in\mathcal{T}}\big(\mathrm{FCD}_{\hat{z}}(\tau,\tau') - \mathrm{FCD}_z(\tau,\tau')\big)^2$$

on the shared window grid. This differentiable surrogate matches each window-pair similarity directly (see Supplementary Material).

Phase-coherence FC (phFC).

Complementing FC based on amplitudes, phFC isolates the *phase synchrony*, a measure that asks how consistently the phase difference between two regions stays fixed over the recording [40, 52, 59]. It is defined as

$$\mathrm{FC}^{\phi}_{ij}(z) := \frac{1}{T}\int_0^T \cos\big(\phi_i(t) - \phi_j(t)\big)\,\mathrm{d}t, \qquad i,j = 1,\dots,n.$$

A value close to 1 indicates that the phase difference $\phi_i - \phi_j$ stays nearly constant in time, while values close to zero indicate that the relative phase drifts uniformly. The matrix FC^{ϕ} is symmetric with unit diagonal and entries in $[-1, 1]$. Agreement between simulated $\mathrm{FC}^{\phi}(\hat{z})$ and empirical $\mathrm{FC}^{\phi}(z)$ is measured by the Pearson correlation between their upper triangles.

Phase FCD (phFCD).

Analogous to the relationship between static FC and FCD, phFCD extends phFC to the time-varying setting, capturing how phase-synchrony patterns reconfigure over time [40, 52, 59]. At each time point t, we form the instantaneous phase-coherence matrix

$$P_{ij}(t) = \cos\big(\phi_i(t) - \phi_j(t)\big),$$

and vectorize its strict upper triangle: $\mathbf{p}(t) = \mathrm{vec}_{\triangle}(P(t)) \in \mathbb{R}^q$. The phFCD similarity kernel is

$$\mathrm{phFCD}(t,t') = \frac{\mathbf{p}(t)^{\top}\mathbf{p}(t')}{\|\mathbf{p}(t)\|_2\,\|\mathbf{p}(t')\|_2}.$$

Unlike windowed FCD, phFCD operates at full temporal resolution with no window-length dependency. Evaluation uses the KS distance on upper-triangular entries; training uses an MSE surrogate.

Amplitude and instantaneous frequency.

The FC and FCD families constrain pairwise relationships between regions but do not feature information about univariate signal properties. We therefore add two simple per-region terms, one ensuring that each region oscillates with the correct average power, the other that it does so at the correct average rate. Writing the analytic signal of region i in polar form as $z_i(t) = a_i(t)\,e^{\mathrm{i}\phi_i(t)}$, the envelope $a_i(t)$ plays the role of an instantaneous oscillation amplitude, and the derivative of the phase $\omega_i(t) := \frac{\mathrm{d}}{\mathrm{d}t}\phi_i(t)$ (with ϕ_i unwrapped modulo 2π) plays the role of an instantaneous angular frequency. Averaging each over the recording gives a per-ROI mean amplitude $\bar{a}_i$ and a per-ROI mean angular frequency $\bar{\omega}_i$. The amplitude loss is the mean-squared error between simulated and empirical $\bar{a}_i$ across regions, and the frequency loss is the analogous mean-squared error on $\bar{\omega}_i$.

Time series fidelity.

The following three diagnostics provide complementary trajectory-level checks; in our experiments they serve as held-out evaluation measures rather than training targets.

All three operate on the real part $\Re(z_i(t))$ of the analytic signal, which we denote $x_i(t)$ throughout this paragraph, and reuse the per-ROI mean μ_i, standard deviation σ_i, and standardized signal $\tilde{x}_i$ introduced for FC.

- *Time series correlation (TS corr)* is the mean per-ROI Pearson temporal correlation between simulated $\hat{x}$ and empirical x, which measures pointwise waveform agreement (higher is better):

$$\mathrm{TScorr}(\hat{x}, x) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{T} \int_0^T \tilde{\hat{x}}_i(t)\, \tilde{x}_i(t)\, \mathrm{d}t.$$

- *Time series power spectral density (TS PSD)* compares the spectral shape of each ROI through the power spectral density (PSD). Let $S_i^x(\xi) := |\mathcal{F}\{x_i\}(\xi)|^2$ denote the one-sided power spectrum and write its L^1-normalization as $\tilde{S}_i^x(\xi) := S_i^x(\xi) \big/ \int_0^{\xi_{\mathrm{Ny}}} S_i^x(\xi')\, \mathrm{d}\xi'$, where ξ_{Ny} is the Nyquist frequency. The TS PSD distance is the mean-squared L^2 gap between normalized spectra, averaged across ROIs and frequencies (lower is better):

$$\mathrm{TSPSD}(\hat{x}, x) := \frac{1}{n\, \xi_{\mathrm{Ny}}} \sum_{i=1}^{n} \int_0^{\xi_{\mathrm{Ny}}} \big(\tilde{S}_i^{\hat{x}}(\xi) - \tilde{S}_i^x(\xi)\big)^2\, \mathrm{d}\xi.$$

- *Autocorrelation (Autocorr)* compares per-ROI temporal memory. Define the lag-τ normalized autocorrelation

$$\rho_i^x(\tau) := \frac{1}{\sigma_i^2\, (T - \tau)} \int_0^{T-\tau} \big(x_i(t) - \mu_i\big)\big(x_i(t+\tau) - \mu_i\big)\, \mathrm{d}t.$$

  The autocorrelation distance is the mean-squared gap between $\rho_i^{\hat{x}}$ and ρ_i^x up to a maximum lag $\tau_{\max}$ (set to 50 time steps, and truncated to $T/2$ for shorter trajectories), which penalizes errors in temporal memory structure (lower is better):

$$\mathrm{Autocorr}(\hat{x}, x) := \frac{1}{n\, \tau_{\max}} \sum_{i=1}^{n} \int_0^{\tau_{\max}} \big(\rho_i^{\hat{x}}(\tau) - \rho_i^x(\tau)\big)^2\, \mathrm{d}\tau.$$

Metastability.

Metastability, i.e. the temporal variability of global phase coherence, captures how much the network fluctuates between synchronized and desynchronized states, a hallmark of resting-state brain dynamics that pairwise connectivity measures do not directly constrain [8, 60]. The Kuramoto order parameter measures instantaneous global phase coherence:

$$R(t) = \left| \frac{1}{n} \sum_{i=1}^{n} e^{\mathrm{i}\phi_i(t)} \right|,$$

and metastability is its temporal standard deviation:

$$\mathrm{Meta}(z) := \sqrt{\frac{1}{T}\int_0^T \big(R(t) - \bar{R}\big)^2 \, \mathrm{d}t}, \qquad \bar{R} := \frac{1}{T}\int_0^T R(t) \, \mathrm{d}t.$$

We penalize the ℓ_1 distance $|\mathrm{Meta}(\hat{z}) - \mathrm{Meta}(z)|$ between simulated and empirical metastability.

Modularity: Spectral clustering and adjusted Rand index (ARI).

In addition to general network matching via FC and phFCD, we calculate modularity, i.e. the partition of ROIs into densely connected communities. We probe this by spectral clustering of the FC matrix. The matrix FC is first mapped to a non-negative affinity $A \in [0,1]^{n\times n}$ via $A = (1 + \mathrm{FC})/2$, with unit diagonal. We then run spectral clustering on A as a precomputed affinity, using the normalized Laplacian eigenmap and k-means label assignment on the leading K eigenvectors, with $K = 2$ chosen to reflect the dominant block structure observed in the empirical reference. Agreement between a simulated partition $\hat{\pi}$ and the corresponding empirical partition π is quantified by the adjusted Rand index

$$\mathrm{ARI}(\hat{\pi}, \pi) := \frac{\mathrm{RI}(\hat{\pi}, \pi) - \mathbb{E}[\mathrm{RI}]}{1 - \mathbb{E}[\mathrm{RI}]},$$

where RI counts the fraction of ROI pairs on which the two partitions agree (both assigned to the same community in $\hat{\pi}$ and π, or both to different ones) and $\mathbb{E}[\mathrm{RI}]$ is its expectation under a random labeling with the same cluster sizes. The ARI is 1 for identical partitions and 0 in expectation under chance, so it directly measures whether the model recovers the empirical community structure rather than just matching the entrywise correlation pattern. Like the time-series fidelity diagnostics, the ARI is used at evaluation only and is not part of the training loss.

Training

Training procedure

Recordings are split into training (70%), validation (15%), and test (15%) partitions. Training windows are sampled from the first temporal half of each training recording; the second half of those same recordings is reserved for intra-subject evaluation, while held-out subjects define the inter-subject test set. The checkpoint with the best validation loss is selected for evaluation.

All models fitted by gradient-based optimization use the Adam optimizer at a fixed learning rate of 10^{-3}. Training runs for up to 20 epochs with early stopping (patience = 15 epochs, monitoring validation loss). Each epoch draws 1,024 windows on-the-fly from the training recordings; windows have length 100 time points (72 s at TR = 0.72 s) and are assembled into mini-batches of size 128. Gradients are clipped to unit norm before each parameter update. SDE trajectories are integrated with Euler–Maruyama at an internal step of $\Delta t = 0.05$ s; for the Hybrid Hopf model the adjoint method [15] is used to reduce GPU memory during the backward pass. All experiments are run on a single NVIDIA GeForce RTX 2080 Ti GPU (11 GB) on an

Intel Xeon Gold 6134 node, using PyTorch 2.10.0 with CUDA 12.8; reported values (mean ± std) are computed across five independent runs with different random seeds. A detailed breakdown of the training and inference wall-clock cost of each model is given in Supplementary Table 6.

Training objective

Let Θ collect all trainable parameters of a model when present. Given an initial condition $z_0^{(k)}$, simulating the corresponding WBM produces a trajectory $\hat{z}_\Theta^{(k)}$. The goal is to find Θ such that the simulated trajectories reproduce the evaluation metrics of the empirical recordings. Formally, we minimize

$$\min_{\Theta} \sum_{k=1}^{N_{\text{exp}}} \mathcal{L}\big(\hat{z}_\Theta^{(k)},\, z^{(k)}\big) + \lambda\, \mathcal{R}(\Theta),$$

where N_{exp} is the number of training recordings, $\mathcal{L}$ is the neuroscience-aligned loss, and $\mathcal{R}(\Theta)$ penalizes model complexity. Individual loss weights are tunable and set per model architecture.

Because FC/phFC and FCD/phFCD neglect the temporal order of the signal, a model can in principle match these metrics while reproducing short-time transitions less faithfully. As a light safeguard we add a finite-dimensional-matching (FDM) regularizer that constrains how the state relates to the state shortly after, as used commonly for fitting of neural networks [55]. For each recording we sample 32 time-point pairs (t_1, t_2) with $1 \leq t_2 - t_1 \leq 50$ steps, form the joint state $\big(\Re z(t_1), \Im z(t_1), \Re z(t_2), \Im z(t_2)\big)$, and penalize the maximum mean discrepancy (MMD) between the empirical and simulated joint-state distributions under a Gaussian radial basis function (RBF) kernel (bandwidth $\sigma\sqrt{d}$, $\sigma = 1$, d the joint-state dimension). In our experiments the FDM term gives a small, consistent improvement in short-time transition fidelity, so we keep it at a low weight.

In sum, our multi-target loss function is described as follows: $w_{\text{FC corr}} = w_{\text{FC MSE}} = w_{\text{FCD}} = w_{\text{phFCD}} = w_{\text{phFC corr}} = w_{\text{Meta}} = w_{\text{amp}} = w_\omega = 1.0$ together with the FDM regularizer $w_{\text{FDM}} = 0.25$ As the temporal-correlation, power-spectrum, and autocorrelation terms are set to zero, these metrics are used as evaluation metrics only.

Model architecture specifications

We consider three model architectures of continuous-time generative models: physics-based, data-driven, and hybrid. We begin with the physics-based Hopf WBM, then introduce a complementary fully data-driven Neural SDE, and finally develop three hybrid architectures that combine the oscillatory prior of the Hopf model with the flexibility of neural networks. Supplementary Table 4 contrasts the three model architectures in terms of their learnable parameters, coupling structure, and parameter complexity.

All three WBMs are formulated as stochastic differential equations (SDEs) with state $z(t) \in \mathbb{C}^n$ collecting the analytic signals of the n ROIs. The stochasticity enters through an n-dimensional complex Brownian motion W.

Coupled Hopf: physics-based baseline.

The supercritical Hopf bifurcation has become the basis of a widely used WBM, as it can generate oscillatory transitions between a noisy equilibrium and self-sustained oscillations —the two dominant dynamical regimes observed in cortical populations [7, 8]. Each node is governed by a local Hopf oscillator f_i^{loc}, extended with a timescale parameter $\kappa \in \mathbb{R}_{>0}$:

$$f_i^{\text{loc}}(z_i) = \big(\kappa a + \mathrm{i}\omega_i - \kappa\,|z_i|^2\big) z_i,$$

where $a \in \mathbb{R}$ is a bifurcation parameter controlling the transition between a stable focus ($a < 0$) and a limit cycle ($a > 0$), and $\omega_i \in \mathbb{R}$ is an intrinsic angular frequency estimated from the peak of the regional power spectrum via the fast Fourier transform (FFT).

By coupling n such oscillators through a fixed connectivity matrix, the *Coupled Hopf* reproduces key features of resting-state fMRI, including functional connectivity, metastability, and frequency-resolved power spectra [11, 12, 32, 45].

Each region i follows the normal-form dynamics

$$\mathrm{d}z_i = \Big[f_i^{\text{loc}}(z_i) + \underbrace{G \sum_{j=1}^{n} C_{ij}\big(z_j - z_i\big)}_{\text{global coupling}}\Big]\,\mathrm{d}t + \sigma\;\mathrm{d}W_i,$$

where $z_i \in \mathbb{C}$ is the complex analytic signal of the i-th region, $G \in \mathbb{R}_{>0}$ is a global coupling strength, $C \in \mathbb{R}^{n\times n}$ is the row-normalized structural connectivity matrix, and $\sigma > 0$ is the noise amplitude. The real part $\Re(z_i)$ serves as a proxy for BOLD amplitude and $\angle z_i$ for the instantaneous phase. The possible learnable parameters are (a, κ, G, σ); intrinsic frequencies ω_i may optionally be learned.

In this paper, the plain Coupled Hopf model serves as the physics-based benchmark against which the stochastic and data-driven extensions are compared.

Neural SDE: fully data-driven.

Neural SDEs extend Neural ODEs [13] to the stochastic setting by parameterizing both the drift and diffusion of an SDE with neural networks [14, 15, 61]. The learned drift governs the deterministic evolution of the state without imposing any oscillatory or physics-based ansatz. And because the diffusion coefficient is learned, the model is inherently *generative* with different sample trajectories, and the learned diffusion controls the variability of the generated dynamics. This is crucial for whole-brain modeling, where trial-to-trial variability is a fundamental feature of neural recordings rather than mere measurement noise.

We choose Neural SDEs as our data-driven baseline because they impose no physics-based assumptions on the dynamics, allowing them to capture complex nonlinear interactions that may be poorly approximated by physics-based WBMs with a small number of parameters. In comparison with the Coupled Hopf model, we expect Neural SDEs to better fit neural activity data, at the potential cost of reduced interpretability and weaker inductive biases for oscillatory dynamics. The model reads:

$$\mathrm{d}z(t) = f_\theta\big(z(t)\big)\;\mathrm{d}t + g_\phi\big(z(t)\big)\;\mathrm{d}W(t),$$

where $f_\theta, g_\phi : \mathbb{C}^n \to \mathbb{C}^n$ are the drift and diffusion networks, respectively. Both operate on a real representation: the complex input is mapped to $\mathbb{R}^{2n}$, processed by a real-valued multilayer perceptron (MLP), and reshaped back to $\mathbb{C}^n$. Specifically, f_θ is a 2-hidden-layer tanh MLP (width 128) and g_ϕ is a 1-hidden-layer tanh MLP (width 64) with Softplus output to ensure non-negative diffusion. All parameters θ and ϕ are learned by gradient-based optimization.

Hybrid Hopf: learned nonlinear coupling.

Our hybrid model architecture replaces the linear diffusive coupling of the Hopf model with a learnable nonlinear transformation. Hybridization can be achieved in multiple ways, and we initially explore three variants—the *Hopf+Neural*, the *Hybrid Hopf*, and the *GNN-Hopf*—compare their performance, and then carry only the best-performing variant forward for further analyses.

The *Hopf+Neural* variant adds a learned neural correction term to the full physics-based Hopf drift, recovering the Coupled Hopf baseline when the correction vanishes; we give its full specification in the Supplementary Material.

We additionally investigate two GNN designs that exploit the connectome structure: the first (Hybrid Hopf) aggregates neighbor information through a learned embedding followed by a nonlinear output network, while the second (GNN-Hopf) uses a standard graph-convolutional layer with learnable connectivity. We detail the Hybrid Hopf below, as it is the variant carried forward; the full GNN-Hopf specification is given in the Supplementary Material.

The Hybrid Hopf is built on the motivation that anatomical connections do not merely transmit a linear difference signal: synaptic transmission involves nonlinear gain, frequency-dependent filtering, and heterogeneous synaptic efficacies that a scalar coupling strength G cannot capture, and recent generative fits suggest the effective inter-regional interactions depart from the structural connectome [42, 43]. By introducing a nonlinear map ψ on the coupling term, we allow the model to learn richer inter-regional interactions while preserving the Hopf local dynamics as an inductive bias. The model reads:

$$m_i = \sum_{j=1}^{n} C_{ij}\big(k_j - k_i\big), \quad \text{where } k_j := \varphi_{\text{key}}(z_j)$$

$$\mathrm{d}z_i = \Big[f_i^{\text{loc}}(z_i) + G\,\psi_{\text{out}}(m_i, z_i)\Big]\,\mathrm{d}t + \sigma\;\mathrm{d}W_i,$$

where $\varphi_{\text{key}} : \mathbb{C} \to \mathbb{C}^{d_k}$ is a per-node key encoder (1-hidden-layer complex MLP, default $d_k = 4$) and $\psi_{\text{out}} : \mathbb{C}^{d_k+1} \to \mathbb{C}$ is a per-node output network (1-hidden-layer complex MLP, near-zero initialization). The coupling matrix is optionally low-rank: $C \approx LR^\top$ with $L, R \in \mathbb{R}^{n\times r}$, reducing the quadratic aggregation cost to linear.

Data availability

The resting-state fMRI data used in this study are publicly available from the Human Connectome Project (HCP) at https://www.humanconnectome.org/study/hcp-young-adult (Young Adult dataset). Access requires completion of the HCP Open

Access Data Use Terms registration. Preprocessed time series files necessary to reproduce the reported results will be made available from the corresponding authors upon reasonable request.

Code availability

All code developed in this study is publicly available as an open-source repository at https://github.com/DCN-FAU-AvH/neuroscience-modeling. The repository includes implementations of all three model architectures (Coupled Hopf, Neural SDE, and Hybrid Hopf), the neuroscience-aligned training objective, evaluation scripts, and instructions for reproducing all reported results.

Acknowledgements

E. Zuazua was partially supported by the European Research Council (ERC) under the European Union's Horizon Europe research and innovation programme (grant agreement No. 101096251-CoDeFeL); by the Alexander von Humboldt Professorship programme; the European Union's Horizon Europe MSCA project ModConFlex (HORIZON-MSCA-2021-DN-01, project 101073558); the Transregio 154 Project "Mathematical Modelling, Simulation and Optimization Using the Example of Gas Networks" of the DFG; the AFOSR 24IOE027 project; the SURE-AI Norwegian Centre for Sustainable, Risk-Averse, and Ethical AI (grant 357482, Research Council of Norway); the Grant PID2023-146872OB-I00-DyCMaMod of MICIU (Spain); and the COST Actions CA24122 – Multiscale Stochastics, Patterns, and Analysis of Combinatorial Environments and CA24136 – Interactions between Control Theory and Machine Learning.

Author contributions

D.L.-M.: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing – original draft. L.L.: Conceptualization, Methodology, Investigation, Writing – review & editing. E.Z.: Conceptualization, Funding acquisition, Project administration, Supervision, Writing – review & editing. X.K.: Conceptualization, Data curation, Resources, Domain expertise, Project administration, Supervision, Writing – review & editing.

Competing interests

The authors declare no competing interests.

Ethics declarations

The Human Connectome Project resting-state fMRI data used in this work were collected under protocols approved by the Washington University Institutional Review Board. All participants provided written informed consent. No new data involving human subjects were collected for this study.

References

[1] Kobeleva, X., Varoquaux, G., Dagher, A., Adhikari, M.H., Grefkes, C., Gilson, M.: Advancing brain network models to reconcile functional neuroimaging and clinical research. NeuroImage: Clinical **36**, 103262 (2022) https://doi.org/10.1016/j.nicl.2022.103262

[2] Jung, K., Florin, E., Patil, K.R., Caspers, J., Rubbert, C., Eickhoff, S.B., Popovych, O.V.: Whole-brain dynamical modelling for classification of Parkinson's disease. Brain Communications **5**(1), 331 (2022) https://doi.org/10.1093/braincomms/fcac331

[3] Little, S., Pogosyan, A., Neal, S., Zavala, B., Zrinzo, L., Hariz, M., Foltynie, T., Limousin, P., Ashkan, K., FitzGerald, J., Green, A.L., Aziz, T.Z., Brown, P.: Adaptive deep brain stimulation in advanced Parkinson disease. Annals of Neurology **74**(3), 449–457 (2013) https://doi.org/10.1002/ana.23951

[4] Bikson, M., Grossman, P., Thomas, C., Zannou, A.L., Jiang, J., Adnan, T., Mourdoukoutas, A.P., Kronberg, G., Truong, D., Boggio, P., Brunoni, A.R., Charvet, L., Fregni, F., Fritsch, B., Gillick, B., Hamilton, R.H., Hampstead, B.M., Jankord, R., Kirton, A., Knotkova, H., Liebetanz, D., Liu, A., Loo, C., Nitsche, M.A., Reis, J., Richardson, J.D., Rotenberg, A., Turkeltaub, P.E., Woods, A.J.: Safety of Transcranial Direct Current Stimulation: Evidence Based Update 2016. Brain Stimulation **9**(5), 641–661 (2016) https://doi.org/10.1016/j.brs.2016.06.004

[5] Stuart, T., Hanna, J., Gutruf, P.: Wearable devices for continuous monitoring of biosignals: Challenges and opportunities. APL Bioengineering **6**(2), 021502 (2022) https://doi.org/10.1063/5.0086935

[6] Wang, H.E., Dollomaja, B., Triebkorn, P., Duma, G.M., Williamson, A., Makhalova, J., Lemarechal, J.-D., Bartolomei, F., Jirsa, V.: Virtual brain twins for stimulation in epilepsy. Nature Computational Science **5**(9), 754–768 (2025) https://doi.org/10.1038/s43588-025-00841-6

[7] Deco, G., Jirsa, V., Robinson, P., Breakspear, M., Friston, K.: The Dynamic Brain: From Spiking Neurons to Neural Masses and Cortical Fields. PLoS computational biology **4**, 1000092 (2008) https://doi.org/10.1371/journal.pcbi.1000092

[8] Deco, G., Jirsa, V.K., McIntosh, A.R.: Emerging concepts for the dynamical organization of resting-state activity in the brain. Nature Reviews Neuroscience **12**(1), 43–56 (2011) https://doi.org/10.1038/nrn2961

[9] Beurle, R.L.: Properties of a mass of cells capable of regenerating pulses. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences **240**(669), 55–94 (1956) https://doi.org/10.1098/rstb.1956.0012

[10] Wilson, H.R., Cowan, J.D.: Excitatory and Inhibitory Interactions in Localized

Populations of Model Neurons. Biophysical Journal **12**(1), 1–24 (1972) https://doi.org/10.1016/S0006-3495(72)86068-5

[11] Deco, G., Kringelbach, M.L., Jirsa, V.K., Ritter, P.: The dynamics of resting fluctuations in the brain: Metastability and its dynamical cortical core. Scientific Reports **7**(1), 3095 (2017) https://doi.org/10.1038/s41598-017-03073-5

[12] Ponce-Alvarez, A., Deco, G.: The Hopf whole-brain model and its linear approximation. Scientific Reports **14**(1), 2615 (2024) https://doi.org/10.1038/s41598-024-53105-0

[13] Chen, R.T.Q., Rubanova, Y., Bettencourt, J., Duvenaud, D.: Neural Ordinary Differential Equations. arXiv (2019). https://doi.org/10.48550/arXiv.1806.07366

[14] Kidger, P., Foster, J., Li, X., Lyons, T.J.: Neural SDEs as Infinite-Dimensional GANs. In: Proceedings of the 38th International Conference on Machine Learning, pp. 5453–5463. PMLR, ??? (2021)

[15] Li, X., Wong, T.-K.L., Chen, R.T.Q., Duvenaud, D.: Scalable Gradients for Stochastic Differential Equations. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, pp. 3870–3882. PMLR, ??? (2020)

[16] Han, K., Yang, Y., Huang, Z., Kan, X., Yang, Y., Guo, Y., He, L., Zhan, L., Sun, Y., Wang, W., Yang, C.: BrainODE: Dynamic Brain Signal Analysis via Graph-Aided Neural Ordinary Differential Equations. https://arxiv.org/abs/2405.00077v1 (2024)

[17] Raissi, M., Perdikaris, P., Karniadakis, G.E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics **378**, 686–707 (2019) https://doi.org/10.1016/j.jcp.2018.10.045

[18] Karniadakis, G.E., Kevrekidis, I.G., Lu, L., Perdikaris, P., Wang, S., Yang, L.: Physics-informed machine learning. Nature Reviews Physics **3**(6), 422–440 (2021) https://doi.org/10.1038/s42254-021-00314-5

[19] Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A., Edelman, A.: Universal Differential Equations for Scientific Machine Learning. https://arxiv.org/abs/2001.04385v4 (2020)

[20] Willard, J., Jia, X., Xu, S., Steinbach, M., Kumar, V.: Integrating Scientific Knowledge with Machine Learning for Engineering and Environmental Systems. https://arxiv.org/abs/2003.04919v6 (2020)

[21] Liverani, L., Steynberg, M., Zuazua, E.: HYCO: Hybrid-Cooperative Learning for Data-Driven PDE Modeling. arXiv (2025). https://doi.org/10.48550/arXiv.2509.

14123

[22] Merkelbach, K., Schweidtmann, A.M., Müller, Y., Schwoebel, P., Mhamdi, A., Mitsos, A., Schuppert, A., Mrziglod, T., Schneckener, S.: HybridML: Open source platform for hybrid modeling. Computers & Chemical Engineering **160**, 107736 (2022) https://doi.org/10.1016/j.compchemeng.2022.107736

[23] Procopio, A., Cesarelli, G., Donisi, L., Merola, A., Amato, F., Cosentino, C.: Combined mechanistic modeling and machine-learning approaches in systems biology - A systematic literature review. Computer Methods and Programs in Biomedicine **240**, 107681 (2023) https://doi.org/10.1016/j.cmpb.2023.107681

[24] Hansen, E.C.A., Battaglia, D., Spiegler, A., Deco, G., Jirsa, V.K.: Functional connectivity dynamics: Modeling the switching behavior of the resting state. NeuroImage **105**, 525–535 (2015) https://doi.org/10.1016/j.neuroimage.2014.11.001

[25] Preti, M.G., Bolton, T.A., Van De Ville, D.: The dynamic functional connectome: State-of-the-art and perspectives. NeuroImage **160**, 41–54 (2017) https://doi.org/10.1016/j.neuroimage.2016.12.061

[26] Elbaz, D., Zibulevsky, M.: Perceptual Audio Loss Function for Deep Learning. arXiv (2017). https://doi.org/10.48550/ARXIV.1708.05987

[27] Martin-Donas, J.M., Gomez, A.M., Gonzalez, J.A., Peinado, A.M.: A Deep Learning Loss Function Based on the Perceptual Evaluation of the Speech Quality. IEEE Signal Processing Letters **25**(11), 1680–1684 (2018) https://doi.org/10.1109/LSP.2018.2871419

[28] Fu, S.-W., Liao, C.-F., Tsao, Y.: Learning With Learned Loss Function: Speech Enhancement With Quality-Net to Improve Perceptual Evaluation of Speech Quality. IEEE Signal Processing Letters **27**, 26–30 (2020) https://doi.org/10.1109/LSP.2019.2953810

[29] Hornik, K., Stinchcombe, M., White, H.: Multilayer feedforward networks are universal approximators. Neural Networks **2**(5), 359–366 (1989) https://doi.org/10.1016/0893-6080(89)90020-8

[30] ElGazzar, A., Gerven, M.: Generative Modeling of Neural Dynamics via Latent Stochastic Differential Equations. arXiv (2026). https://doi.org/10.48550/arXiv.2412.12112

[31] Rockne, R.C., Andersen, M., Anderson, A.R.A., Basanta, D., Bentivegna, A., Benzekry, S., Branciamore, S., Brüningk, S.C., Conte, M., Farahpour, F., Karolak, A., Köhn-Luque, A., Lorenzo, G., Manookian, B., Rodin, A.S., Schmalenstroer, L., Soler, J., Tomasetti, C., Urbaniak, K.: The future of mathematical oncology in the age of AI. npj Systems Biology and Applications **12**(1), 22 (2026) https:

//doi.org/10.1038/s41540-026-00656-9

[32] Lombardi, F., Pepić, S., Shriki, O., Tkačik, G., De Martino, D.: Statistical modeling of adaptive neural networks explains co-existence of avalanches and oscillations in resting human brain. Nature Computational Science **3**(3), 254–263 (2023) https://doi.org/10.1038/s43588-023-00410-9

[33] Sanz Leon, P., Knock, S.A., Woodman, M.M., Domide, L., Mersmann, J., McIntosh, A.R., Jirsa, V.: The Virtual Brain: A simulator of primate brain network dynamics. Frontiers in Neuroinformatics **7** (2013) https://doi.org/10.3389/fninf.2013.00010

[34] Sanz-Leon, P., Knock, S.A., Spiegler, A., Jirsa, V.K.: Mathematical framework for large-scale brain network modeling in The Virtual Brain. NeuroImage **111**, 385–430 (2015) https://doi.org/10.1016/j.neuroimage.2015.01.002

[35] West, T.O., Berthouze, L., Farmer, S.F., Cagnan, H., Litvak, V.: Inference of brain networks with approximate Bayesian computation - assessing face validity with an example application in Parkinsonism. NeuroImage **236**, 118020 (2021) https://doi.org/10.1016/j.neuroimage.2021.118020

[36] Ziaeemehr, A., Woodman, M., Domide, L., Petkoski, S., Jirsa, V., Hashemi, M.: Virtual Brain Inference (VBI), a flexible and integrative toolkit for efficient probabilistic inference on whole-brain models. eLife **14**, 106194 (2025) https://doi.org/10.7554/eLife.106194.4

[37] Boelts, J., Lueckmann, J.-M., Gao, R., Macke, J.H.: Flexible and efficient simulation-based inference for models of decision-making. eLife **11**, 77220 (2022) https://doi.org/10.7554/eLife.77220

[38] Wang, Y.C., Rudi, J., Velasco, J., Sinha, N., Idumah, G., Powers, R.K., Heckman, C.J., Chardon, M.K.: Multimodal parameter spaces of a complex multi-channel neuron model. Frontiers in Systems Neuroscience **16**, 999531 (2022) https://doi.org/10.3389/fnsys.2022.999531

[39] Wischnewski, K.J., Jarre, F., Eickhoff, S.B., Popovych, O.V.: Exploring dynamical whole-brain models in high-dimensional parameter spaces. PLOS One **20**(5), 0322983 (2025) https://doi.org/10.1371/journal.pone.0322983

[40] Leone, R., Geysen, S., Deco, G., Kobeleva, X., Alzheimer's Disease Neuroimaging Initiative: Beyond Focal Lesions: Dynamical Network Effects of White Matter Hyperintensities. Human Brain Mapping **45**(17), 70081 (2024) https://doi.org/10.1002/hbm.70081

[41] Perl, Y.S., Pallavicini, C., Piccinini, J., Demertzi, A., Bonhomme, V., Martial, C., Panda, R., Alnagger, N., Annen, J., Gosseries, O., Ibañez, A., Laufs, H., Sitt, J.D., Jirsa, V.K., Kringelbach, M.L., Laureys, S., Deco, G., Tagliazucchi, E.: Low-dimensional organization of global brain states of reduced consciousness. Cell

Reports **42**(5), 112491 (2023) https://doi.org/10.1016/j.celrep.2023.112491

[42] Gilson, M., Moreno-Bote, R., Ponce-Alvarez, A., Ritter, P., Deco, G.: Estimation of Directed Effective Connectivity from fMRI Functional Connectivity Hints at Asymmetries of Cortical Connectome. PLOS Computational Biology **12**(3), 1004762 (2016) https://doi.org/10.1371/journal.pcbi.1004762

[43] Luppi, A.I., Sanz Perl, Y., Vohryzek, J., Ali, H., Mediano, P.A.M., Rosas, F.E., Milisav, F., Suárez, L.E., Gini, S., Gutierrez-Barragan, D., Yee, Y., Froudist-Walsh, S., Gozzi, A., Misic, B., Deco, G., Kringelbach, M.L.: Competitive interactions shape mammalian brain network dynamics and computation. Nature Neuroscience **29**(4), 915–933 (2026) https://doi.org/10.1038/s41593-026-02205-3

[44] Smith, S.M., Beckmann, C.F., Andersson, J., Auerbach, E.J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D.A., Griffanti, L., Harms, M.P., Kelly, M., Laumann, T., Miller, K.L., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A.Z., Vu, A.T., Woolrich, M.W., Xu, J., Yacoub, E., Uğurbil, K., Van Essen, D.C., Glasser, M.F.: Resting-state fMRI in the Human Connectome Project. NeuroImage **80**, 144–168 (2013) https://doi.org/10.1016/j.neuroimage.2013.05.039

[45] Kobeleva, X., López-González, A., Kringelbach, M.L., Deco, G.: Revealing the Relevant Spatiotemporal Scale Underlying Whole-Brain Dynamics. Frontiers in Neuroscience **15**, 715861 (2021) https://doi.org/10.3389/fnins.2021.715861

[46] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020) https://doi.org/10.1016/j.inffus.2019.12.012

[47] Jirsa, V., Wang, H., Triebkorn, P., Hashemi, M., Jha, J., Gonzalez-Martinez, J., Guye, M., Makhalova, J., Bartolomei, F.: Personalised virtual brain models in epilepsy. The Lancet Neurology **22**(5), 443–454 (2023) https://doi.org/10.1016/S1474-4422(23)00008-X

[48] Rudolph, M., Kurz, S., Rakitsch, B.: Hybrid modeling design patterns. Journal of Mathematics in Industry **14**(1), 3 (2024) https://doi.org/10.1186/s13362-024-00141-0

[49] Rackauckas, C., Ma, Y., Martensen, J., Warner, C., Zubov, K., Supekar, R., Skinner, D., Ramadhan, A., Edelman, A.: Universal Differential Equations for Scientific Machine Learning. arXiv (2021). https://doi.org/10.48550/arXiv.2001.04385

[50] Kipf, T.N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. arXiv (2017). https://doi.org/10.48550/arXiv.1609.02907

[51] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., Bengio, Y.: Graph Attention Networks. arXiv (2018). https://doi.org/10.48550/arXiv.1710.10903

[52] Deco, G., Cruzat, J., Cabral, J., Tagliazucchi, E., Laufs, H., Logothetis, N.K., Kringelbach, M.L.: Awakening: Predicting external stimulation to force transitions between different brain states. Proceedings of the National Academy of Sciences **116**(36), 18088–18097 (2019) https://doi.org/10.1073/pnas.1905534116

[53] Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.T.: Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral Cortex **28**(9), 3095–3114 (2018) https://doi.org/10.1093/cercor/bhx179

[54] Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A.: Neurophysiological investigation of the basis of the fMRI signal. Nature **412**(6843), 150–157 (2001) https://doi.org/10.1038/35084005

[55] Zhang, J., Viktorov, J., Jung, D., Pitler, E.: Efficient Training of Neural Stochastic Differential Equations by Matching Finite Dimensional Distributions. arXiv (2025). https://doi.org/10.48550/arXiv.2410.03973

[56] Massey, F.J.: The Kolmogorov-Smirnov Test for Goodness of Fit. Journal of the American Statistical Association **46**(253), 68–78 (1951) https://doi.org/10.1080/01621459.1951.10500769

[57] Smirnov, N.: Table for Estimating the Goodness of Fit of Empirical Distributions. The Annals of Mathematical Statistics **19**(2), 279–281 (1948) https://doi.org/10.1214/aoms/1177730256

[58] Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S.: Functional connectivity in the motor cortex of resting human brain using echo-planar mri. Magnetic Resonance in Medicine **34**(4), 537–541 (1995) https://doi.org/10.1002/mrm.1910340409

[59] Glerean, E., Salmi, J., Lahnakoski, J.M., Jääskeläinen, I.P., Sams, M.: Functional Magnetic Resonance Imaging Phase Synchronization as a Measure of Dynamic Functional Connectivity. Brain Connectivity **2**(2), 91–101 (2012) https://doi.org/10.1089/brain.2011.0068

[60] Shanahan, M.: Metastable chimera states in community-structured oscillator networks. Chaos: An Interdisciplinary Journal of Nonlinear Science **20**(1), 013108 (2010) https://doi.org/10.1063/1.3305451

[61] Liu, X., Xiao, T., Si, S., Cao, Q., Kumar, S., Hsieh, C.-J.: Neural SDE: Stabilizing Neural ODE Networks with Stochastic Noise. arXiv (2019). https://doi.org/10.48550/arXiv.1906.02355

Supplementary Material

Mathematical relation between surrogate measures and distribution statistics (FCD MSE and KS)

Let $a, b \in \mathbb{R}^N$ denote the strict upper-triangular entries of the simulated and empirical FCD, respectively, on the same window grid (the same argument applies to phFCD), with empirical measures $\mu_a = N^{-1}\sum_i \delta_{a_i}$, $\mu_b = N^{-1}\sum_i \delta_{b_i}$ and CDFs F_a, F_b. The training surrogate and the Kolmogorov–Smirnov (KS) evaluation metric are

$$L_{\mathrm{MSE}}(a,b) := \frac{1}{N}\sum_{i=1}^{N}(a_i - b_i)^2, \qquad D_{\mathrm{KS}}(\mu_a,\mu_b) := \sup_{x\in\mathbb{R}}\big|F_a(x) - F_b(x)\big|,$$

and we measure the local concentration of μ_b at scale $r > 0$ by $Q_b(r) := \sup_{x\in\mathbb{R}} \mu_b\big((x, x+r]\big)$.

Proposition. For every $r > 0$, $D_{\mathrm{KS}}(\mu_a,\mu_b) \le L_{\mathrm{MSE}}(a,b)/r^2 + Q_b(r)$.

Proof Let $I \sim \mathrm{Unif}\{1,\dots,N\}$ and set $X := a_I \sim \mu_a$, $Y := b_I \sim \mu_b$, so $\mathbb{E}|X-Y|^2 = L_{\mathrm{MSE}}(a,b)$. Markov's inequality applied to $|X-Y|^2$ gives

$$\mathbb{P}(|X-Y| > r) = \mathbb{P}\big(|X-Y|^2 > r^2\big) \le \mathbb{E}|X-Y|^2/r^2 = L_{\mathrm{MSE}}(a,b)/r^2. \tag{3}$$

Fix $x \in \mathbb{R}$. Splitting both $\{X \le x\}$ and $\{Y \le x\}$ on $\{X \le x, Y \le x\}$ yields $F_a(x) - F_b(x) = \mathbb{P}(X \le x < Y) - \mathbb{P}(Y \le x < X)$, and since both terms are non-negative,

$$|F_a(x) - F_b(x)| \le \max\big\{\mathbb{P}(X \le x < Y),\ \mathbb{P}(Y \le x < X)\big\}. \tag{4}$$

If $X \le x < Y$, then either $Y - X > r$ (so $|X-Y| > r$) or $Y - X \le r$, in which case $X \le x$ gives $Y \le x + r$, hence $Y \in (x, x+r]$; symmetrically, if $Y \le x < X$ and $X - Y \le r$, then $X > x$ gives $Y > x - r$, hence $Y \in (x-r, x]$. Thus

$$\{X \le x < Y\} \subseteq \{|X-Y| > r\} \cup \{Y \in (x, x+r]\},$$
$$\{Y \le x < X\} \subseteq \{|X-Y| > r\} \cup \{Y \in (x-r, x]\}.$$

and combining with (3) bounds each of $\mathbb{P}(X \le x < Y)$ and $\mathbb{P}(Y \le x < X)$ by $L_{\mathrm{MSE}}(a,b)/r^2 + Q_b(r)$. Substituting into (4) and taking the supremum over x proves the claim. □

Corollary. If $Q_b(r) \le \rho r$ for some $\rho > 0$ on the relevant range (equivalently, F_b is ρ-Lipschitz at scale r) and $L := L_{\mathrm{MSE}}(a,b) > 0$, then minimizing $g(r) := L/r^2 + \rho r$ on $(0,\infty)$ via $g'(r) = -2L/r^3 + \rho = 0$, $g'' > 0$, gives $r_\star = (2L/\rho)^{1/3}$ and

$$D_{\mathrm{KS}}(\mu_a,\mu_b) \le g(r_\star) = \underbrace{2^{-2/3}\rho^{2/3}L^{1/3}}_{L/r_\star^2} + \underbrace{2^{1/3}\rho^{2/3}L^{1/3}}_{\rho r_\star} = 3\cdot 2^{-2/3}\,\rho^{2/3}\,L_{\mathrm{MSE}}(a,b)^{1/3}.$$

The case $L_{\mathrm{MSE}}(a,b) = 0$ is trivial since then $\mu_a = \mu_b$. Thus small entrywise FCD/phFCD MSE controls the KS distance under a mild anti-concentration condition. The two metrics are not equivalent: KS ignores the ordering of matrix entries, whereas the MSE surrogate also enforces the window-pair correspondence.

Full model comparison table

Table 3 reports all five model variants across all evaluation metrics on the HCP resting-state dataset (94 subjects; 100 regions of interest, ROIs). The main text (Table 1) shows a subset of three architectures and metrics for clarity.

Table 3 Full comparison of all model architectures on the Human Connectome Project (HCP) resting-state dataset. Functional connectivity (FC) correlation; FC mean-squared error (MSE); phase functional connectivity (phFC) correlation; functional connectivity dynamics (FCD) Kolmogorov–Smirnov (KS) distance; phase FCD (phFCD) KS distance; metastability absolute error (Meta $|\Delta|$); time-series (TS) correlation; autocorrelation. Arrows indicate the direction of improvement (↑ higher is better, ↓ lower is better). Bold entries denote the best value per column among models fitted by gradient-based optimization.

Model	FC corr ↑	FC MSE ↓	phFC corr ↑	FCD KS ↓	phFCD KS ↓	Meta $\|\Delta\|$ ↓	TS corr ↑	Autocorr ↓
Coupled Hopf	0.911 ± 0.001	**0.018 ± 0.000**	0.783 ± 0.003	0.235 ± 0.006	0.254 ± 0.012	$9.5 \times 10^{-3} \pm 2.7{\times}10^{-3}$	0.097 ± 0.003	**0.148 ± 0.004**
Neural SDE	0.915 ± 0.005	0.020 ± 0.005	0.911 ± 0.001	**0.049 ± 0.006**	0.305 ± 0.001	$6.7 \times 10^{-3} \pm \mathbf{1.3}{\times}10^{-3}$	0.027 ± 0.013	0.158 ± 0.000
Hybrid Hopf	0.916 ± 0.005	0.023 ± 0.001	0.915 ± 0.005	0.082 ± 0.001	**0.052 ± 0.008**	0.013 ± 0.002	0.058 ± 0.008	0.262 ± 0.007
GNN-Hopf	**0.926 ± 0.009**	0.029 ± 0.006	**0.923 ± 0.010**	0.133 ± 0.002	0.126 ± 0.017	0.031 ± 0.000	0.071 ± 0.000	0.308 ± 0.004
Hopf+Neural	0.839 ± 0.014	0.021 ± 0.003	0.649 ± 0.013	0.223 ± 0.003	0.195 ± 0.006	0.018 ± 0.002	0.071 ± 0.003	0.180 ± 0.001

Trajectory metrics

Beyond the network-structure and network-dynamics metrics reported in the main text, we evaluate *trajectory-level* metrics that quantify point-by-point fidelity of the simulated time series rather than time-averaged or dynamic connectivity. Temporal (time-series) correlation measures the agreement between empirical and simulated regional time courses, while autocorrelation captures how well each region's intrinsic temporal structure is reproduced. The Coupled Hopf retains the strongest trajectory-level metrics, in particular temporal correlation and autocorrelation (Table 3), reflecting that its explicit physics-based local drift is a strong inductive bias for reproducing point-by-point time-series structure.

Model architecture overview

Supplementary Table 4 contrasts the three model architectures introduced in the Methods (Model specifications) in terms of their learnable parameters, coupling structure, and asymptotic parameter complexity.

Table 4 Overview of model architecture properties. Param. complexity refers to the asymptotic number of learnable parameters; h denotes the relevant multilayer perceptron (MLP) hidden width and r is the coupling low-rank dimension.

Model	Learnable params	Coupling	Param. complexity
Coupled Hopf	a, κ, G, σ	fixed C	$O(1)$
Neural SDE	θ, ϕ	none (fully learned)	$O(n\,h)$
Hybrid Hopf	$a, \kappa, G, \sigma, L, R, \varphi, \psi$	low-rank $LR^\top$	$O(nr)$

Additional hybrid model variants

Beyond the Hybrid Hopf carried forward in the main text, we explored two further hybrid variants. Both are specified here for completeness.

Hopf+Neural: physics-based base with learned residual.

An alternative hybridization strategy is to keep the full physics-based Hopf dynamics—including its linear coupling—as a *base model*, and add a Neural SDE correction term that learns to compensate for systematic errors. This approach is motivated by the observation that the Coupled Hopf model captures the coarse oscillatory structure well but consistently misses fine-grained Pearson-FC correlations. Rather than modifying the coupling nonlinearity (as in the Hybrid Hopf), we let a neural network learn an additive correction over the entire drift, which can capture arbitrary missing dynamics—including higher-order interactions, non-diffusive coupling, and region-heterogeneous effects. The model reads:

$$\mathrm{d}z_i = \Big[f_i^{\mathrm{loc}}(z_i) + G \sum_j C_{ij}(z_j - z_i) + \underbrace{f_\theta(z)_i}_{\substack{\text{neural} \\ \text{correction}}} \Big]\, \mathrm{d}t + \sigma\, \mathrm{d}W_i,$$

where $f_\theta : \mathbb{C}^n \to \mathbb{C}^n$ is a 2-hidden-layer tanh MLP (width 64) acting on the state z, with near-zero output initialization, similar in architecture to the Neural SDE drift but with smaller capacity. The model starts in the pure-Hopf regime ($f_\theta \approx 0$) and learns corrections where the physics-based model falls short.

GNN-Hopf: graph neural network coupling.

The GNN-Hopf variant replaces the coupling function with a graph neural network (GNN) message-passing layer [50]. While the Hybrid Hopf applies a pointwise nonlinearity to each node's aggregated message, the GNN-Hopf processes the message and the node state jointly through a complex-valued MLP ψ_θ, allowing the model to learn message transformations that depend on the receiving node's current state. Additionally, the coupling matrix C itself is made learnable (initialized from the structural connectome), so the model can adapt the effective connectivity structure beyond what the structural connectome prescribes. This is motivated by the well-documented discrepancy between structural and functional connectivity: the effective interactions between brain regions during ongoing dynamics are shaped by neuromodulation, synaptic plasticity, and polysynaptic pathways that are not directly reflected in the diffusion-MRI tractogram [8]. The model reads:

$$\begin{aligned} m_i &= \sum_{j=1}^{n} C_{ij}\big(z_j - z_i\big), \\ \mathrm{d}z_i &= \Big[f_i^{\mathrm{loc}}(z_i) + G\, \psi_\theta(m_i, z_i) \Big]\, \mathrm{d}t + \sigma\, \mathrm{d}W_i, \end{aligned}$$

where C is a learnable coupling matrix (initialized from the structural connectome) and $\psi_\theta : \mathbb{C}^2 \to \mathbb{C}$ is a complex-valued node-wise MLP (2 hidden layers, width 32,

phase-preserving activations). The activation $\sigma_{\mathrm{pp}}(z) = \frac{z}{|z|}\tanh(|z|)$ preserves phase while bounding magnitude. The output layer is initialized near zero so the model starts in the pure-Hopf regime.

Complexity measures

We additionally probe whether the simulated dynamics reproduce the *information content* of empirical connectivity-dynamics matrices through two evaluation-only diagnostics. Both operate on the strict upper triangle of a symmetric similarity matrix $M \in \{\mathrm{FCD}, \mathrm{phFCD}\}$ obtained from either an empirical recording or a simulated trajectory. The matrix size $m \times m$ is determined by the number of FCD windows or phFCD time points. For each model we draw $N_{\mathrm{trial}} = 10$ trials, where each trial samples 2 held-out subjects, simulates a fresh trajectory of 120 time points per subject under independent noise, and recomputes all matrices.

Shannon entropy.

The Shannon entropy of M measures how broadly its entries are distributed: a low value means the matrix is dominated by a few correlation regimes, while a high value indicates a richer mixture. Let $v \in \mathbb{R}^q$, $q = m(m-1)/2$, denote the finite upper-triangular entries of M. We discretize v into a histogram of $B = \min(32, \max(4, \lfloor q/4 \rfloor))$ equal-width bins spanning $[\min v, \max v]$, normalize the bin counts to a probability vector $p \in \Delta^B$, and define

$$H(M) := -\sum_{b:\, p_b > 0} p_b \log_2 p_b \quad \text{(bits)}.$$

Lempel–Ziv–Welch (LZW) complexity.

LZW complexity captures structural redundancy in the same upper-triangular sequence: it asks how many distinct patterns are needed to losslessly describe v. We first quantize v into $B' = 4$ symbols using equal-frequency (quantile) bins, so that each symbol is, by construction, equally likely under the empirical marginal and any deviation from a uniform i.i.d. source is attributable to higher-order structure. Running the standard LZW dictionary encoder on the resulting symbol stream $s \in \{0, \ldots, B'-1\}^q$ produces $c(s)$ output codes; we report the normalized complexity

$$\mathrm{LZW}(M) := \frac{c(s)\,\log_2 q}{q\,\log_2 b},$$

where $b \le B'$ is the number of symbols actually used. The denominator is the Lempel–Ziv asymptotic bound for an i.i.d. source over b symbols, so $\mathrm{LZW}(M) \to 1$ for maximally unstructured sequences and decreases as the matrix becomes more compressible. These diagnostics are markers of consciousness, arousal, and pathology in the clinical literature [45, 52].

Higher-order modular structure and signal complexity

To complement the main-text modularity analysis (Figure 4c), which is shown at $K = 2$, we report two additional descriptors of brain-network activity in Supplementary Figure 5.

First, we repeated the spectral clustering of the empirical FC at a higher partition resolution ($K = 3$) and reordered each simulated FC with the resulting partition (Supplementary Figure 5a). The Hybrid Hopf and Neural SDE again recover the finer block-diagonal community structure, while the Coupled Hopf yields a flatter reordering with weaker inter-community contrast, consistent with the $K = 2$ result in the main text.

Second, we probed the complexity of the simulated dynamics by computing the Shannon entropy and LZW complexity of the FCD and phFCD matrices (Supplementary Figure 5b). For both Shannon entropy and LZW complexity, the Hybrid Hopf matches the empirical complexity across both FCD and phFCD. The Coupled Hopf systematically underestimates phFCD entropy and overshoots phFCD LZW, while the Neural SDE matches FCD complexity well but underestimates the phFCD complexity.

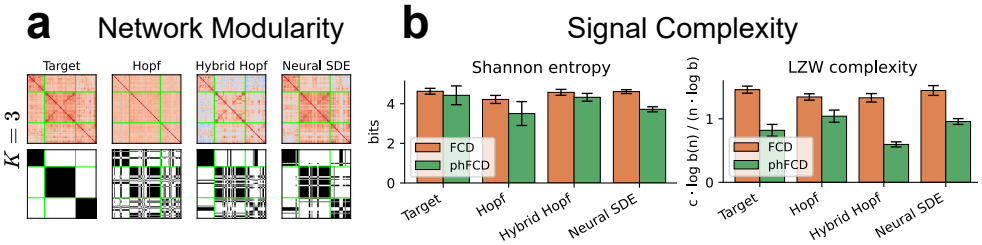

Fig. 5 Higher-order modular structure and signal complexity. (a) Network modularity at $K = 3$: brain clustering analysis over the empirical functional connectivity (FC) matrix and reordering of the simulated FCs using the obtained classes, complementing the $K = 2$ analysis in Figure 4c. **(b)** Signal complexity analysis showing the Shannon entropy and Lempel–Ziv–Welch (LZW) complexity of the functional connectivity dynamics (FCD) and phase FCD (phFCD) matrices for the empirical target and the three model architectures.

Coupled Hopf: grid search vs. gradient-based optimization

Supplementary Table 5 reports the empirical comparison between grid search and gradient-based optimization of the Coupled Hopf model on the HCP resting-state dataset, summarized in the Results (Figure 2c).

Computational cost

Table 6 reports the wall-clock cost of model fitting and inference for all five model architectures on the HCP resting-state dataset. Fitting cost is the mean ± standard deviation of the time for one fitting epoch — 1,024 windows drawn on the fly and processed in 8 mini-batches of 128, with forward SDE integration and backpropagation — averaged over four measured epochs after two warm-up epochs. Inference cost is the

Table 5 Coupled Hopf model: grid search vs. gradient-based optimization. Functional connectivity (FC) correlation; phase FC (phFC) correlation; functional connectivity dynamics (FCD) Kolmogorov–Smirnov (KS) distance; phase FCD (phFCD) KS distance; metastability absolute error (Meta $|\Delta|$); time-series (TS) correlation. Arrows: ↑ higher is better, ↓ lower is better. Bold and * indicate the better value per column.

Method	FC corr ↑	phFC corr ↑	FCD KS ↓	phFCD KS ↓	Meta $\lvert\Delta\rvert$ ↓	TS corr ↑
Grid search	**0.921 ± 0.014**	**0.812*** **± 0.033**	0.250 ± 0.015	0.406 ± 0.169	0.030 ± 0.025	**0.104 ± 0.007**
Gradient	0.912 ± 0.012	0.776 ± 0.018	**0.240*** **± 0.009**	**0.181*** **± 0.051**	**0.01*** $\pm 4.8\times10^{-3}$	0.102 ± 0.006

mean ± standard deviation over 20 forward-only rollouts (no gradient tracking) of one mini-batch of 128 windows, each simulated for 100 time points (72 s at repetition time TR = 0.72 s). Peak GPU memory is the maximum device memory allocated during a fitting epoch. All measurements were taken on a single NVIDIA GeForce RTX 2080 Ti GPU (11 GB) with an Intel Xeon Gold 6134 CPU, running PyTorch 2.10.0 with CUDA 12.8 under Python 3.13. A full reproduction of all reported experiments requires approximately 20.7 GPU-hours on a single consumer GPU.

Table 6 Computational cost of each model architecture. Inference time is for one forward of a mini-batch of 128 windows simulated for 100 time points.

Model	Learnable params	Fitting (s/epoch)	Inference (s)	GPU mem (GB)
Coupled Hopf	2	24.91 ± 0.04	2.24 ± 0.03	4.51
Neural SDE	220,176	28.18 ± 0.05	2.29 ± 0.04	4.77
Hybrid Hopf	2,215	127.04 ± 1.55	3.20 ± 0.05	3.91
GNN-Hopf	10,067	100.64 ± 0.07	2.58 ± 0.04	3.91
Hopf+Neural	30,127	34.23 ± 0.04	2.52 ± 0.05	4.86

The Coupled Hopf model is additionally fitted by an exhaustive grid search over the global coupling G, bifurcation parameter a, and coupling exponent κ ($5 \times 4 \times 2 = 40$ parameter combinations ranked by the training objective). On the same hardware this costs 2.32 s per trial and 92.7 s in total.