Linearizing and forecasting: a reservoir computing route to digital twins of the brain

Gabriele Di Antonio[1,2,3], Tommaso Gili[4], Andrea Gabrielli[1,2], and Maurizio Mattia[3,*]

[1] *"Enrico Fermi" Research Center - CREF, 00184 Rome, Italy*
[2] *Dip. di Ingegneria Civile, Informatica e delle Tecnologie Aeronautiche, Università degli Studi "Roma Tre", 00146 Rome, Italy*
[3] *Natl. Center for Radiation Protection and Computational Physics, Istituto Superiore di Sanità, 00161 Rome, Italy*
[4] *Networks Unit, IMT Scuola Alti Studi Lucca, 55100 Lucca, Italy*
[*] *Corresponding author: maurizio.mattia@iss.it*

October 22, 2024

Abstract

Exploring the dynamics of a complex system, such as the human brain, poses significant challenges due to inherent uncertainties and limited data. In this study, we enhance the capabilities of noisy linear recurrent neural networks (lRNNs) within the reservoir computing framework, demonstrating their effectiveness in creating autonomous in silico replicas – digital twins – of brain activity. Our findings reveal that the poles of the Laplace transform of high-dimensional inferred lRNNs are directly linked to the spectral properties of observed systems and to the kernels of auto-regressive models. Applying this theoretical framework to resting-state fMRI, we successfully predict and decompose BOLD signals into spatiotemporal modes of a low-dimensional latent state space confined around a single equilibrium point. lRNNs provide an interpretable proxy for clustering among subjects and different brain areas. This adaptable digital-twin framework not only enables virtual experiments but also offers computational efficiency for real-time learning, highlighting its potential for personalized medicine and intervention strategies.

Introduction

The study of complex dynamical systems has a central role in contemporary scientific research. Over recent decades, a plethora of data-driven methodologies has emerged, denoting a field in rapid evolution that is capable to addressing systems with increasing complexity. Among these methodologies, machine learning techniques have become increasingly popular due to their effectiveness in modelling complex datasets and deliver accurate predictions [1, 2, 3, 4]. However, in this rapid march toward prediction, we have lost the clarity of understanding. Indeed, while the inferred models are impressive in their accuracy, they often lack interpretability [5, 6].

To address this limitation, Koopman theory offers a valuable framework (Fig. 1a). By mapping nonlinear systems into higher-dimensional spaces, it offers clearer insights into the system dynamics [7, 8, 9, 10]. This approach involves the development of data-driven methods aimed at computing finite-size linear representations of nonlinear systems within a functional space. Once an equivalent linear system is inferred, interpretability becomes straightforward, allowing to gain understanding of the time scales and dynamical modes of the observed systems.

Reservoir computing (RC) is an alternative machine learning approach (Fig. 1b). It employs high-dimensional recurrent neural networks (RNNs) with fixed internal couplings, which nonlinearly transform the time series received as input. This leads to a high dimensional nonlinear representation of the input, enabling in principle to predict future samples of the same time series by reading out the current state of the network. The power of this approach lies in the fact that a successful readout can be conceptualized as a projection of the network state onto a manifold learned through linear regression [11, 12]. This solution is reminiscent of the efficient coding approach adopted by biological networks of neurons in the intact brain [13, 14, 15]. Consequently, it is not surprising that RNNs have been extensively studied in neuroscience for analytical purposes and hypothesis generation [16, 17, 18, 19].

Having machinery capable to predict observations related to the future state of a dynamical system by using past observations as input allows, in principle, to build a replica of the same system. Indeed, predicted observations can be fed back as input resulting in a generative model. Both Koopman theory and RC approaches have the potential to generate such a 'digital copy,' producing time series statistically equivalent to those generated by the original system. This machinery can then serve as a 'digital twin' – a concept that has emerged as a transformative paradigm in the study of dynamical systems [20, 21]. Such digital twins can be effective tools for analyzing and simulating their physical counterparts. Successful examples include digital copies of brain activity designed to assist neurosurgeons in dealing with drug-resistant epileptic patients [22, 23]. Furthermore, these tools can provide a complementary workbench for designing stimulation approaches to be tested *in silico*, ultimately leading to control systems for neurorehabilitation [24].

Here, we investigate the potential of RNNs with linear units to implement effective digital twins of single-subject brain dynamics, measured as metabolic activity (i.e., the BOLD signal) extracted from functional magnetic resonance imaging (fMRI). We demostrate the autoregressive nature of this approach and its effectiveness in representing whole-brain activity into a low-dimensional latent state space. Such a reduced representational complexity mitigates the risk of overfitting often associated with one-to-one fitted physical models, especially in realistic scenarios where only a limited number of observations are available and only a partial view of the system can be experimentally accessed. The integration of endogenous memoryless fluctuations is another essential component that enables the digital twin to self-sustain a noisy dynamics statistically equivalent to the one experimentally measured. Finally, we show that the linearity of the model allows for stable and interpretable decomposition consistent with the Koopman theory.

Results

Linear RNNs can copy nonlinear systems

Here we investigate the reservoir computing abilities of a RNN with N linear units (lRNN). The state of the lRNN is the vector $\boldsymbol{r}(t)$ whose elements are the activities $r_i(t)$ of each unit, following the linear dynamics

$$\tau \frac{d\boldsymbol{r}(t)}{dt} = \boldsymbol{W}\boldsymbol{r}(t) + \boldsymbol{W}^{\text{in}}\boldsymbol{\omega}(t) - \boldsymbol{r}(t)\,. \tag{1}$$

The synaptic matrix $\boldsymbol{W} \in \mathbb{R}^{N\times N}$ determines the recurrent input $\boldsymbol{W}\boldsymbol{r}$, which, together with the contributions from the external environment $\boldsymbol{W}^{\text{in}}\boldsymbol{\omega}(t)$, composes the total input $\boldsymbol{h}$ that each unit receives. In general, these currents are nonlinear transformed by an input-output gain function $\Phi(\boldsymbol{h})$ we assume here to be $\Phi(\boldsymbol{h}) = \boldsymbol{h}$ – a condition usually associated with unit activities perturbatively fluctuating around the quiescent state (see Methods). Network units receive a source of external input due to the M observables $\omega_i(t)$ of the systems under investigation, mediated by the synaptic matrix $\boldsymbol{W}^{\text{in}} \in \mathbb{R}^{N\times M}$. From Eq. (1) the network state at any time t results then to be

$$\boldsymbol{r}(t) = \int_{-\infty}^{t} e^{(\boldsymbol{W}-\boldsymbol{I})(t-t')/\tau}\boldsymbol{W}^{\text{in}}\boldsymbol{\omega}(t')dt'\,, \tag{2}$$

where $\boldsymbol{I} \in \mathbb{R}^{N\times N}$ is the identity matrix.

If the lRNN is capable to display the so-called 'generalized synchronization' property in response to the input $\boldsymbol{W}^{\text{in}}\boldsymbol{\omega}$ [25, 26], $\boldsymbol{r}(t)$ effectively embed the dynamics of the 'whole' observed system into its state space. This because the lRNN incorporates the 'echo' [27, 11, 28] of past observations implementing a dimensional embedding *a la* Takens [29, 30]. In this RC framework, for sufficiently high N a simple linear transformation capable to approximately map the network state to a target output exists

$$\boldsymbol{\omega}(t) \approx \boldsymbol{W}^{\text{out}}\boldsymbol{r}(t) = \int_{0}^{\infty} \boldsymbol{G}(s)\boldsymbol{\omega}(t-s)ds \tag{3}$$

with synaptic weights $\boldsymbol{W}^{\text{out}} \in \mathbb{R}^{M\times N}$ computed via a simple linear regression [11, 12]. The linear kernels $\boldsymbol{G}(s) = \boldsymbol{W}^{\text{out}}e^{(\boldsymbol{W}-\boldsymbol{I})(s)/\tau}\boldsymbol{W}^{\text{in}} \in \mathbb{R}^{M\times M}$ (i.e., the Green functions) resulting from Eq. (2) reveals the autoregressive nature of the RC approach in this lRNN case. This concept can be better clarified by resorting to the eigendecomposition of the synaptic matrix $\boldsymbol{W} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{-1}$. Here $\boldsymbol{\Lambda}$ is the diagonal matrix of the eigenvalues

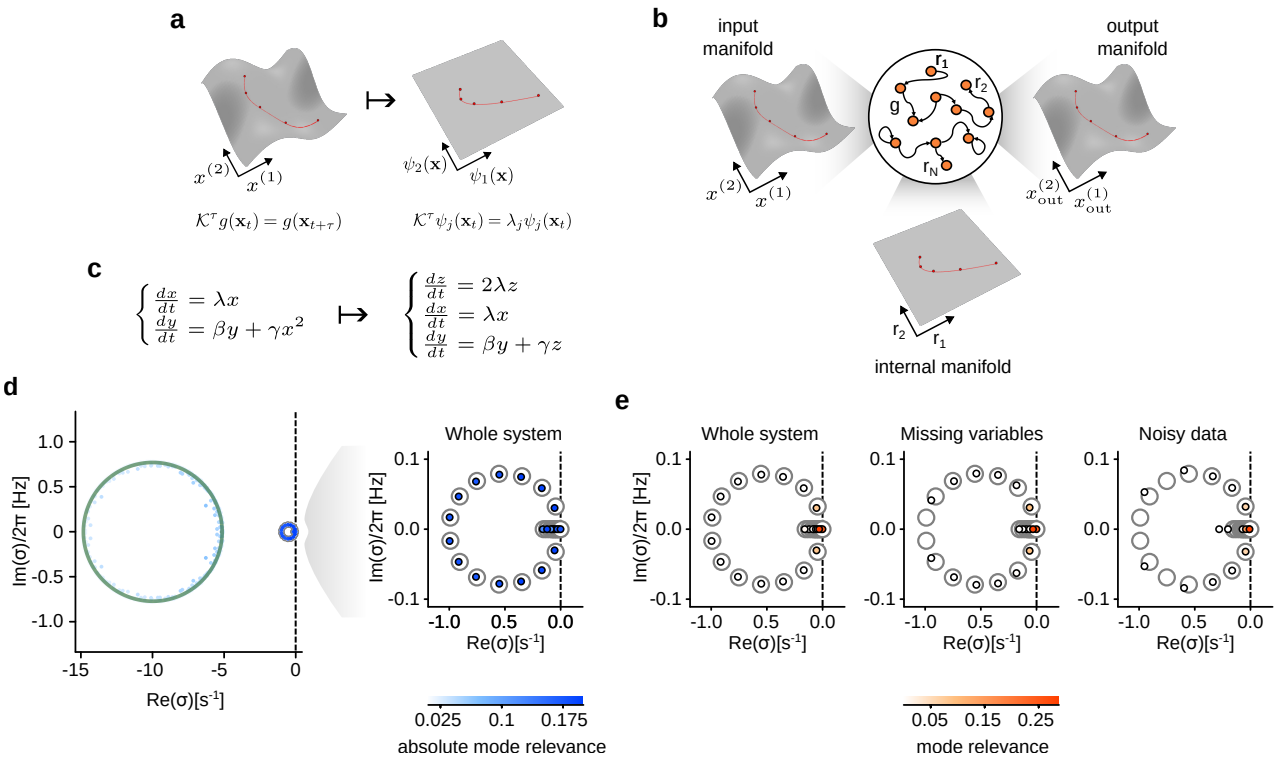

Figure 1: **IRNN encodes a linear representation of the input system**. **a,** Nonlinear dynamics can be projected onto linear manifolds. Within the infinite functional space of the system observables, the eigenfunctions of the Koopman operator establish a basis for decomposing nonlinear dynamics into simpler linear processes. **b,** Reservoir computing (RC) with recurrent neural networks (RNNs). After 'learning', the readout weights $\boldsymbol{W}^{\text{out}}$ are set to reproduce the system observables provided as input. The resulting network is composed of linear units (IRNN) and its post-learning autonomous dynamics encapsulates a linear representation of the original system. **c,** Example of a 2-dimensional nonlinear system that can be decomposed into a set of 3 linearly evolving variables. **d,** Spectrum of a IRNN with $N = 500$ units trained to mimic a 15-dimensional version of the system in (c) composed of 5 linear x variables and 10 nonlinear y variables. Green circles, spectrum of the IRNN before training. Gray circles, theoretically-derived eigenvalues of the 20-dimensional linearized system to be replicated (see text for details). Small circles, Laplace transform poles from the post-learning autonomous IRNN replicating the example system. Filling colors code the absolute mode relevance $\langle |\Xi_{jk}| \rangle_j$. Right panel, zoom in of the left panel showing the 20 most significant modes. **e,** Same spectrum as in (d) with filling colors coding the mode relevance $\langle |\Xi_{jk} v_k(0)| \rangle_j$ (left panel, see text for details). Central panel, spectrum of a IRNN trained on incomplete information (i.e., omitting all x and 5 y variables). Right panel, spectrum of a IRNN trained on the same observables but perturbed by noise.

($\Lambda_{ii} = \lambda_i$) and the eigenvectors are the columns of $\boldsymbol{Q}$: $\boldsymbol{W} |Q_i\rangle = \lambda_i |Q_i\rangle$. By applying this decomposition to the kernel expression, a superposition of N exponential modes results: $\boldsymbol{G}(s) = \sum_{n=1}^{N} \boldsymbol{J}_n e^{(\lambda_n - 1)s/\tau}$, where $\boldsymbol{J}_n$ are rank-1 matrices defined as $\boldsymbol{J}_n = \boldsymbol{W}^{\text{out}} |Q_n\rangle \langle Q_n^{-1}| \boldsymbol{W}^{\text{in}}$ and $\langle Q_n^{-1}|$ is the n-th row of the inverse matrix of the eigenvectors (see Methods). To have a smooth and compact kernel the spectral radius $\rho = \max_{n\in[1,N]} \text{Re}\,\lambda_n$ of $\boldsymbol{W}$ must be smaller than 1. Under this condition Eq. (3) describes a continuous-time auto-regressive (AR) model with order N (number of exponential modes in $\boldsymbol{G}(s)$). This relationship between RC with IRNN and AR models generalizes partial results obtained in the discrete-time domain [31, 32]. As we will see later, the order of the equivalent AR model can be much smaller than N as many $\boldsymbol{J}_n$ are usually close to $\boldsymbol{0}$.

If we now replace the observables $\boldsymbol{\omega}(t)$ provided as input with the reconstructed once from Eq. (3), the IRNN follows the autonomous dynamics

$$\tau \frac{d\boldsymbol{r}}{dt} = \tilde{\boldsymbol{W}} \boldsymbol{r} - \boldsymbol{r} \tag{4}$$

where $\tilde{\boldsymbol{W}} = \boldsymbol{W} + \boldsymbol{W}^{\text{in}}\boldsymbol{W}^{\text{out}}$ is the updated (i.e., 'learned') synaptic matrix. If the network can correctly predict the future of the target output, Eq. (4) provides a linear representation of the system under analysis. Relying as above on the eigenmode decomposition $\tilde{\boldsymbol{W}} = \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{Q}}^{-1}$, the projections $\boldsymbol{v}(t) = \tilde{\boldsymbol{Q}}^{-1}\boldsymbol{r}(t)$ result to evolve in time as a set of decoupled variables:

$$v_k(t) = v_k(0)e^{(\tilde{\lambda}_k - 1)t/\tau} \quad \forall k \in [1, N]. \tag{5}$$

Analogously to the Koopman decomposition, a suited linear combination of these modes implements an effective decomposition of the time series $\boldsymbol{\omega}(t)$:

$$\boldsymbol{\omega}(t) = \boldsymbol{W}^{\text{out}}\boldsymbol{r}(t) \equiv \boldsymbol{\Xi}\boldsymbol{v}(t)\,, \tag{6}$$

with $\boldsymbol{\Xi} = \boldsymbol{W}^{\text{out}}\tilde{Q}$. The Laplace transform of the reconstructed observable will then have N poles $\sigma_k = (\tilde{\lambda}_k - 1)/\tau$ with their related residues, allowing us to represent the dynamical properties of the 'digital twin' (i.e., the autonomous IRNN), and linking it to AR models and finite approximations of the Koopman operator (see Methods). Indeed, the k-th residue is proportional to $\Xi_{jk}v_k(0)$ and provides in absolute value the relevance of the k-th mode in describing the observable ω_j, given the initial network state $\boldsymbol{r}(0)$.

As a testing ground for the theoretical framework derived above we consider the following nonlinear system:

$$\begin{cases} \dot{\boldsymbol{x}} = \boldsymbol{C}_x\boldsymbol{x} \\ \dot{\boldsymbol{y}} = \boldsymbol{C}_y\boldsymbol{y} + \boldsymbol{C}_{yx}\boldsymbol{x} \odot \boldsymbol{x} \end{cases} \tag{7}$$

Choosing $\boldsymbol{C}_x$ as a diagonal matrix, this system can be mapped onto a linear one by introducing a set of new observables $\boldsymbol{z} = \boldsymbol{x} \odot \boldsymbol{x}$ (Fig. 1c). In this case, the related infinitesimal generator of the Koopman operators is equivalent to the finite matrix $\boldsymbol{K}$ with diagonal blocks $\boldsymbol{C}_x$, $2\boldsymbol{C}_x$, $\boldsymbol{C}_y$.

We trained a IRNN to replicate this system from the observables $\boldsymbol{\omega} = [\boldsymbol{x}; \boldsymbol{y}]$ with $\boldsymbol{x} \in \mathbb{R}^5$ and $\boldsymbol{y} \in \mathbb{R}^{10}$. As initial synaptic matrix $\boldsymbol{W}$ we chosen the one associated to ring-like connectivity (i.e., $W_{i,i+1} = W_{N,1} = \rho$ are the only non-zero elements) known to have optimal performances in the case of linear units [33]. In Fig. 1d the poles σ_k of the reconstructed observables from Eq. (6) are shown together with the related mode relevance. Despite the nonlinearity of the system to replicate, the overlap between the poles σ_k and the theoretical spectrum of $\boldsymbol{K}$ is remarkable. Furthermore, only 20 out of the $N = 500$ modes have $\langle|\Xi_{jk}|\rangle_k$ significantly different from 0 (filled circles).

Decomposition modes display damped oscillations as $\tilde{\lambda}_k \in \mathbb{C}$ and a few of them, the slowest ones, contribute significantly to the dynamics (Fig. 1e, red-colored small circles). Such low-dimensionality is not altered even if during learning the IRNN receives noisy-contaminated observables or some of them are not provided as input. Indeed, the autonomous IRNN continues to predict accurately the system's future also under these conditions, although the less relevant poles move away from the imaginary axis (missing match between small and large circles).

IRNNs successfully forecast fMRI time series

To explore the applicability of the RC with IRNN, we examined time series data from a real-world physical system. We considered the brain activity of 20 healthy subjects at rest (i.e., in a state of quiet wakefulness), measured through fMRI. The experimental time series was derived from the BOLD signals recorded from a subset of voxels, which serve as proxies for neuronal activity across 11 cortical areas belonging to the language network (Fig. 2a, see Methods).

The BOLD time series underwent preprocessing using principal component analysis (PCA) to reduce redundancy and achieve dimensionality reduction, which was found to be relatively high (see Fig. 2b, c). Specifically, we assessed the correlation between the information lost during dimensionality reduction and the actual signals. Our analysis revealed that the first principal components (PCs), which explain 99.5% of the variance (Fig. 2d,e), excluded a residual activity that was indistinguishable from white noise. Based on this criterion, we determined that, on average, only four PCs per cortical area should be considered. Consequently, the average dimensionality of the dataset was $M \approx 44$ PCs per subject (refer to Methods for details). This relatively high-dimensional time series served as the input for the IRNN. The readout synaptic matrix $\boldsymbol{W}^{\text{out}}$ was computed over the first 280 seconds (Fig. 2f).

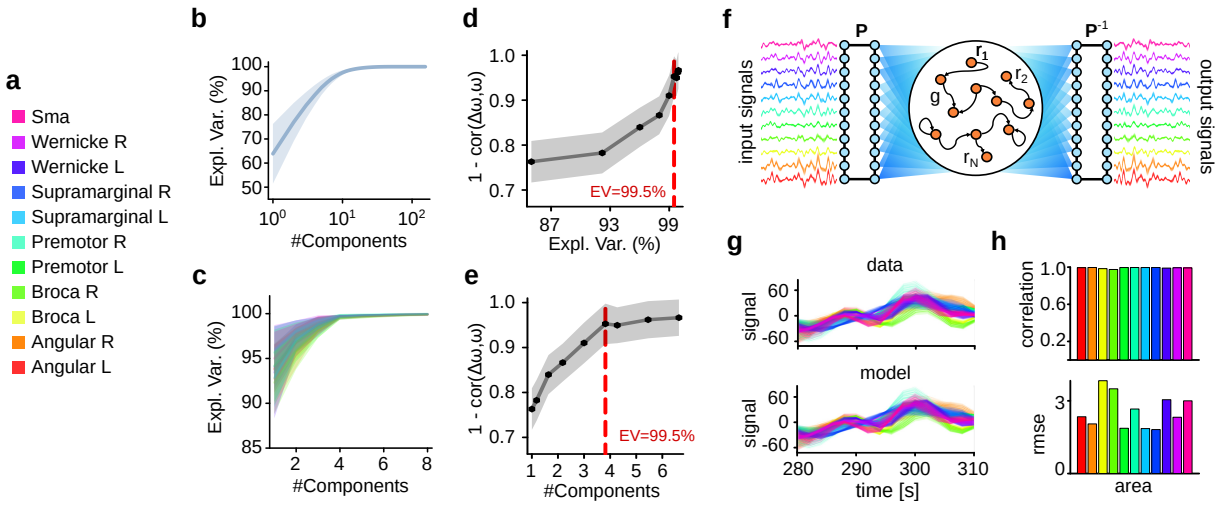

Figure 2: **Compression and prediction of resting state fMRI activity**. **a,** Color legend associated with the 11 cortical areas taken from the language network. **b,** Explained variance (EV) from PCA of the fMRI BOLD signals of the voxels in the language network. Solid line, average across the 20 considered subjects. Shaded area, standard deviation across subjects. **c,** EV of the PCA for each cortical area. Colors as in (a). **d-e,** 1 minus the correlation between the residuals from the PCA reconstruction ($\Delta\omega$) and the original data (ω) per single area in the last (unseen) 30 seconds. A value equal to 1 means that the remaining information is just white noise. Vertical dashed lines, explained variance (d) and number of first PCs (e) required to reach the EV of 99.5%. **f,** Schematic design of the used IRNN. The input from each area is compressed to have an EV of 99.5%. $\boldsymbol{P}$ is the matrix operating this linear transform leading to an average of 44 observables. The IRNN output reconstructing these observable is then uncompressed applying $\boldsymbol{P}^{-1}$. **g,** Example of forecasted (bottom) and original (top) activities during the test window (last 30 seconds). Each curve is the BOLD signal of a voxel with color indicating the cortical area of belonging. **h,** Forecast performances of the post-learning autonomous IRNN per area. Correlations between predicted and experimental activity (top) and root mean square error (bottom).

The decision to utilize a transformation of brain activity that results in information loss may appear as counter-productive. However, projecting onto the first PCs was expected to be less sensitivity to the uncertainties inherent in the measured BOLD signals. By employing these first PCs as inputs, we provided the IRNN with denoised representations of brain activity. Moreover, thanks to the echo state property [27, 11, 28], the missing information regarding the state of the observed system can be effectively recovered through the dimensional embedding that the IRNN performs.

With the learned $\boldsymbol{W}^{\text{out}}$ – which varies from subject to subject – we established a closed loop between the input and output of the IRNN, effectively creating an autonomous digital twin of the brain network under investigation. This approach yielded a remarkable overlap between the reconstructed and measured time series (see Fig. 2g for an example subject). At the population level, we evaluated the forecasting performance by calculating the correlations between the replicated and experimental BOLD signals, as well as the root mean square errors (RMSEs) for the final unseen 30 seconds of the time series (Fig. 2h). From this perspective, the inferred digital twin demonstrated high-quality reconstruction of brain activity, further validating its effectiveness.

IRNN as a digital twin of brain activity

The IRNN dynamics described thus far is dissipative, meaning that the state of the network ultimately relaxes to $\mathbf{0}$. To counteract this, we introduced an endogenous noise into the activity of each network unit, creating a generative mechanism that persists over time. This approach models the system under investigation using a set of Ornstein-Uhlenbeck processes $\boldsymbol{\eta}(t)$, leading to the following stochastic differential equation (SDE):

$$\begin{cases} \tau\dot{\boldsymbol{r}} = -\boldsymbol{r} + \tilde{\boldsymbol{W}}\,\boldsymbol{r} + \boldsymbol{\eta} \\ \tau\dot{\boldsymbol{\omega}} = -\boldsymbol{\omega} + \boldsymbol{W}^{\text{out}}\tilde{\boldsymbol{W}}\boldsymbol{r} + \boldsymbol{W}^{\text{out}}\boldsymbol{\eta} \end{cases} . \tag{8}$$

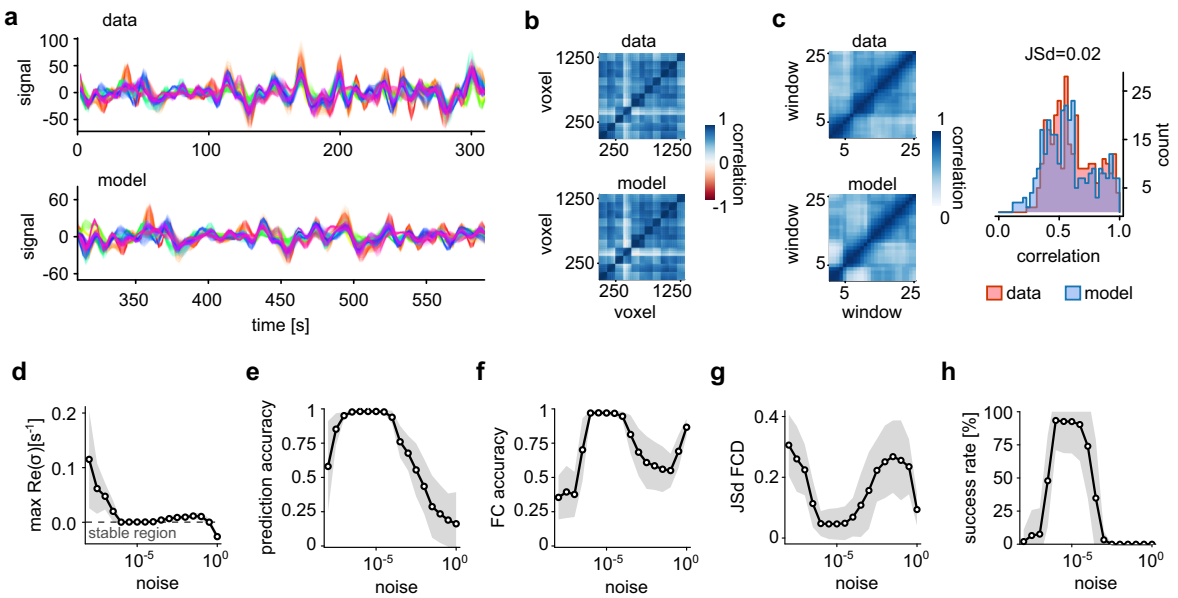

Figure 3: **Endogenous noise in IRNN is needed to replicate brain activity**. **a,** Comparison of actual brain activity (top) with the corresponding output from a trained autonomous IRNN over the subsequent prediction interval (bottom). **b-c,** Analysis of functional connectivity (FC) and functional connectivity dynamics (FCD) computed from the activity depicted in (a). Right panel, distribution of FCD values for IRNN-generated and exeprimental time series with the related Jensen-Shannon distance. **d-g,** Characterization of simulated subject profiles under varying amplitude of endogenous noise, detailing the maximum pole real part (d), correlation between data and model prediction (e), FC correlation with actual data (f), and Jensen-Shannon distance between FCDs. **h,** Success rate in reproducing above-threshold performances of the features in panels (e-g) (see main text for details).

As a result, the reconstructed $\boldsymbol{\omega}$ share the same sources of noise transformed by the linear map $\boldsymbol{W}^{\mathrm{out}}$. Note that, since the system is linear, this formulation does not change the theoretical framework derived above; in this case, it applies to the expectation values $\mathbb{E}(\boldsymbol{r})$ and $\mathbb{E}(\boldsymbol{\omega})$.

To validate the effectiveness of this stochastic version of the IRNN in accurately replicating brain activity, we numerically integrated the network dynamics for an additional 310 seconds (Fig. 3a). We then compared the functional connectivity (FC) – defined as the correlation between BOLD signals of all possible voxel pairs – calculated from this simulation to that obtained from the experimental data (Fig. 3b). Furthermore, the temporal evolution of the FC, referred to as functional connectivity dynamics (FCD), closely aligned with the experimental observations. This similarity was quantified by the Jensen-Shannon distance (JSd) between the distribution values of the associated matrices (Fig. 3c).

Finally, we conducted a comprehensive analysis of the optimal level of noise across all subjects (Fig. 3), focusing on the accuracy of BOLD signal forecasting and the IRNN capability to reproduce both FC and FCD. To evaluate the quality of these aspects simultaneously, we introduced a success rate defined by applying fixed thresholds to the performance metrics: $\mathrm{TH}_{\mathrm{acc}} = 80\%$, $\mathrm{TH}_{\mathrm{FC}} = 85\%$ and $\mathrm{TH}_{\mathrm{JSd}} = 0.1$. Additionally, we assessed the stability of the inferred IRNN by examining the maximum real part of the poles σ_k (Fig. 3d). This analysis enabled us to identify a common optimal level of endogenous noise across subjects with a standard deviation of 10^{-6} (see Fig. 3h). Interestingly, the existence of an amplitude optimality of the injected noise displays some similarities with the stochastic resonance phenomenon found in other studies [34], which in our case spontaneously emerged in all the subjects.

Dynamical properties of the digital twin

We showed that RC with IRNNs can effectively reproduce resting state fMRI activity. Each simulated voxel can be described as a decomposition of linearly evolving modes. The dynamical properties of this decomposition can be

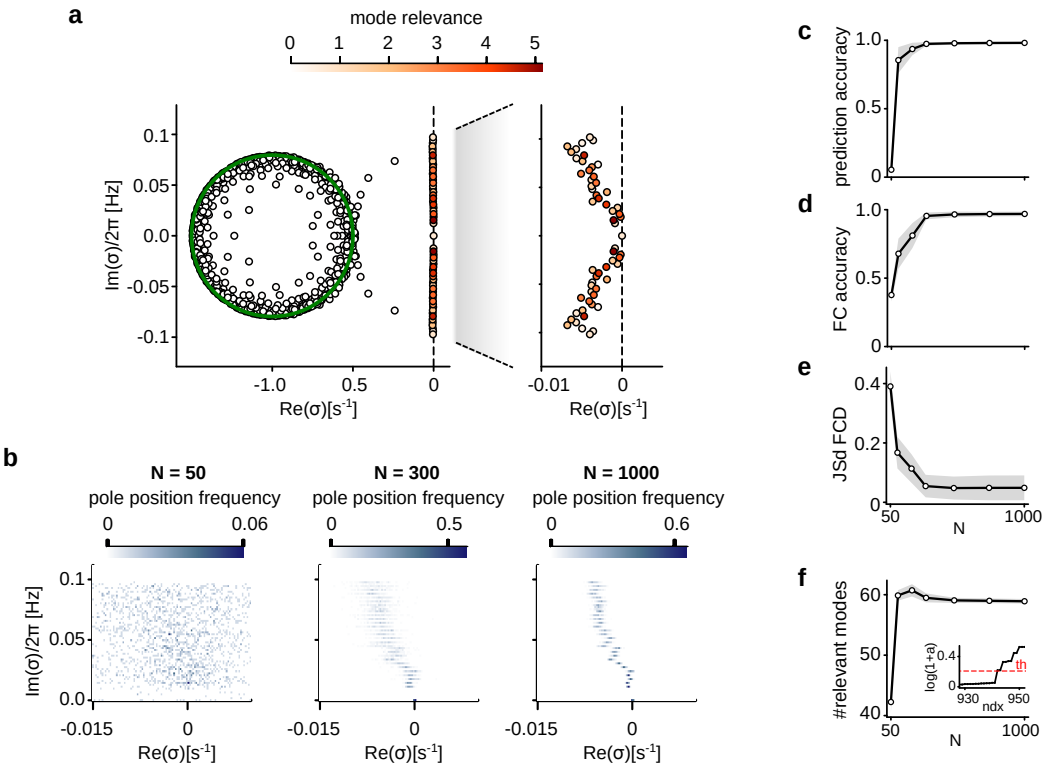

Figure 4: **Convergence of the mode decomposition with increasing IRNN size. a,** Mode decomposition of a single cortical area for an example subject. Eigenvalues of $\tilde{\boldsymbol{W}}$ are color-coded as in Fig. 1e. **b,** Distribution of eigenvalues for three distinct network sizes N, analyzed across 100 randomly sampled IRNNs. **c**, Number of relevant modes (i.e., with relevance greater than 5%). Inset, sorted mode relevance and related threshold (dashed line). **d-f,** Accuracy of forecasted activity (d), reproduction fidelity of FC (e) and of FCD (f). Gray shadings and black lines, standard deviation and mean across subjects, respectively.

effectively represented by the poles σ_k of the Laplace transform of the inferred IRNN, which are the eigenvalues $\tilde{\lambda}_k$ of the learned synaptic matrix $\tilde{\boldsymbol{W}}$. This spectrum of eigenvalues is shown in Fig. 4a for an example subject. For each of the modes, the relevance is averaged across all voxels and it is color coded, showing a pattern similar to what seen in Fig. 1e where only a subset of poles are moved towards the imaginary axis. These poles are the most relevant and display a specific organization like a rotated parabola. The modes with the highest relevance (i.e., the darkest) appear to be distributed at specific frequencies falling into the range of infra-slow oscillations (< 0.1 Hz), a typical footprint of the resting state and of the unconscious brain activity [35, 36, 37, 38]. As a representative example, in Fig. 4a, the most relevant mode is one of the most persistent, characterized by a relatively small $\mathrm{Re}\,\sigma_k$, and exhibits a resonant frequency of approximately 0.02 Hz.

Subsequently, we investigated the stability of this representation exploring how the spectrum of $\tilde{\boldsymbol{W}}$ and the mode relevance change according to the number N of units in the IRNN. As shown in Fig. 4b, the density of eigenvalues in the complex plane $(\mathrm{Re}\,\sigma, \mathrm{Im}\,\sigma)$ becomes less and less scattered with increasing network size. Besides, all the performance measures converged to fixed values corresponding to a high quality of the data replicate by the digital twin (Fig. 4c-e), proving that a finite size IRNN can reach asymptotically high performances. Intriguingly, as convergence is approached the average number of relevant modes stands at about 60 independent components (Fig. 4f). This number is significantly higher than the number of experimental PCs provided as input to the IRNN (44 on average). Thus, brain activity replicated by the inferred IRNN appear to live in a latent state space whose dimensionality is larger than the one determined by its observation.

Functional connectivity illustrated in Fig. 5a-top for an example subject is then fully replicated by the inferred (and stochastic) IRNN. This effective copy occurs even though the dimensionality of the latent state space of the digital twin is significantly smaller than the number of voxels encompassing the examined language network. To gain a deeper understanding of this result, we 'opened the box' by examining the linear transformation $\boldsymbol{Z} \equiv$

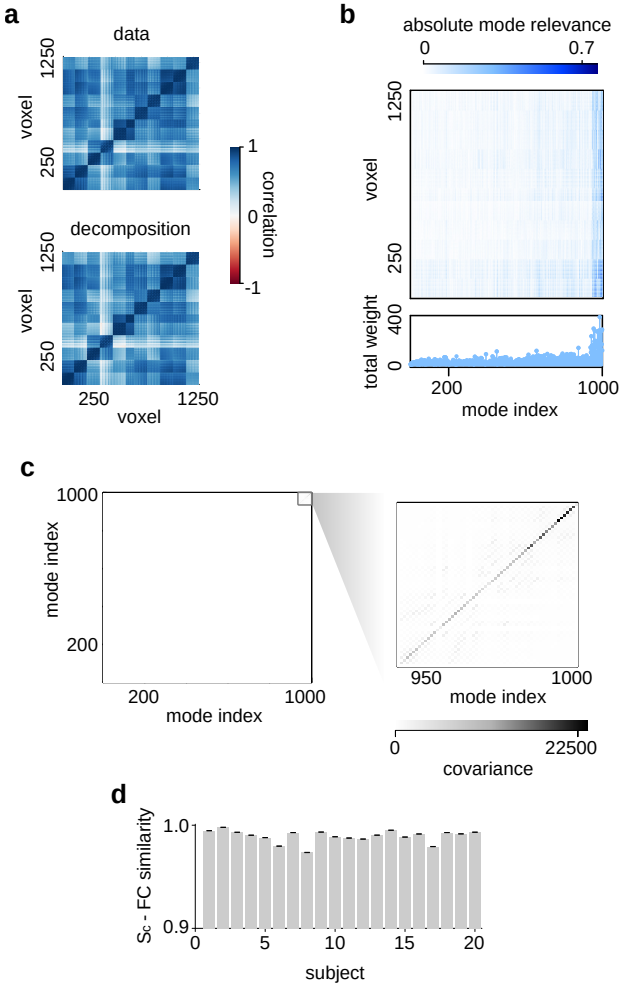

Figure 5: **Low-dimensional mode decomposition explains functional connectivity. a**, Functional connectivity of an example subject (top) and cosine similarity (bottom) between rows of the matrix $\boldsymbol{Z}$ mapping eigenmode projections to voxel-wise BOLD activity from the inferred lRNN with $N = 1000$ units (see main text). **b**, Absolute mode relevance $|Z_{jk}|$ for the same lRNN together with the total weight $\sum_j |Z_{jk}|$ of each of the N modes (bottom). Modes are sorted according to the real part of the associated eigenvalue ($\mathrm{Re}\,\tilde{\lambda}_k$). **c**, Covariance matrix of the eigenmode projections $v_k(t)$ (see Eq. (6)) during the extended autonomous phase (310 seconds post learning in Fig. 3a-bottom). Right, zoomed-in view on the slow and most relevant modes. **d** Correlation between the upper diagonal elements of the functional connectivity (FC) and the cosine similarity ($\mathrm{S_C}$) of the rows of $\boldsymbol{Z}$ for each subject. Averaged over 100 noise realizations. Error bars, SEM.

$\boldsymbol{P}^{-1}\boldsymbol{\Xi} \in \mathbb{R}^{L\times N}$. According to Eq. (6), this matrix facilitates the mapping of the N eigenmode projections $\boldsymbol{v}(t)$ onto the L-dimensional voxel-wise BOLD activity, represented as $\boldsymbol{x}(t) = \boldsymbol{Z}\boldsymbol{v}(t)$. The absolute values of the elements in this matrix are displayed in Fig. 5b for the same subject. Unsurprisingly, the most significant contributions arise from the slowest (rightmost) modes, specifically those with the largest $\mathrm{Re}\,\tilde{\lambda}_k$. Additionally, by measuring the covariance between the eigenmode projections $v_k(t)$, it becomes evident from Fig. 5c that they are largely uncorrelated. Only about 60 of the slowest modes exhibit significant variability, as indicated by the dark diagonal elements. This empirical evidence suggests that functional connectivity can be estimated directly using the cosine similarity $\mathrm{S_C}$ between the rows of the matrix $\boldsymbol{Z}$ (see Methods):

$$\mathrm{FC}_{jk} \approx \mathrm{S_C}\left(\langle Z_j|\,,\langle Z_k|\right)\,. \tag{9}$$

In Fig. 5a-bottom, the similarity matrix is presented, showcasing a remarkable overlap with the experimental functional connectivity $\mathbf{FC}$. It is important to note that in computing $\boldsymbol{Z}$, we considered only the columns of $\boldsymbol{\Xi}$

associated with the relevant modes. This further confirms that the remaining $N - 60$ degrees of freedom are nearly irrelevant.

IRNN as a proxy to characterize subjects and brain areas

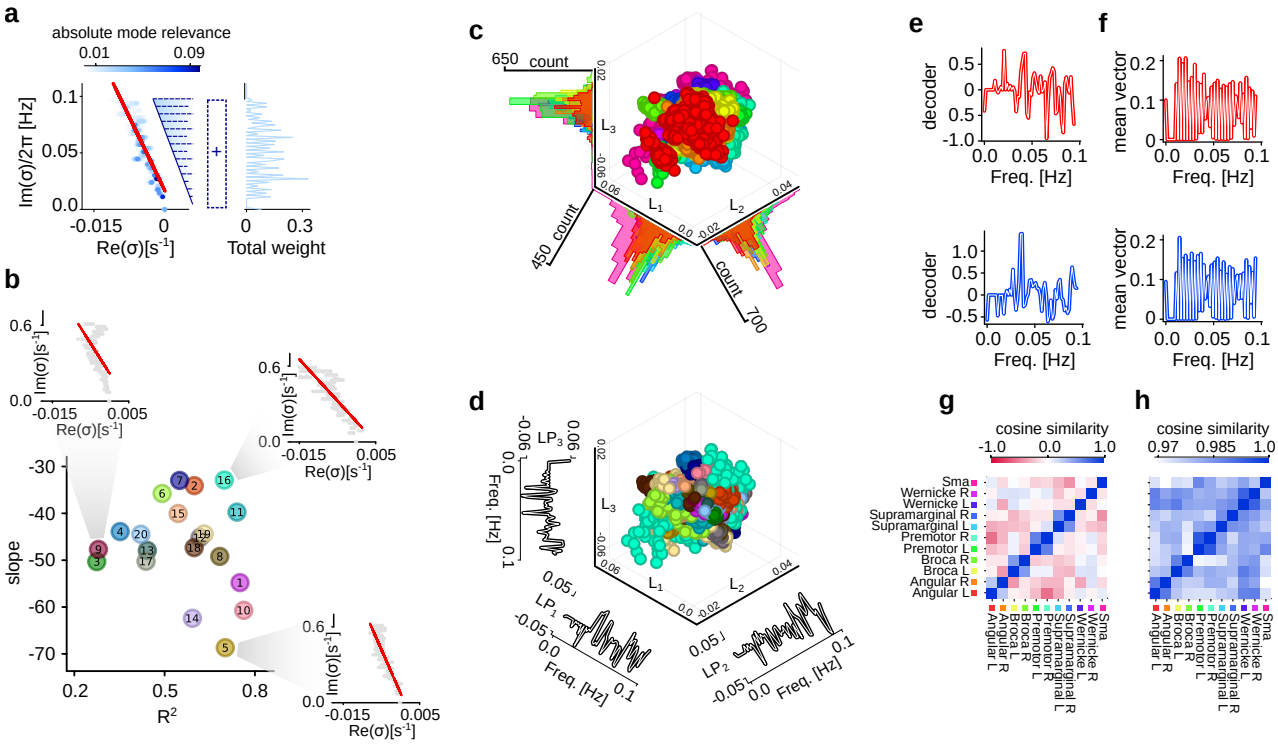

Figure 6: **Characterization of subjects and brain areas. a,** Left: Spectrum of relevant eigenmodes for a sample subject. A linear regression (red line) is performed between the real and imaginary parts of the poles, σ_k, with the slope serving as the first feature that characterizes the subject. The coefficient of determination, R^2, from this regression represents the second feature. Right: Frequency vector, i.e., the histogram of Im σ_k each counted with multiplicity given by the averaged absolute mode relevance across voxels in a single brain area: $\langle |\Xi_{jk}| \rangle_{j \in \mathrm{BA}}$. **b,** Slope of linear regression and R^2 of the 20 subjects in the dataset. Insets: eigenmode spectra of three representative subjects with the related linear regressions (red lines, see main text for details). **c,** First three latent variables derived from the linear discriminant analysis (LDA) applied to the frequency vectors of 11 brain areas across all 20 subjects. The densities of circles for each latent variable are also plotted. For each subject, 100 IRNNs are inferred by sampling different endogenous noise in their stochastic dynamics, resulting in a total of 22,000 circles. Colors code brain areas according to Fig. 2a. **d,** Same as (c) but with circles colored according with the subject color as in (b). Composition of the three latent variables LP_1, LP_2, LP_3 are show on the side of their respective axes. **e-f,** Linear decoders (e) discriminating the different areas in the frequency-vector space and average frequency vectors (f) for two sample brain areas (red and blue). **g-h,** Cosine similarity between linear decoders (g) and between frequency vectors (h) per area averaged across subjects.

Given that inferred IRNNs reliably replicate observed brain activity and their modal decomposition defines their dynamical properties, a critical question emerges: Can IRNNs serve as proxies for understanding the similarities and differences among subjects and cortical areas? To address this, we characterized the spectrum of relevant eigenvalues by focusing on two key features: the linear relationship between their real and imaginary parts, and the overall significance of each oscillation frequency, as illustrated in Fig. 6a. The slope of this linear relationship indicates the correlation between decay times and oscillation frequencies, while the coefficient of determination (R^2) assesses the goodness of fit for this linear model. In this two-dimensional parameter space, subjects are systematically distributed among three distinct regions (Fig. 6b). Those with high R^2 values display either steep

or shallow slopes, whereas subjects with low R^2 values indicate a poor linear fit and suggest the presence of persistent isolated high-frequency oscillatory modes. Within this representational space, spectra characterized by steeper slopes (located at the bottom) are associated with poles that are closer to the imaginary axis. This suggests that those IRNNs (and consequently subjects) exhibit longer relaxation time scales, which may be linked to brain activity approaching a critical point where the resting state could become unstable and displaying a more complex dynamics.

Shifting our focus to the frequency vectors derived from the inferred IRNN, we sought statistical regularities that reveal invariant features associated with individual brain areas and subjects. The frequency vector for a given brain area in a subject is represented by the histogram of Im σ_k from the corresponding IRNN, each counted with multiplicity proportional to its absolute mode relevance averaged across the voxels in that area: $\langle|\Xi_{jk}|\rangle_{j\in\mathrm{BA}}$. We then conducted a linear discriminant analysis (LDA) in the frequency-vector space to identify patterns that effectively discriminate between different brain areas (Fig. 6c). Despite originating from distinct subjects, the resulting data points clustered according to their respective brain areas. It is important to note that the compositions of the latent variables (L_1, L_2, and L_3) correspond to specific frequency patterns (Fig. 6d). Among these, the pattern associated with L_1 appears most relevant, as data points on this linear manifold are distinctly clustered based on the brain area of belonging (Fig. 6c).

To further isolate the frequencies at which each brain area exhibits unique dynamics, we computed the corresponding linear decoders. The elements of the coefficient vectors characterizing each decoder are assigned high absolute values if the associated frequency plays a crucial role in distinguishing that area from others (Fig. 6e). The resulting patterns are markedly distinct from the average frequency vector, which, in contrast, reveals a continuum of relevant frequencies (Fig. 6f). In fact, the frequency vectors show significant similarities, making it challenging to differentiate between areas even if their are located in different hemispheres (Fig. 6h). In contrast, the decoders exhibit a lower degree of similarity, except when comparing corresponding areas from both hemispheres (Fig. 6g). Interestingly, there is an exception to this trend: the decoder for the right Wernicke area appears to highlight frequencies that differ from those identified for the same area in the left hemisphere.

In short, these findings demonstrate that inferred IRNNs serve as powerful proxies for characterizing the unique dynamics of subjects and brain areas, capturing their specific dynamical footprints. By revealing distinct frequency patterns and enhancing our understanding of neural variability, this approach paves the way for deeper insights into the intricate workings of the brain and its functional architecture.

Discussion

The successful application of linear recurrent neural networks within the reservoir computing framework represents a significant methodological contribution to model resting-state fMRI data. Our results show that IRNNs effectively capture brain dynamics, suggesting that resting-state activity can be approximated as a linear system in a high-dimensional space, supporting recent studies advocating for linear models in this context [39]. This is consistent with the hypothesis of a resting activity fluctuating around a fixed point poised at the edge of instability [40, 41]. Despite its apparent complexity, it can be effectively studied relying on simple approaches.

The presented method provides valuable insights into the functional connectivity structure as similarity patterns between vectors associated with the eigenmode decomposition of the inferred IRNN. This factorization could move the attention to the modes referential frame focusing the voxel-wise analysis on the shared and coherent dynamical components among the brain areas.

Notably, the mode dynamics described in Eq. (5) provides an equivalent linear representation for the observables of the system under investigation. The resulting vector of all time derivatives (see Methods) serves as an alternative descriptor of the system's state [42], potentially encapsulating the trajectory evolution at any given time. This concept is reminiscent of the dynamic mode decomposition (DMD) with Hankel matrix [43, 44], but with a significantly reduced dimensionality constraint. Unlike the DMD approach, which requires estimating a number of parameters that scales quadratically with the number M of observables – specifically $(M \times T)^2$ parameters (where T denotes the number of delayed time samples) – our reservoir computing method only necessitates inferring N readout weights, corresponding to the size of the network. This characteristic makes the presented RC approach particularly advantageous and flexible.

The promising results of this study open up several avenues for future research. Applying this modeling framework to pathological brain states could provide valuable insights into the alterations in functional connectiv-

ity and dynamics associated with various neurological disorders, potentially enhancing diagnosis and treatment strategies. For instance, the emergence of unstable modes may be linked to transitions into pathological states. By stimulating the system to suppress these "pathology-related" modes, we may gain a means to intervene and mitigate the occurrence of such transitions.

In line with this, as in a classic digital twin, the inferred IRNN can incorporate any additional available information about the system simply as new input. An external stimulation can potentially act as a selector for different dynamical regimes, allowing for two distinct representations with a single network. The functional connectivity itself could be a proxy for state transition. Indeed, a significant change in FC might indicate that the decomposition has also changed, suggesting that the observed system is close to a different equilibrium point.

While this study demonstrates the efficacy of IRNNs in modeling resting-state brain activity, several limitations must be acknowledged. The assumption of linearity, although supported by our results, may not fully capture the complexity of brain dynamics under all conditions. Nonlinear phenomena, transient states, and the influence of external stimuli could require more sophisticated modeling approaches. Our analysis utilized fMRI data from healthy subjects as a test bed. Future studies should validate the model's performance across diverse datasets, including those from patients with neurological disorders, to assess its robustness and clinical utility.

We believe that our findings can represent a significant advancement in modeling resting-state brain activity, providing a new tool for understanding the brain's intrinsic dynamics. Importantly, the implications of this work extend beyond neuroscience, highlighting the potential for reservoir computing with IRNNs to analyze other complex systems across various fields.

Methods

Dataset description

Subjects. The study included 20 healthy, right-handed subjects (mean age $\pm$ std = 37 ± 12; 7 females, 13 males) with no history of neurological disorders. The study was approved by the Institutional Review Board at Memorial Sloan Kettering Cancer Center, and informed consent was obtained from each participant. During the resting-state condition, subjects were instructed to lie in the scanner, keep their eyes open, try to think of nothing in particular, and maintain fixation on a central cross on the screen.

MRI methods. The MRI data were acquired using a GE 3T scanner (General Electric, Milwaukee, Wisconsin, USA) and a standard quadrature head coil. Functional images covering the whole brain were obtained using a (T2*)-weighted imaging sequence sensitive to the BOLD signal (repetition time, TR/TE = 2500/40 ms; slice thickness = 4.5 mm; matrix = 128 × 128; FOV = 240 mm; volumes = 160). Anatomical T1-weighted images (TR/TE = 600/8 ms; slice thickness = 4.5 mm) were also acquired for co-registration purposes.

Data preprocessing. The functional MRI data were preprocessed and analyzed using the Analysis of Functional NeuroImages (AFNI) software [45]. Head motion correction was performed using 3D rigid-body registration, with the first volume selected as the reference. The motion profile was saved and regressed out. Spatial smoothing was applied using a Gaussian filter with a 4 mm full width of half maximum to improve the signal-to-noise ratio. Corrections for linear trend and high-frequency noise were also applied. The components correlated with cerebrospinal fluid (CSF) and white matter (WM) were removed to address potential sources of artifacts and noise. The data were detrended, demeaned, and band-pass filtered (frequency range 0.01-0.1 Hz). Finally, all fMRI data were registered to the standard Montreal Neurological Institute (MNI152) space.

Areas and voxels definition. The regions of interest (ROIs) considered in this study are areas implicated in language processing. Specifically, these include the Supplementary Motor Area (SMA), the Premotor Area (left and right), the Supramarginal Gyrus (left and right), Broca's Area (left and right), the Angular Gyrus (left and right), and Wernicke's Area (left and right). BOLD time series data were extracted from spheres with a radius of 6 mm, centered on target MNI152 coordinates corresponding to each ROI, resulting in a total of 1311 voxel signals.

fMRI dimensionality analysis

Principal Component Analysis (PCA) is a widely used dimensionality reduction technique used to transform a dataset into a lower-dimensional space while preserving most of the information in the original data [46]. In the context of the given problem, PCA is applied to the set of BOLD signals to reduce redundancy and keep only the relevant information.

The PCA-based dimensional reduction of BOLD is carried out based on the initial 280 s (learning period) of each subject's data. Referring to Fig. 2, the number of principal components taken into account are those capturing 99.5% of the variance in the time series of each area. The resulting transformation is represented by a block matrix $\boldsymbol{P}_0$, where each block's columns $\boldsymbol{p}_i$ correspond to the eigenvectors of the covariance matrix of the respective cortical area:

$$\boldsymbol{P}_0 = \begin{bmatrix} \boldsymbol{p}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{p}_{11} \end{bmatrix} \tag{10}$$

The output undergoes orthogonalization through an additional PCA step ($\boldsymbol{P}_1$) utilizing all components. Subsequently, normalization is performed by applying the diagonal matrix $\boldsymbol{N}$, which contains the reciprocals of the maximum values of the compressed signals observed during the learning period. This approach ensures that all currents enter the reservoir with the same maximum strength. The final transformation matrix $\boldsymbol{P}$ is computed as:

$$\boldsymbol{P} = \boldsymbol{N}\boldsymbol{P}_1\boldsymbol{P}_0 \tag{11}$$

While the antitransformation matrix $\boldsymbol{P}^{-1}$ is calculated as:

$$\boldsymbol{P}^{-1} = \boldsymbol{P}_0^T\boldsymbol{P}_1^T\boldsymbol{N}^{-1} \tag{12}$$

The data is effectively reduced and then restored, ensuring the preservation of essential information while reducing dimensionality for the analysis. The decomposition for voxel activity is derived by applying $\boldsymbol{P}^{-1}$ to the decomposition matrix for the actual network inputs.

Koopman Operator

The discrete-time Koopman operator $\mathcal{K}^1 : \mathcal{F} \to \mathcal{F}$ is a linear operator defined in an infinite-dimensional space $\mathcal{F}$ of observables of the system state [47, 7, 8]. The operation of the Koopman operator on an observable g is described by the equation:

$$\mathcal{K}^1 g(\boldsymbol{x}_t) = g(\boldsymbol{x}_{t+1})\,. \tag{13}$$

Here, $\boldsymbol{x}$ represents a point in the state space. The eigenfunctions of the Koopman operator are potential observables of the system themselves, and have the peculiar property of evolving linearly in time:

$$\mathcal{K}^1\psi_k(\boldsymbol{x}_t) = \psi_k(\boldsymbol{x}_{t+1}) = \lambda_k^1\psi_k(\boldsymbol{x}_t)\,. \tag{14}$$

These eigenfunctions $\psi_k(\boldsymbol{x}_t)$ constitute a basis of the space $\mathcal{F}$, enabling the representation of all the possible observables of the system as a linear combination of linearly evolving functions.

In general there exists a family of Koopman operators, $\mathcal{K}^t$, that advances a function forward by a time t:

$$\mathcal{K}^t g(\boldsymbol{x}) = g(\mathcal{S}^t(\boldsymbol{x}))\,. \tag{15}$$

Here, $\mathcal{S}^t$ denotes the time flow operator, with $\boldsymbol{x}(t) = \mathcal{S}^t[\boldsymbol{x}(0)]$. The infinitesimal generator of the Koopman operator family $\mathcal{L}g = \lim_{t\to 0}\frac{\mathcal{K}^t g - g}{t}$ is the Lie operator which evaluates the temporal change of observables:

$$\mathcal{L}g(\boldsymbol{x}(t)) = \frac{d}{dt}g(\boldsymbol{x}(t)) \tag{16}$$

The family of Koopman operators can then be expressed in this term $\mathcal{K}^t = e^{\mathcal{L}t}$.

Reservoir computing

The reservoir computing approach involves a RNN with fixed and random internal couplings (namely the 'reservoir') whose state is fed forward to a second set of 'readout' units [27, 11]. In this framework the required 'echo-state property' spontaneously emerges from the RNN collective dynamics. The units composing the RNN are intended to model homogeneous and local neuronal assemblies of a cortical network [48, 49, 50]. Each of the N units in the network has activity state $r_j(t)$ evolving in time as

$$\tau \dot{r_j} = \Phi(h_j) - r_j \tag{17}$$

with $j \in [1, N] \subset \mathbb{Z}$. The decay time constant τ and the sigmoidal activation function $\Phi(h_j)$ is the same for all units. Here we consider the limiting case of weak recurrent coupling where the activation function can be approximated to a linear function: $\Phi(h_j) = h_j$. The synaptic input h_j is the weighted sum

$$h_j(t) = \sum_{k=1}^{N} W_{jk} r_k(t) + h_j^{\text{in}}(t)$$

where W_{jk} are the elements of the synaptic matrix (i.e., the internal couplings) $\boldsymbol{W} \in \mathbb{R}^{N\times N}$, and $h_j^{\text{in}}(t)$ is the external input received by the unit j. In reservoir compunting, the external input is driven by be the measured observable ω of the inspected system, which is fed into the network using random synaptic couplings: $h_j^{\text{in}}(t) = W_j^{\text{in}}\,\omega(t)$. Due to the high dimensionality of the trajectories in the RNN state space, a simple linear transformation represented by $\boldsymbol{W}^{\text{out}}$ is often sufficient to accurately map the RNN state into a desired output time series. More precisely, given $\boldsymbol{R} \in \mathbb{R}^{N\times T}$ whose k-th row represent the time series of the k-th unit during a learning period lasting T time steps, and $\boldsymbol{\Omega} \in \mathbb{R}^{M\times T}$ has rows given by the M observables $\omega_k(t)$ to be replicated, the linear map is computed by a ridge regression:

$$\boldsymbol{W}^{\text{out}} = \boldsymbol{\Omega}\boldsymbol{R}^T(\boldsymbol{R}\boldsymbol{R}^T + \beta\boldsymbol{I})^{-1} \tag{18}$$

where β is a regularization parameter and $\boldsymbol{I} \in \mathbb{R}^{N\times N}$ the identity matrix. If the target output is the external input itself $\boldsymbol{\omega}(t)$, the resulting linear map can be used to simulate the external stimulation and predict the future steps of the data. This is equivalent to update the synaptic matrix of the reservoir. In particular $\boldsymbol{W}^{\text{in}}\boldsymbol{\omega}(t) \approx \boldsymbol{W}^{\text{in}}\boldsymbol{W}^{\text{out}}\boldsymbol{r}(t)$ leading to the autonomous system:

$$\tau\dot{\boldsymbol{r}} = \boldsymbol{\Phi}\left[\tilde{\boldsymbol{W}}\boldsymbol{r}\right] - \boldsymbol{r}\,, \tag{19}$$

where $\tilde{\boldsymbol{W}} = \boldsymbol{W} + \boldsymbol{W}^{\text{in}}\boldsymbol{W}^{\text{out}}$ is the updated synaptic matrix.

Linear recurrent neural networks (IRNN) and autoregressive models

In the established framework of reservoir computing with linear RNNs (IRNN, i.e., with a linear activation function) by setting as initial condition $r_k(-\infty) = 0$, the system Eq. (17) has the following solution:

$$\boldsymbol{r}(t) = \sum_{k=1}^{M} \int_{-\infty}^{t} \omega_k(t') e^{(\boldsymbol{W}-\boldsymbol{I})(t-t')/\tau} \boldsymbol{W}_k^{\text{in}} dt' \tag{20}$$

As above $\omega_k(t)$ denotes the k-observable measured from the inspected system. Assuming a time t^* exists after which a matrix $\boldsymbol{W}^{\text{out}}$ maps the network state to the input data $\boldsymbol{\omega}(t) = \boldsymbol{W}^{\text{out}}\boldsymbol{r}(t)$, we can rewrite the above equation as:

$$\boldsymbol{\omega}(t) = \sum_{k=1}^{M} \int_{-\infty}^{t} \omega_k(t') \boldsymbol{W}^{\text{out}} e^{(\boldsymbol{W}-\boldsymbol{I})(t-t')/\tau} \boldsymbol{W}_k^{\text{in}} dt' \tag{21}$$

By substituting $s = t - t'$, we obtain an autoregressive description of the signals:

$$\omega_j(t) = \sum_{k=1}^{M} \int_{0}^{\infty} G_{jk}(s)\omega_k(t-s) ds \tag{22}$$

Here, $G_{jk}(s) = \boldsymbol{W}_j^{\text{out}} e^{(\boldsymbol{W}-\boldsymbol{I})s/\tau} \boldsymbol{W}_k^{\text{in}}$. These kernels can be expressed in terms of the synaptic matrix spectrum. The diagonal representation of the exponential matrix is given by:

$$e^{(\boldsymbol{W}-\boldsymbol{I})s/\tau} = \sum_{n=1}^{N} e^{(\lambda_n - 1)/\tau} \left|Q_n\right\rangle \left\langle Q_n^{-1}\right| . \tag{23}$$

In this representation, λ_n denotes the eigenvalues and $|Q_n\rangle\langle Q_n^{-1}|$ the outer product between the right and left eigenvectors of $\boldsymbol{W}$. This representation allows us to rewrite the kernel as the weighted sum over exponentially decaying components:

$$G_{jk}(s) = \sum_{n=1}^{N} J_{jk}^{(n)} e^{(\lambda_n - 1)s/\tau} , \tag{24}$$

where the tensor of weights is defined by the set of rank-1 matrices $\boldsymbol{J}_n = \boldsymbol{W}^{\text{out}} \left|Q_n\right\rangle \left\langle Q_n^{-1}\right| \boldsymbol{W}^{\text{in}}$.

IRNN and the Laplace transform of system state

Assuming the existence of a time t^* such that a matrix $\boldsymbol{W}^{\text{out}}$ maps the IRNN state into the system observable given as input $\omega(t) = \boldsymbol{W}^{\text{out}} \boldsymbol{r}(t)$, the open- and closed-loop dynamics are in principle equivalent [27, 26]. This property allows us to link the kernel of the autoregressive model (closed-loop dynamics) to the spectrum of the learned synaptic matrix $\tilde{\boldsymbol{W}}$, and consequently, to the resulting finite Koopman approximation, as we will see in the following.

Open-loop Representation

The autoregressive formula expressed in Eq. (3) can be reformulated as:

$$\omega_j(t) = \sum_{k=1}^{M} \int_0^t G_{jk}(t-t')\omega_k(t')dt' + \omega_j(0) . \tag{25}$$

The convolution theorem for the Laplace transform leads to:

$$\text{LT}\{\omega_j\}(\sigma) = \sum_{k=1}^{M} \text{LT}\{G_{jk}\}(\sigma)\text{LT}\{\omega_k\}(\sigma) + \frac{\omega_j(0)}{\sigma} \tag{26}$$

Here, $\text{LT}\{G_{jk}\} = \int_0^\infty G_{jk}(t)e^{-\sigma t}dt$ denotes the Laplace transform of the kernel that is

$$\text{LT}\{G_{jk}\}(\sigma) = \sum_{n=1}^{N} \frac{J_{jk}^{(n)}}{\sigma - \frac{\lambda_n - 1}{\tau}} . \tag{27}$$

Thus, in matrix formalism the Laplace transforms of the system observables result to be

$$\text{LT}\{\boldsymbol{\omega}\}(\sigma) = \frac{\boldsymbol{L}^{-1}(\sigma)}{\sigma}\boldsymbol{\omega}(0) , \tag{28}$$

where the matrix $\boldsymbol{L}(\sigma) \in \mathbb{R}^{M\times M}$ has elements $L_{ij}(\sigma) = \delta_{ij} - \text{LT}\{G_{ij}\}(\sigma)$.

Closed-loop representation

The closed-loop network evolves according to Eq. (4). Assuming a diagonalizable synaptic matrix $\tilde{\boldsymbol{W}}$, it can be factorized as $\tilde{\boldsymbol{W}} = \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{Q}}^{-1}$, where $\tilde{\boldsymbol{Q}}$ is a matrix whose columns are the right eigenvectors of $\tilde{\boldsymbol{W}}$ and $\tilde{\boldsymbol{\Lambda}}$ is a diagonal matrix containing the associated eigenvalues. The spectral decomposition of the learned synaptic matrix leads to a decoupled set of linearly evolving projections $\boldsymbol{v}(t) = \tilde{\boldsymbol{Q}}^{-1}\boldsymbol{r}(t)$ whose N elements evolves independently as

$$v_n(t) = v_n(0)e^{(\tilde{\lambda}_n - 1)t/\tau} . \tag{29}$$

These variables serve as a basis to decompose the system observables $\omega_k(t)$ as the sum of linearly evolving modes, resembling the Koopman decomposition. Indeed, given $\boldsymbol{r} = \tilde{\boldsymbol{Q}}\boldsymbol{v}$ we can write

$$\omega_j(t) = \sum_{n=1}^{N} \Xi_{jn} v_n(t) = \sum_{n=1}^{N} \Xi_{jn} v_n(0) e^{(\tilde{\lambda}_n - 1)t/\tau}\,, \tag{30}$$

with matrix $\boldsymbol{\Xi} \in \mathbb{R}^{M\times N}$ defined as $\boldsymbol{\Xi} = \boldsymbol{W}^{\text{out}}\boldsymbol{Q}$ such that $\boldsymbol{\omega}(t) = \boldsymbol{\Xi}\boldsymbol{v}(t)$. The Laplace transform of the replicated observables is then

$$\hat{\omega}_j(\sigma) = \sum_{n=1}^{N} \frac{\Xi_{jn} v_n(0)}{\sigma - \frac{(\tilde{\lambda}_n - 1)}{\tau}} \tag{31}$$

whose poles coincide with the eigenvalues of the estimated Koopman matrix.

Given the equivalence between the open and closed representations, $L_{ij}(\sigma)/\sigma$ shares the same poles, providing a link between the autoregressive modeling and the linear dynamic representation.

IRNNs and the spectrum of the Koopman operator

The closed-loop representation provides a decomposition of the signal as linear evolving modes. We can explicitly express the functional form of the basis starting from open loop dynamics. By applying $\tilde{\boldsymbol{Q}}^{-1}$ to both hand sides of Eq. (2) yields:

$$\boldsymbol{v}(t) = \sum_{k=1}^{M} \int_0^\infty \omega_k(t-s) \tilde{\boldsymbol{Q}}^{-1} e^{(\boldsymbol{W}-\boldsymbol{I})s/\tau} \boldsymbol{W}_k^{\text{in}} ds\,. \tag{32}$$

By resorting to the spectral decomposition in Eq. (23), the above eigenmode dynamics reduces to:

$$v_j(t) = \sum_{k=1}^{M}\sum_{n=1}^{N} \left[\tilde{\boldsymbol{Q}}^{-1} \left|Q_n\right\rangle \left\langle Q_n^{-1}\right| \boldsymbol{W}^{\text{in}}\right]_{jk} \int_0^\infty \omega_k(t-s) e^{(\lambda_n - 1)s/\tau} ds\,. \tag{33}$$

We can then expand in Taylor series the observables: $\omega_k(t-s) = \sum_{r=0}^{\infty} \frac{d^r \omega_k(t)}{dt^r}(-1)^r s^r / r!$. This allows to solve the above integral leading to a functional description of the estimated eigenmodes

$$v_j(t) = \psi_j\left[\boldsymbol{\omega}(t), \dot{\boldsymbol{\omega}}(t), ..\right] = \sum_{n=1}^{N}\sum_{k=1}^{M} A_{jk}^{(n)} \sum_{r=0}^{\infty} \frac{d^r \omega_k(t)}{dt^r} \left(\frac{\tau}{\lambda_n - 1}\right)^r\,. \tag{34}$$

The vector of all the derivatives $\frac{d^r \omega_k(t)}{dt^r}$ fully determine the state of the observed system [29] and ψ_j is the eigenfunction of the approximated Koopman operator depending on the full state of the system. The corresponding eigenvalue $\sigma_j = (\tilde{\lambda}_j - 1)/\tau$, and $A_{jk}^{(n)}$ is a tensor of complex weights defined by the rank-1 matrices $\boldsymbol{A}_n = \frac{\tau}{1-\lambda_n} \tilde{\boldsymbol{Q}}^{-1} \left|Q_n\right\rangle \left\langle Q_n^{-1}\right| \boldsymbol{W}^{\text{in}}$.

Stochastic IRNN

The deterministic neural network described can be extended to the stochastic case, where a white noise input stimulates the reservoir, maintaining it out of equilibrium as a continuous source of new energy. Each unit receives a total input defined by the equation:

$$h_j(t) = \sum_{k=1}^{N} W_{jk} r_k(t) + h_j^{\text{in}}(t) + \eta_j(t) \tag{35}$$

where $\eta_j(t)$ represents an Ornstein-Uhlenbeck process (colored noise) with a relatively small correlation time such that for the purpose of this work it can be considered as memory-less: $\langle \eta_i(t)\eta_j(t')\rangle = \gamma^2 \delta_{ij}\delta(t-t')$. An optimal standard deviation γ can be identified to maximize the accuracy of the linear map $\boldsymbol{W}^{\text{out}}$.

In the closed-loop formulation, the input can be expressed as:

$$h_j(t) = \sum_{k=1}^{N} \tilde{W}_{jk} r_k(t) + \eta_j(t) \tag{36}$$

Consequently, the associated Langevin equation for the reservoir is given by:

$$\tau \frac{d\boldsymbol{r}}{dt} = -\boldsymbol{r} + \tilde{\boldsymbol{W}}\, \boldsymbol{r} + \boldsymbol{\eta} \tag{37}$$

In this context, the introduced autoregressive description and linear decomposition remain valid when considering the expected values of the system.

By applying the linear map $\boldsymbol{W}^{\mathrm{out}}$, we derive the resulting stochastic differential equation that describes the signal under analysis:

$$\tau \frac{d\boldsymbol{\omega}}{dt} = -\boldsymbol{\omega} + \boldsymbol{W}^{\mathrm{out}} \tilde{\boldsymbol{W}} \boldsymbol{r} + \boldsymbol{W}^{\mathrm{out}} \boldsymbol{\eta} \tag{38}$$

This modeling approach incorporates realistic scenarios involving common sources of noise η that are shared among the components of the system, as dictated by the matrix $\boldsymbol{W}^{\mathrm{out}}$.

Functional Connectivity and Functional Connectivity Dynamics

Functional Connectivity (FC) is a measure of the degree of co-activation over time of different brain regions. Although FC does not allow to infer the directional flow of information among the nodes of the brain network, it has proven to provide valuable insights into the functional organization of the central nervous system [51]. Mathematically, the FC between two brain regions labeled A and B, is usually measured as the correlation function

$$\mathrm{FC}_{AB}(t;T) = \mathrm{cor}[X_A(t;T), X_B(t;T)]\,, \tag{39}$$

where cor is the Pearson correlation coefficient, and $X_A(t;T)$ and $X_B(t;T)$ are the BOLD signals from region A and B, respectively, in the time window $[t, t+T]$.

Temporal changes of functional connectivity are evaluated by computing the Functional Connectivity Dynamics (FCD) [52, 53]. FCD is defined as the correlation between vectorized FC matrices evaluated in time-shifted chunks of BOLD activity:

$$\mathrm{FCD}_{A,B}(\delta t, t; T) = \mathrm{ucor}[\mathrm{FC}_A(t;T), \mathrm{FC}_B(t+\delta t;T)]\,, \tag{40}$$

where δt is a constant time shift, and ucor is the correlation between the vectors composed of the upper-diagonal elements of the two FC matrices involved.

Throughout the whole paper FC is computed with $t = 0$ s and $T = 310$ s, while FCD uses $T = 60$ s time shift δt multiples of 10 seconds.

Functional Connectivity in IRNN

Once a IRNN capable to reproduce the systems observables $\omega_j(t)$ is available, the elements of the FC matrix can be inferred directly from the IRNN parameters. The computed Functional Connectivity of the simulated signals depends on the covariance matrix estimated from time t over a time window of length T. Considering large time windows T and assuming a zero time average $\langle \omega_j(t) \rangle_t \simeq 0$, it holds:

$$\mathrm{cov}_{ij}(t;T) = \int_t^{t+T} \omega_i^*(t')\omega_j(t')dt' = \sum_{r,s} \Xi_{ir}^* \Xi_{js} \int_t^{t+T} v_r^*(t') v_s(t') dt' = \sum_{r,s} \Xi_{ir}^* \Xi_{js} \Omega_{rs}(t;T) \tag{41}$$

Here we are making use of the spectral decomposition in Eq. (6) and we define $\Omega_{rs}(t;T)$ as the estimated covariance between the OU modes v_r, v_s. From the simulations, it results that only a few slow uncorrelated modes have a significant weight Ξ in the decomposition. In this scenario, the covariance matrix is well approximated by the dot product between the representational vectors in the functional space:

$$\mathrm{cov}_{ij}(t;T) \approx \boldsymbol{\Xi}_i \cdot \boldsymbol{\Xi}_j \tag{42}$$

Consequently for the correlation matrix it holds

$$\mathrm{FC}_{ij}(t;T) \approx \mathrm{S_C}\left(\boldsymbol{\Xi}_i, \boldsymbol{\Xi}_j\right) \tag{43}$$

where $\mathrm{S_C}$ indicates the cosine similarity function, i.e., the Pearson correlation between row and column vectors composing $\boldsymbol{\Xi}^*$ and $\boldsymbol{\Xi}$, respectively, under the assumption of a zero mean for the system observables.

Simulation parameters

Network parameters have been selected through grid search methods to optimize prediction accuracy. The reservoir state is initialized in a zero-activity state and is disrupted from equilibrium by the input for a transient period of $20\,s$ before learning. This period corresponds to the t^* mentioned in the preceding paragraphs. The input signal is linearly interpolated, and the system's dynamics are resolved using the Euler–Maruyama method with a time step that is one-hundredth of the sampling interval. The spectral radius ρ of the synaptic matrix, used for fMRI analysis, is set to 0.5, and the time constant τ is 1.0. The standard deviation g^{in} of the input matrix $\boldsymbol{W}^{\mathrm{in}}$ is set to 1. The thermal noise shown in Fig. 3 has a regularizing effect. Consequently, we do not employ further regularization for the readout regression. For the example system depicted in Fig. 1, the parameters are $\rho = 0.5$, $\tau = 0.1$ with no regularizers, and g^{in} equal to the inverse of the square root of the number of inputs. The stochastic scenario is characterized by thermal noise with a standard deviation of 10^{-5} and ridge regression with a regularization of 10^{-7} . For cases with missing variables, the parameters are set to $\rho = 0.6$, $\tau = 0.25$.

Clustering analysis

The clustering analysis is performed on the resulting spectrum of the data by discretizing the complex space into a grid with a real resolution of $0.00025s^{-1}$, and an imaginary resolution of $0.01s^{-1}$.

To further reduce the dimensionality of the data and get an interpretable 3 dimensional representation, Linear Discriminant Analysis (LDA) is employed. LDA is a widely used technique in machine learning [46]. It is a supervised learning method that aims to reduce the dimensionality of the data while preserving the discriminatory information between different classes. The core idea is to find a linear transformation that maximizes the ratio of the between-class scatter $\boldsymbol{S}_b$ to the within-class scatter $\boldsymbol{S}_w$. This optimization process equals to solve the following generalized eigenvalue problem

$$\boldsymbol{S}_b \boldsymbol{l} = \lambda \boldsymbol{S}_w \boldsymbol{l} \tag{44}$$

The optimal projection matrix $\boldsymbol{L}$ to a subspace of dimension k is given by the eigenvectors $\boldsymbol{l}$ corresponding to the largest k eigenvalues.

Acknowledgments

This research has received funding from the Italian National Recovery and Resilience Plan (PNRR), M4C2, funded by the European Union–NextGenerationEU (Project IR0000011, CUP B51E22000150006, "EBRAINS-Italy"), to MM.

References

[1] Christopher M. Bishop. *Pattern recognition and machine learning*. Springer, New York, NY, 2006.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[3] Bernhard Mehlig. *Machine learning with neural networks: an introduction for scientists and engineers*. Cambridge University Press, 2021.

[4] Keshav Srinivasan, Nolan Coble, Joy Hamlin, Thomas Antonsen, Edward Ott, and Michelle Girvan. Parallel machine learning for forecasting the dynamics of complex networks. *Phys. Rev. Lett.*, 128(16):0–4, 2022. doi: $10.1103/\mathrm{PhysRevLett.128.164101}$.

[5] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 417–431. Springer, 2020.

[6] Radwa ElShawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Comput. Intell.*, 37(4):1633–1650, 2021.

[7] Alexandre Mauroy, Y Susuki, and Igor Mezić. *Koopman operator in systems and control*. Springer, 2020.

[8] Steven L. Brunton, Marko Budišić, Eurika Kaiser, and J. Nathan Kutz. Modern koopman theory for dynamical systems. *SIAM Rev.*, 64(2):229–340, 2022. doi: $10.1137/21\mathrm{M}1401243$.

[9] Peter J Schmid. Dynamic mode decomposition and its variants. *Annu. Rev. Fluid Mech.*, 54:225–254, 2022.

[10] Gabriele Di Antonio and Gianni Valerio Vinci. Koopman correlations underlie linear response and causality. *arXiv preprint arXiv:2410.08708*, 2024.

[11] Herbert Jaeger and Harald Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80, 2004.

[12] Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Comput. Sci. Rev.*, 3(3):127–149, 2009.

[13] H B Barlow. Possible principles underlying the transformation of sensory messages. In W A Rosenblith, editor, *Sensory Communication*, pages 217–34. MIT Press, Cambridge, MA, 1961.

[14] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annu. Rev. Neurosci.*, 24(1):1193–1216, 2001. doi: $10.1146/\mathrm{annurev.neuro.24.1.1193}$.

[15] Stefano Fusi, Earl K Miller, and Mattia Rigotti. Why neurons mix: high dimensionality for higher cognition. *Curr. Opin. Neurobiol.*, 37:66–74, 2016. doi: $10.1016/\mathrm{j.conb.2016.01.010}$.

[16] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, 2013.

[17] David Sussillo and O. Barak. Opening the black box: Low-dimensional dynamics in high-dimensional recurrent neural netw. *Neural Computation*, 25:626–649, 2013. doi: $10.1162/\mathrm{NECO_a_00409}$.

[18] Daniel Durstewitz, Georgia Koppe, and Max Ingo Thurm. Reconstructing computational system dynamics from neural data with recurrent neural networks. *Nat. Rev. Neurosci.*, 24(11):693–710, 2023.

[19] Gabriele Di Antonio, Sofia Raglio, and Maurizio Mattia. A geometrical solution underlies general neural principle for serial ordering. *Nat. Commun.*, 15(1):8238, 2024.

[20] Fei Tao, Bin Xiao, Qinglin Qi, Jiangfeng Cheng, and Ping Ji. Digital twin modeling. *J. Manuf. Syst.*, 64: 372–389, 2022.

[21] Eric VanDerHorn and Sankaran Mahadevan. Digital twin: Generalization, characterization and implementation. *Decis. Support Syst.*, 145:113524, 2021.

[22] Michael Schirner, Simon Rothmeier, Viktor K Jirsa, Anthony Randal McIntosh, and Petra Ritter. An automated pipeline for constructing personalized virtual brains from multimodal neuroimaging data. *NeuroImage*, 117: 343–357, 2015.

[23] Viktor Jirsa, Huifang Wang, Paul Triebkorn, Meysam Hashemi, Jayant Jha, Jorge Gonzalez-Martinez, Maxime Guye, Julia Makhalova, and Fabrice Bartolomei. Personalised virtual brain models in epilepsy. *Lancet Neurol.*, 22(5):443–454, 2023. doi: $10.1016/\mathrm{S1474\text{-}4422(23)00008\text{-}X}$.

[24] Lucius Samo Fekonja, Robert Schenk, Emily Schröder, Rosario Tomasello, Samo Tomšič, and Thomas Picht. The digital twin in neuroscience: from theory to tailored therapy. *Front. Neurosci.*, 18:1–8, 2024. doi: $10.3389/\mathrm{fnins.2024.1454856}$.

[25] Ljupco Kocarev and U Parlitz. Generalized synchronization, predictability, and equivalence of unidirectionally coupled dynamical systems. *Phys. Rev. Lett.*, 76(11):1816–9, 1996. doi: 10.1103/PhysRevLett.76.1816.

[26] Zhixin Lu, Brian R Hunt, and Edward Ott. Attractor reconstruction by machine learning. *Chaos*, 28(6), 2018. doi: 10.1063/1.5039508.

[27] Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks. Technical report, German National Research Center for Information Technology, 2001.

[28] Izzet B Yildiz, Herbert Jaeger, and Stefan J Kiebel. Re-visiting the echo state property. *Neural Netw.*, 35: 1–9, 2012. doi: 10.1016/j.neunet.2012.07.005.

[29] Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381, Berlin, Heidelberg, 1981. Springer Berlin Heidelberg. ISBN 9783540389453.

[30] J Stark, D S Broomhead, M E Davies, and J Huke. Delay embeddings for forced systems. II. Stochastic forcing. *J. Nonlinear Sci.*, 13:519–577, 2003. doi: 10.1007/s00332-003-0534-4.

[31] Erik Bollt. On explaining the surprising success of reservoir computing forecaster of chaos? the universal machine learning dynamical system with contrast to var and dmd. *Chaos*, 31(1), 2021.

[32] Daniel J Gauthier, Erik Bollt, Aaron Griffith, and Wendson AS Barbosa. Next generation reservoir computing. *Nat. Commun.*, 12(1):1–8, 2021.

[33] Pietro Verzelli, Cesare Alippi, Lorenzo Livi, and Peter Tiňo. Input-to-state representation in linear reservoirs dynamics. *IEEE Trans. Neural Netw. Learn. Syst.*, 33(9):4598–4609, 2021.

[34] Zheng-Meng Zhai, Ling-Wei Kong, and Ying-Cheng Lai. Emergence of a resonance in machine learning. *Phys. Rev. Res.*, 5(3):033127, 2023.

[35] E. J W Van Someren, Y. D. Van Der Werf, Pieter R Roelfsema, H. D. Mansvelder, and F. H. Lopes da Silva. Slow brain oscillations of sleep, resting state, and vigilance. *Prog. Brain Res.*, 193:3–15, 2011. doi: 10.1016/B978-0-444-53839-0.00001-6.

[36] Sandro Lecci, Laura M J Fernandez, Frederik D Weber, Romain Cardis, Jean Yves Chatton, Jan Born, and Anita Lüthi. Coordinated infraslow neural and cardiac oscillations mark fragility and offline periods in mammalian sleep. *Sci. Adv.*, 3(2), 2017. doi: 10.1126/sciadv.1602026.

[37] Daniel Gutierrez-Barragan, M Albert Basson, Stefano Panzeri, and Alessandro Gozzi. Infraslow state fluctuations govern spontaneous fMRI network dynamics. *Curr. Biol.*, 29(14):2295–2306.e5, 2019. doi: 10.1016/j.cub.2019.06.017.

[38] Núria Tort-Colet, Cristiano Capone, Maria V Sanchez-Vives, and Maurizio Mattia. Attractor competition enriches cortical dynamics during awakening from anesthesia. *Cell Rep.*, 35(12):109270, 2021. doi: 10.1016/j.celrep.2021.109270.

[39] Erfan Nozari, Maxwell A Bertolero, Jennifer Stiso, Lorenzo Caciagli, Eli J Cornblath, Xiaosong He, Arun S Mahadevan, George J Pappas, and Dani S Bassett. Macroscopic resting-state brain dynamics are best described by linear models. *Nat. Biomed. Eng.*, pages 1–17, 2023.

[40] Gustavo Deco and Viktor K Jirsa. Ongoing cortical activity at rest: criticality, multistability, and ghost attractors. *J. Neurosci.*, 32(10):3366–75, 2012. doi: 10.1523/JNEUROSCI.2523-11.2012.

[41] Gustavo Deco, Adrián Ponce-Alvarez, Dante Mantini, Gian Luca Romani, Patric Hagmann, and Maurizio Corbetta. Resting-state functional connectivity emerges from structurally and dynamically shaped slow linear fluctuations. *J. Neurosci.*, 33(27):11239–52, 2013. doi: 10.1523/JNEUROSCI.1091-13.2013.

[42] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*. Cambridge university press, 2003.

[43] Hassan Arbabi and Igor Mezic. Ergodic theory, dynamic mode decomposition, and computation of spectral properties of the koopman operator. *SIAM J. Appl. Dyn. Syst.*, 16(4):2096–2126, 2017.

[44] Steven L Brunton, Bingni W Brunton, Joshua L Proctor, Eurika Kaiser, and J Nathan Kutz. Chaos as an intermittently forced linear system. *Nat. Commun.*, 8(1):19, 2017.

[45] Robert W Cox. AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.*, 29(29):162–173, 1996. doi: 10.1006/cbmr.1996.0014.

[46] Richard O Duda, Peter E Hart, and David D Stork. *Pattern classification*. John Wiley & Sons, New York, NY, 2nd edition edition, 2001.

[47] Bernard O Koopman. Hamiltonian systems and transformation in hilbert space. *Proc. Natl. Acad. Sci. U.S.A.*, 17(5):315–318, 1931.

[48] Alessandro Treves. Mean-field analysis of neuronal spike dynamics. *Network*, 4(3):259–84, 1993.

[49] Bruce W Knight. Dynamics of encoding in neuron populations: some general mathematical features. *Neural Comput.*, 12(3):473–518, 2000.

[50] Maurizio Mattia and Paolo Del Giudice. Population dynamics of interacting spiking neurons. *Phys. Rev. E*, 66(5 Pt 1):051917, 2002. doi: 10.1103/PhysRevE.66.051917.

[51] Edward T Bullmore and Olaf Sporns. The economy of brain network organization. *Nat. Rev. Neurosci.*, 13 (5):336–349, 2012. doi: 10.1038/nrn3214.

[52] Enrique C A Hansen, Demian Battaglia, Andreas Spiegler, Gustavo Deco, and Viktor K Jirsa. Functional connectivity dynamics: Modeling the switching behavior of the resting state. *NeuroImage*, 105:525–535, 2015. doi: 10.1016/j.neuroimage.2014.11.001.

[53] Gustavo Deco, Morten L Kringelbach, Viktor K Jirsa, and Petra Ritter. The dynamics of resting fluctuations in the brain: Metastability and its dynamical cortical core. *Sci. Rep.*, 7(1):1–14, 2017. doi: 10.1038/s41598-017-03073-5.