The attractor states of the functional brain connectome

Robert Englert
University Medicine Essen

Tamas Spisak[1]
University Medicine Essen

Friday 14th July, 2023

Abstract

Abstract:
todo

Keywords

Key Points:

- We propose a high-level computational model of "activity flow" across brain regions
- The model considers the funcional brain connectome as an already-trained Hopfield neural network
- It defines an energy level for any arbitrary brain activation patterns
- and a trajectory towards one of the finite number of stable patterns (attractor states) that minimize this energy
- The model reproduces and explains the dynamic repertoir of the brain's spontanous activity at rest
- It conceptualizes both task-induced and pathological changes in brain activity as a shift on the "attractor landscape"
- We validate our findings on healthy and clinical samples (~2000 participants)

1 Introduction

Brain function is characterized by the continuous activation and deactivation of anatomically distributed neuronal populations. While the focus of related research is often on the direct mapping between changes in the activity of a single brain area and a specific task or condition, in reality, regional activation never seems to occur in isolation (). Regardless of the presence or absence of explicit stimuli, brain regions seem to work in concert, resulting in a rich and complex spatiotemporal fluctuation (). This fluctuation is neither random, nor stationary over time ; . It shows quasi-periodic properties (), with a limited number of recurring patterns known as "brain states" (; ; Liu and Duyn [2013]; Richiardi et al. [2011]).

Whole-brain dynamics have previously been characterized with various descriptive techniques (; Vidaurre et al. [2017]; ;), providing accumulating evidence not only for the existence, but also for the high neurobiological and clinical significance, of such dynamics (; ;). However, due to the nature of such studies, the underlying driving forces remain elusive.

Questions regarding the mechanisms that cause these remarkable dynamics can be addressed through computational models, which have the potential to shift our understanding from mere associations to causal explanations. Conventional computational approaches try to solve the puzzle by delving all the way down to

[1]Correspondence to: tamas.spisak@uk-essen.de

the biophysical properties of single neurons and then aim to construct a model of larger neural populations, or even the entire brain (). While such approaches have demonstrated numerous successful applications (; Heinz et al. [2018]), the estimation of all the free parameters in such models presents a grand challenge. This limitation hampers the ability of these techniques to effectively bridge the gap between explanations at the level of single neurons and the complexity of behavior ().

An alternative approach, known as "neuroconnectomism" (Doerig et al. [2023]) shifts the emphasis from "biophysical fidelity" of models to "cognitive/behavioral fidelity" (), by using artificial neural networks (ANNs) that were trained to perform various tasks, as brain models. While this novel approach has already significantly contributed to expanding our understanding of the general computational principles of the brain (see), the requirement of training ANNs for specific tasks poses inherent limitations in their capacity to explain the spontaneous macro-scale dynamics of neural activity (Richards et al. [2019]).

In this work, we adopt a middle ground between traditional computational modeling and neuroconnectionism to investigate the phenomenon of brain dynamics. On one hand, similar to neuroconnectionism, our objective is not to achieve a comprehensive bottom-up understanding of neural mechanisms. Instead, we utilize an artificial neural network (ANN) as a high-level computational model of the brain (Figure 1A). On the other hand, we do not train our ANN for a specific task. Instead, we empirically set its weights based on data about the "activity flow" (;) across regions within the functional brain connectome, as measured with functional magnetic resonance imaging (fMRI, Figure 1B). We employ a neurobiologically motivated ANN architecture, a continuous-space Hopfield network (;).

Within this architecture, the topology of the functional connectome naturally defines an energy level for any arbitrary activation patterns and a trajectory towards one of the finite number of stable patterns that minimize this energy, the so-called attractor states. Our model also offers a natural explanation for brain state dynamics. In the presence of weak noise, the system does not converge into an equilibrium state but undergoes "bifurcation", enabling it to traverse extensive regions of the state space, moving on a path restricted by the "gravitational pull" of different attractor states (Figure 1C).

In this simplistic yet powerful framework, both spontaneous and task-induced brain dynamics can be conceptualized as a high-dimensional path that meanders on the reconstructed energy landscape in a way that is restricted by the "gravitational pull" of the attractors states. The framework provides a generative model for both resting state and task-related brain dynamics, offering novel perspectives on the mechanistic origins of resting state brain states and task-based activation maps.

In the present work, we first explore the attractor states of the functional brain connectome and construct a low-dimensional representation of the energy landscape. Subsequently, we rigorously test the proposed model through a series of experiments conducted on data obtained from 7 studies encompassing a total of n≈2000 individuals.

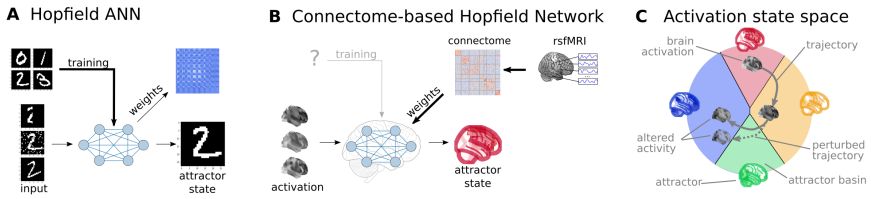
These analyses include evaluation of robustness and replicability, testing the model's ability to reconstrcut various characteristics of resting state brain dynamics as well as its capacity to detect and explain changes induced by tasks or pathological conditions.

These experiments provide converging evidence for the validity of connectome-based Hopfield networks as models of brain dynamics and highlight their potential to provide a fresh perspective on a wide range of research questions in basic and translational neuroscience.

2 Results

2.1 Connectome-based Hopfield network as a model of brain dynamics

First, we explored the attractor states of the functional brain connectome in a sample of n=41 healthy young participants (Table **??**). We estimated interregional activity flow (Cole et al. [2016];) as the study-level average of regularized partial correlations among the resting state fMRI timeseries of m = 122 functionally defined brain regions (BASC brain atlas, see Methods for details). We then used the standardized functional connectome as the w_{ij} weights of a continuous-state Hopfield network (,) consisting of m neural units, each having an activity $a_i \in [-1, 1]$. Hopfield networks can be initialized by an arbitrary activation pattern (m activations) and iteratively updated, until convergence ("relaxation"), according to the following equation:



A Hopfield artificial neural networks (ANNs) are a form of recurrent ANNs that serve as content-addressable ("associative") memory systems. Hopfield networks can be trained to store a finite number of patterns (e.g. via Hebbian learning). During the training procedure, the weights of the Hopfield ANN are trained so that the stored patterns become stable attractor states of the network. Thus, when the trained network is presented partial or noisy variations of the stored patterns, it can effectively reconstruct the original pattern via an iterative relaxation procedure that converges to the attractor states. **B** We consider regions of the brain as nodes of a Hopfield network. Instead of training the Hopfield network to specific tasks, we use the set its weights empirically, with the interregional activity flow estimated via functional brain connectivity. Following form the strong analogies between the relaxation rule of Hopfield networks and the activity flow principle that links activity to connectivity in brain networks, we propose the constructed connectome-based Hopfield (CBH) network as a computational model for macro-scale brain dynamics. **C** The proposed computational framework assigns an energy level, an attractor state and a position in a low-dimensional embedding to brain activation patterns. Additionally, it models how the whole state-space of viable activation patterns is restricted by the dynamics of the system how alterations in activity and/or connectivity modify these dynamics.

A Hopfield artificial neural networks (ANNs) are a form of recurrent ANNs that serve as content-addressable ("associative") memory systems. Hopfield networks can be trained to store a finite number of patterns (e.g. via Hebbian learning). During the training procedure, the weights of the Hopfield ANN are trained so that the stored patterns become stable attractor states of the network. Thus, when the trained network is presented partial or noisy variations of the stored patterns, it can effectively reconstruct the original pattern via an iterative relaxation procedure that converges to the attractor states. **B** We consider regions of the brain as nodes of a Hopfield network. Instead of training the Hopfield network to specific tasks, we use the set its weights empirically, with the interregional activity flow estimated via functional brain connectivity. Following form the strong analogies between the relaxation rule of Hopfield networks and the activity flow principle that links activity to connectivity in brain networks, we propose the constructed connectome-based Hopfield (CBH) network as a computational model for macro-scale brain dynamics. **C** The proposed computational framework assigns an energy level, an attractor state and a position in a low-dimensional embedding to brain activation patterns. Additionally, it models how the whole state-space of viable activation patterns is restricted by the dynamics of the system how alterations in activity and/or connectivity modify these dynamics.

Figure 1: **Connectome-based Hopfield networks as models of macro-scale brain dynamics.**

A Hopfield artificial neural networks (ANNs) are a form of recurrent ANNs that serve as content-addressable ("associative") memory systems. Hopfield networks can be trained to store a finite number of patterns (e.g. via Hebbian learning). During the training procedure, the weights of the Hopfield ANN are trained so that the stored patterns become stable attractor states of the network. Thus, when the trained network is presented partial or noisy variations of the stored patterns, it can effectively reconstruct the original pattern via an iterative relaxation procedure that converges to the attractor states. **B** We consider regions of the brain as nodes of a Hopfield network. Instead of training the Hopfield network to specific tasks, we use the set its weights empirically, with the interregional activity flow estimated via functional brain connectivity. Following form the strong analogies between the relaxation rule of Hopfield networks and the activity flow principle that links activity to connectivity in brain networks, we propose the constructed connectome-based Hopfield (CBH) network as a computational model for macro-scale brain dynamics. **C** The proposed computational framework assigns an energy level, an attractor state and a position in a low-dimensional embedding to brain activation patterns. Additionally, it models how the whole state-space of viable activation patterns is restricted by the dynamics of the system how alterations in activity and/or connectivity modify these dynamics.

$$\dot{a}_i = S(\beta \sum_{j=1}^{m} w_{ij} a_j - b_i) \tag{1}$$

where $\dot{a}_i$ is the activity of neural unit i in the next iteration and $S(a_j)$ is the sigmoidal activation function $S(a) = tanh(a)$ and b_i is the bias of unit i and β is the so-called temperature parameter. For the sake of simplicity, we set $b_i = 0$ in all our experiments. Importantly, in our implementation, the relaxation of the Hopfield network can be conceptualized as the repeated application of the activity flow principle, simultaneously for all regions: $\dot{a}_i = \sum_{j=1}^{m} w_{ij} a_j$. The update rule also exhibits strong analogies with the inner workings of neural mass models () as applied e.g. in dynamic causal modelling (see Discussion for more details).

Hopfiled networks assign an energy to every possible activity configurations (see Methods), which decreases during the relaxation procedure until reaching an equilibrium state with minimal energy (Figure 2.1A, top panel, ;). We used a large number of random initializations to obtain all possible attractor states of the connectome-based Hopfield network in Table **??** (Figure 2.1A, bottom panel).

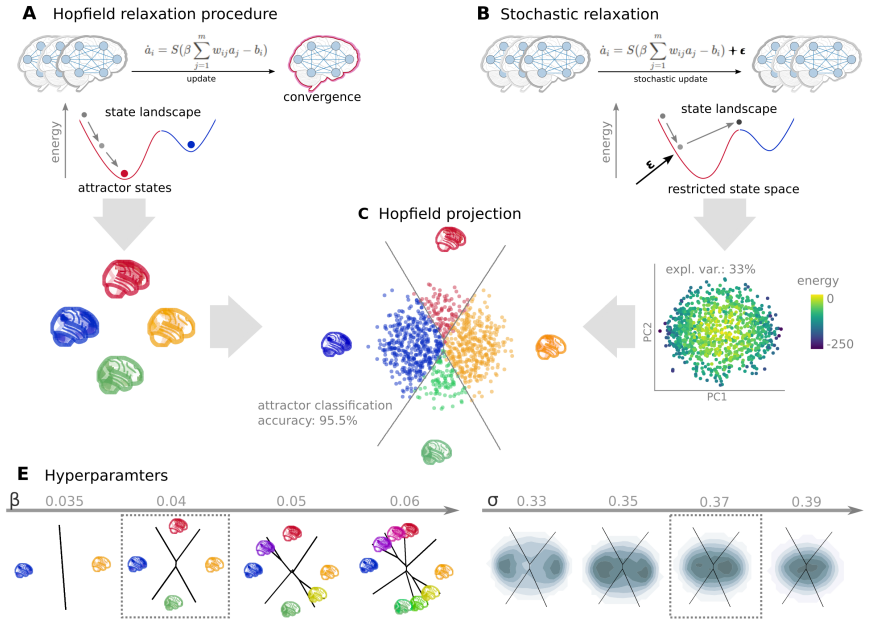
We observed that, in line with theory, increasing the temperature parameter β results in an increasing number of attractor states ((Figure 2.1E, left) appearing in symmetric pairs (i.e. $a_i = -a_j$). For the sake of simplicity, we set $\beta = 0.4$ for the rest of the paper, resulting in 4 distinct attractor states (2 symmetric pairs).

Without modifications, connectome-based Hopfield networks always converge to an equilibrium state. To account for stochastic fluctuations in neuronal activity (Robinson et al. [2005]), we add weak Gaussian noise to the connectome-based Hopfield network, to prevent the system reaching equilibrium. This approach, similarly to Stochastic DCM ()), induces a "stochastic walk" of the internal state (activity pattern) of the network that may traverse extensive regions of the state space, determined by the "gravity field" (basins) of multiple attractor states (Figure 2.1B).

We hypothesise that the resulting dynamics reflect many important characteristics of spontaneous activity fluctuations in the brain and may serve as a useful generative computational model of large scale brain dynamics. To sample the resulting state space, we obtained 100.000 iterations (starting from a random seed pattern) of the stochastic relaxation procedure with a Hopfield network initialized with the mean functional connectome in study 1 (n=44). Next, to increase interpretability, we obtained the first two components from a principal component analysis (PCA) on the resulting state space sample to construct a low-dimensional embedding. Largely independent on the free parameter σ (variance of the noise), the first two principal components (PCs) explained around 15% of the variance in the state space, with low energy states (attractor states) located at the extremes of the PCs (Figure 2.1B, bottom plot). The PCA embedding was found to be largely consistent across different values of β and σ (Figure 2.1E). For all further analyses, we fixed $\sigma = 0.37$, as a result of a coarse optimization procedure to reconstruct the bimodal distribution of empirical data on the same projection (Figure 2.1E, see Methods for details) On the low-dimensional embedding, which we refer to as the *Hopfield projection*, we observed a clear separation of the attractor states (Figure 2.1C), with the two symmetric pairs of attractor states located at the extremes of the first and second PC. To map the attractor basins onto the space spanned by the first two PCs (Figure 2.1C), we obtained the attractor state of each point visited during the stochastic relaxation and fit a multinomial logistic regression model to predict the attractor state from the first two PCs. The resulting model achieved a high prediction accuracy (out-of-sample accuracy 96.5%). Attractor bases were visualized based on the decision boundaries of this model (Figure 2.1C). .We propose the Hopfield projection depicted on (Figure 2.1C) as a simplified representation of brain dynamics, as modelled by connectome-based Hopfield networks, and use it as a basis for all subsequent analyses in this work.

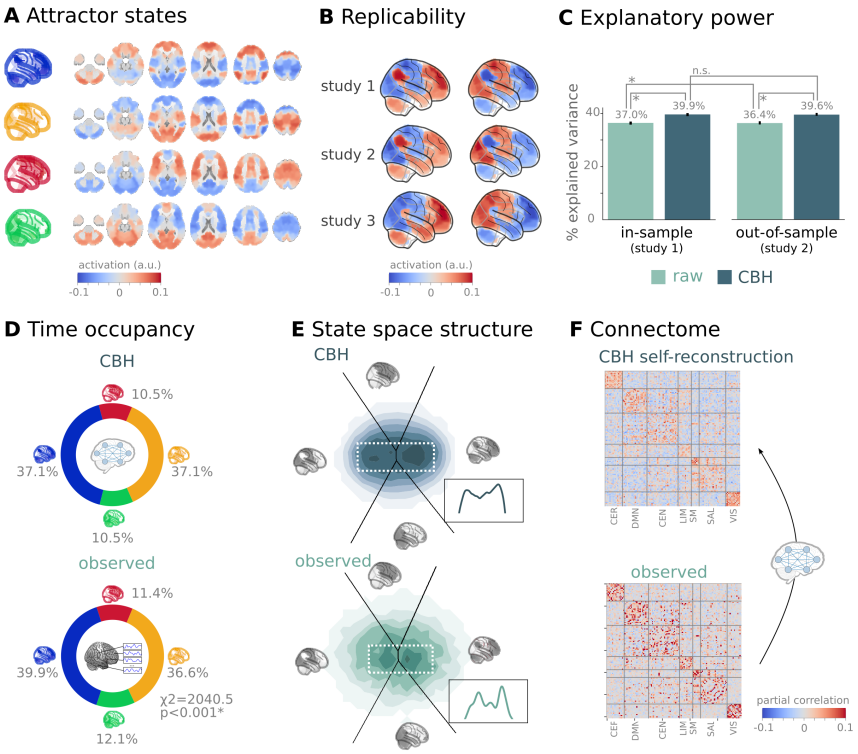
2.2 Reconstruction of resting state brain dynamics

The obtained attractor states resemble familiar, neurobiologically highly plausible patterns (Figure 2.2A). The first pair of attractors (mapped on PC1) resemble the two complementary “macro” systems described by and Cioli et al. [2014] as well as the two primary brain states previously described by . This state-pair has previously been described as an “extrinsic” system that is more directly linked to the immediate sensory environment and an “intrinsic” system whose activity preferentially relates to changing higher-level, internal context (a.k.a the default mode network). The other attractor pair spans an orthogonal axis between regions commonly associated with active (motor) and passive inference (visual).



A Top: During so-called relaxation procedure, activities in the nodes of a connectome-based Hopfield (CBH) network are iteratively updated based on the activity of all other regions and the connectivity between them. The energy of a connectome-based Hopfield (CBH) network decreases during the relaxation procedure until reaching an equilibrium state with minimal energy, i.e. an attractor state. Bottom: Four attractor states obtained by a CBH initialized with a group-level functional connectivity matrix from Table **??** (n=44). **B** Top: Similarly to stochastic dynamic causal modelling, in presence of weak noise (stochastic update), the system does not converge to equilibrium anymore. Instead, it the system transverses on the state landscape in a way restricted by the topology of the connectome and the "gravitational pull" of the attractor states. Bottom: We sample the state space by running the stochastic relaxation procedure for an extended amount of time (e.g. 100.000 consecutive stochatsic updates). To construct a low-dimensional representation of the state space, we take the first principal components of the simulated activity patterns. The first two principal components explain approximately 55-85% of the variance of state energy (depending on the noise parameter σ, see Supplementary Material **X**). **C** We map the attractors for all states (color-coded) of the state space sample with the conventional Hopfield relaxation procedure (A). The four attractor states are also visualized in their corresponding position on the PCA-based projection. The first two principal components yield a clear separation of the attractive state basins (cross-validated classification accuracy: 95.5%, Supplementary Material **X**). We refer to the resulting visualization as the Hopfiled projection and use it to visualize CBH-derived and empirical brain dynamics throughout the rest of the manuscript. **E** At its simpliest form, the CBH framework entails only two free hyperparamters: the temperature parameter β (left) that controls the number of attractor states and the noise parameter of the stochastic relaxation σ. To avoid overfitting these parameters to the empirical data, we set $\beta = 0.4$ and $\sigma = 0.01$ for the rest of the paper.

A Top: During so-called relaxation procedure, activities in the nodes of a connectome-based Hopfield (CBH) network are iteratively updated based on the activity of all other regions and the connectivity between them. The energy of a connectome-based Hopfield (CBH) network decreases during the relaxation procedure until reaching an equilibrium state with minimal energy, i.e. an attractor state. Bottom: Four attractor states obtained by a CBH initialized with a group-level functional connectivity matrix from Table **??** (n=44). **B** Top: Similarly to stochastic dynamic causal modelling, in presence of weak noise (stochastic update), the system does not converge to equilibrium anymore. Instead, it the system transverses on the state landscape in a way restricted by the topology of the connectome and the "gravitational pull" of the attractor states. Bottom: We sample the state space by running the stochastic relaxation procedure for an extended amount of time (e.g. 100.000 consecutive stochatsic updates). To construct a low-dimensional representation of the state space, we take the first principal components of the simulated activity patterns. The first two principal components explain approximately 55-85% of the variance of state energy (depending on the noise parameter σ, see Supplementary Material **X**). **C** We map the attractors for all states (color-coded) of the state space sample with the conventional Hopfield relaxation procedure (A). The four attractor states are also visualized in their corresponding position on the PCA-based projection. The first two principal components yield a clear separation of the attractive state basins (cross-validated classification accuracy: 95.5%, Supplementary Material **X**). We refer to the resulting visualization as the Hopfiled projection and use it to visualize CBH-derived and empirical brain dynamics throughout the rest of the manuscript. **E** At its simpliest form, the CBH framework entails only two free hyperparamters: the temperature parameter β (left) that controls the number of attractor states and the noise parameter of the stochastic relaxation σ. To avoid overfitting these parameters to the empirical data, we set $\beta = 0.4$ and $\sigma = 0.01$ for the rest of the paper.



A The four attractor states of the connectome-based Hopfiled (CBH) network from study 1 reflect brain activation patterns with a high neurobiologicasl relevance, resembling to sub-systems previously described as being associated for "internal context" (blue), "external context" (yellow), "action" (red) and "perception" (green) (Golland et al. [2008]; ; Chen et al. [2018];). **B** The attractor states show excellent replicability in two external datasets (Table **??**, mean correlation *XX*). **C** The Hopfield projection (first two PCs of the CBH state space) explains more variance (p<0.0001) in the real resting state fMRI data than principal components derived from the real resting state data itself and generalizes better (p<0.0001) to out-of-sample data (study 2). Error bars denote 99% bootstrapped confidence intervals. **D** The CBH analysis accurately predicts (p<0.0001) the fraction of time spent on the basis of the four attractor states in real restring state fMRI data (study 1) and **E** reconstructs the characteristic bimodal distribution of the real resting state data. **F** CBH networks perform self-reconstruction: the timeseries resulting from the stochastic relaxation procedure mirror the co-variance structure of the functional connectome the CBH network was initialized with.

A The four attractor states of the connectome-based Hopfiled (CBH) network from study 1 reflect brain activation patterns with a high neurobiologicasl relevance, resembling to sub-systems previously described as being associated for "internal context" (blue), "external context" (yellow), "action" (red) and "perception" (green) (Golland et al. [2008]; ; Chen et al. [2018];). **B** The attractor states show excellent replicability in two external datasets (Table **??**, mean correlation *XX*). **C** The Hopfield projection (first two PCs of the CBH state space) explains more variance (p<0.0001) in the real resting state fMRI data than principal components derived from the real resting state data itself and generalizes better (p<0.0001) to out-of-sample data (study 2). Error bars denote 99% bootstrapped confidence intervals. **D** The CBH analysis accurately predicts (p<0.0001) the fraction of time spent on the basis of the four attractor states in real restring state fMRI data (study 1) and **E** reconstructs the characteristic bimodal distribution of the real resting state data. **F** CBH networks perform self-reconstruction: the timeseries resulting from the stochastic relaxation procedure mirror the co-variance structure of the functional connectome the CBH network was initialized with.

Figure 3: **Connectome-based Hopfield networks reconstruct characteristics of real resting state brain activity.**

A The four attractor states of the connectome-based Hopfiled (CBH) network from study 1 reflect brain activation patterns with a high neurobiologicasl relevance, resembling to sub-systems previously described as being associated for "internal context" (blue), "external context" (yellow), "action" (red) and "perception" (green) (Golland et al. [2008]; ; Chen et al. [2018];). **B** The attractor states show excellent replicability in two external datasets (Table **??**, mean correlation *XX*). **C** The Hopfield projection (first two PCs of the CBH state space) explains more variance (p<0.0001) in the real resting state fMRI data than principal components derived from the real resting state data itself and generalizes better (p<0.0001) to out-of-sample data (study 2). Error bars denote 99% bootstrapped confidence intervals. **D** The CBH analysis accurately predicts (p<0.0001) the fraction of time spent on the basis of the four attractor states in real restring state

Hopfield networks are known to exhibit remarkable robustness to noisy input () and even to corrupted weights (**ref**). We found that this property renders connectome-based Hopfiled networks as a strikingly robust tool (Supplementary Analysis X), showing a remarkable replicability (mean Pearson's correlation **XX**) across the discovery datasets (study 1) and two independent replication datasets (Table **??**, Figure 2.2C).

Further analysis in study 1 demonstrated that connectome-based Hopfield models very accurately reconstruct several characteristics of true resign state data. First, the Hopfield projection explained a large amount of variance in the real resting state fMRI data in study 1 (mean $R^2 = 0.15$) and generalized well to study 2 (mean $R^2 = 0.13$) and study 3 (mean $R^2 = 0.12$) (Figure 2.2E). Explained variance significantly exceeded that of a PCA performed on the real resting state fMRI data itself (Figure 2.2E).

Second, during stochastic relaxation, the connectome-based Hopfield network spends three-quarter of the time on the basis of the first two attractor states (equally distributed across the two) and one-quarter on the basis of the second pair (again equally distributed). To test if this characteristic can also be found in real resting state data, we obtained normalized and cleaned mean timeseries in $m = 122$ regions from all participants in study 1 obtained the attractor state of each time-frame via the connectome-based Hopfield network. We observed highly similar temporal occupancies to those predicted by the model (^{2}-test of equal occupancies: $p<0.00001$, Figure 2.2B).

Third, during the stochastic relaxation procedure, connectome-based Hopfield models generate regional timeseries that retain the partial correlation structure of the real functional connectome the network was initialized with, indicating a high-level of construct validity (Figure 2.2D). To

Finally, our connectome-based Hopfield model also accurately reconstructs the bimodal distribution of the real resting state fMRI data on the Hopfield projection (Figure 2.2F).

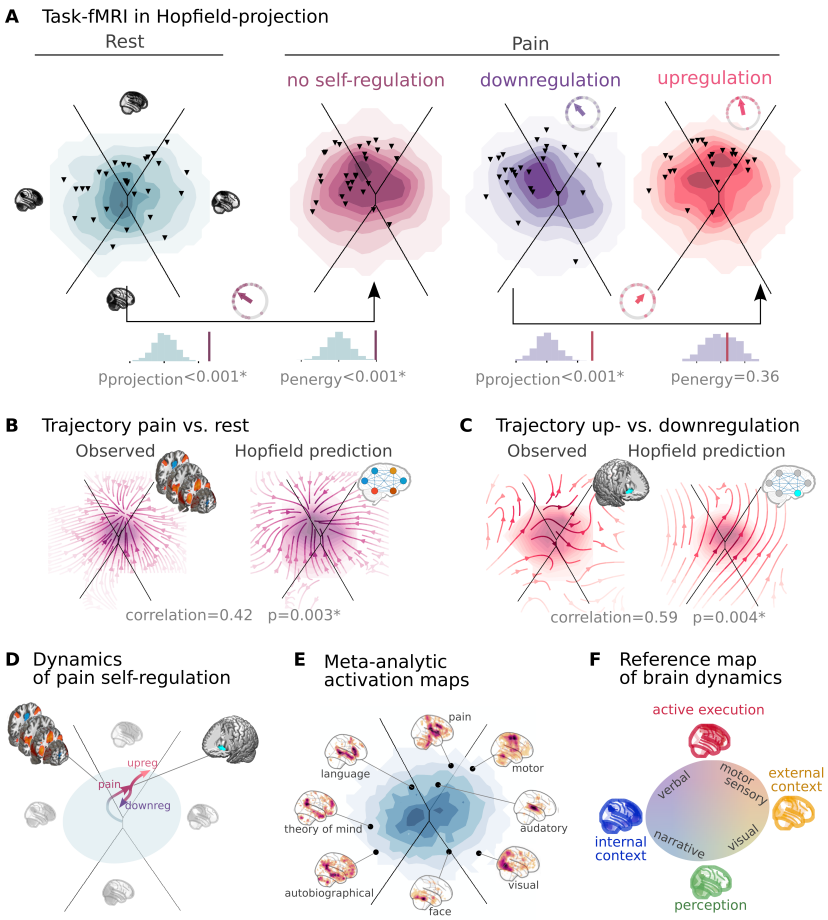
The ability of the connectome-based Hopfield model to reconstruct such characteristics of remarkable, given that the model was neither trained to reconstruct nor informed about any spatial (bi-model distribution, explanatory performance) or temporal patterns (temporal state occupancy) of the brain. The only information the model was provided with was the functional connectome, which was used to initialize the network and to constrain the dynamics of the network during stochastic relaxation. The fact that the model is able to reconstruct such characteristics of resting state brain dynamics, which are not explicitly encoded in the connectome, suggests that the connectome-based Hopfield model captures important relationships between the topology of brain connectome of the dynamics of the brain activation.

2.3 An explanatory framework for task-based brain activity

The proposed framework provides a natural account for how activation patterns in the brain dynamically emerge form the underlying functional connectivity. To illustrate this, we obtained task-based fMRI data from a study by (Table **??**, n=33, see Figure 2.2), invetsigating the neural correlates of pain, with focus on self-regulation. We found that time-frames from obtained from periods with pain stimulation (taking into account hemodynamics, see Methods for details) locate significantly differently on the Hopfield projection than time-frames obtained from periods without pain stimulation (permutation test, $p<0.001$, Figure 2.3A, left). Energies, as defined by the Hopfield model, were also significantly different between the two conditions (permutation test, $p<0.001$), with higher energies during pain stimulation. The Hopfield-projections thus provide an intuitive account for how the underlying functional connectivity of the brain can give rise to different activation patterns, depending on the current input. Change in input (i.e. task) does not switch to the brain into a distinct mode of operation but acts as a perturbation of the system's dyanmics, resulting in mean activations changes that are only reliable measurable over an extended period of time, as done by conventional task-based fMRI analyses.

Participants were instructed to up- or down-regulate their pain sensation (resulting in increased and decreased pain reports and differential brain activity in the nucleus accumbens, NAc, see details), which resulted in further changes of the location of momentary brain states on the Hopfield-projection (permutation test, $p<0.001$, Figure 2.3A, right). Interestingly, self-regulation did not manifest in significant energy changes (permutation test, $p=0.36$). This suggest that visualizong data on the Hopfiled projection can also capture changes in brain activity that originate from intrinsic modulation, rather than from changes in external input.

The proposed framework offers much more than visualization and inference of resting state and task based data on the Hopfield projection. It can provide a generative model for observed activity changes that can be used to predict brain activity under different conditions. To illustrate this, we used the Hopfield model to simulate brain activity during pain stimulation and self-regulation. First we registered the frame-to-frame



A Functional MRI time-frames during pain stimulation from Table **??** (second Hopfiled projection plot) and self-regulation (third and fourth) locate significantly differently on the Hopfield projection than brain states during rest (first projection, permutation test, p<0.001 for all). Energies, as defined by the Hopfield model, are also significantly different between rest and the pain conditions (permutation test, p<0.001), with higher energies during pain stimulation. Triangles denote participant-level mean activations in the various blocks (corrected for hemodynamics). Circle plots show the directions of the change for each individual (points) as well as the mean direction across participants (arrow), as compared to the reference state (downregulation for the last circle plot, rest for all other circle plots). **B** The average difference between the characteristic directions of the single time-frames on the Hopfield projection reveal a non-linear flow difference between pain the brain dynamics during pain and rest (left). When introducing weak pain-related signal in the CBH network during stochastic relaxation, it accurately reproduces these non-linear flow differences (right). **C** Similarly simulating activity in the nucleaus accumbens (NAc) reconstructs a non-linear flow difference between up- and downregulation (left). When introducing weak self-regulation-related signal similar to the observed dynamics (characterized by NAc activation differences, as observed by . **D** Schematic representation of brain dynamics during pain and its up- and downregulation, visualized on the Hopfiled projection. Pain shifts spontanous brain dynamics towards the "action" subsystem, converging to a putative "ghost attractor of pain". Up-regulation by NAc de-activation exerts force towards a similar dircetionm while down-regulation by NAc activation exhibist an opposite effect on brain dynamics, leading to the brain less frequent "visiting" pain-associated states. **E** Visualizing meta-analyitic activation maps on the Hopfiled projection informs pur theoretical interpretative framework **F** for spontaneous and task-based brain dynamics. In the proposed framework, taks-based activity is not a mere response to external stimuli in certain brain locations but a perturbation of the brain's charcteristic dynamic trajectories. In this framework, conventional task-based fMRI analyses capture mean differences of the whole brain dyanmics, resulting in the widely reported focal "activation maps" thought to be specific to various taks and stimuli. In the CBH framework, the brain's characteristic trajectories are constrained by the underlying functional connectivity and only perturbed by external inpout, rather than predestinated.

A Functional MRI time-frames during pain stimulation from Table **??** (second Hopfiled projection plot) and self-regulation (third and fourth) locate significantly differently on the Hopfield projection than brain states during rest (first projection, permutation test, p<0.001 for all). Energies, as defined by the Hopfield model, are also significantly different between rest and the pain conditions (permutation test, p<0.001), with higher energies during pain stimulation. Triangles denote participant-level mean activations in the various blocks (corrected for hemodynamics). Circle plots show the directions of the change for each individual (points) as well as the mean direction across participants (arrow), as compared to the reference state (downregulation

transitions in the real fMRI data (all four conditions: rest, pain without self-regulation, downregulation, upregulation) and converted those into the Hopfield embedding (resulting in a 2-dimensional vector on the Hopfield projection for each transition). Then, we assessed the mean direction in various segments of the projection (on a 6x6 grid). Next we took the difference of these mean directions between rest and pain (no regulation) (Figure 2.3B, left side), as well as between down- and upregulation (Figure 2.3C, left side). This analysis revealed remarkable non-linear trajectory patterns, showing the most likely direction the brain proceeds towards from a given state (activity pattern) in a given condition (pain without delf-regulation or upregulation), as compared to the reference state (rest and downregulation, respectively). In case of pain vs. rest, brain activity is pulled toward a "ghost attractor" located in the proximity of the Hopfiled projection typical pain activation map, as observed via conventional task-based fMRI analyses. In terms of attractor states, this belongs to the basin of attractor corresponding to sensory and motor processes (active inference). In case of up vs. downregulation, brain activity is pulled generally towards a similar direction, although with non-linear local perturbations and the lack of a clear ghost attractor.

Next, we aimed to assess, how much these non-linear dyanmics can be reconstructed by the proposed framework. To simulate how brain dynamics alter during pain stimulation, we obtained a meta-analytic pain activation map (Zunhammer et al. [2021]) and introduced it with as additional signal on top of the Gaussian noise during the stochastic relaxation procedure. Not that, while adding such signal naturally results in a slight, linear shift on the Hopfield projection for each state generated during the stochastic relaxation procedure, that alone could only very weakly account for the observed nonlinear dyanmics in the real data (Supplementary material X). After optimizing across 5 different signal-to-noise (SNR) values (logarithmically spaced between 0.001 and 0.1) we found that, with a very low amount of signal added (SNR=0.01) the connectome-based Hopfield model is able to provide a highly accurate reconstruction of the observed non-linear differences in brain dynamics between the pain and rest conditions, including the "ghost attractor" of pain (Spearman's $\rho = 0.42$, p=0.003, Figure 2.3B, right side).

Interestingly, the same model was also able to reconstruct the observed non-linear differences in brain dynamics between the up- and downregulation conditions (Spearman's $\rho = 0.59$, p=0.004) with a very simple change; the addition (downregulation) or subtraction (upregulation) of activation in the NAc (the region in which Woo et al. [2015] observed significant chnages between up- and downregulation). Importantly, in this analysis, we did not have to optimize any paramters of the model, we simply used the same low SNR for the NAC that we already found optimal in the previous analysis (SNR=0.01, Figure 2.3C, right side).

These results provide a fresh perspective on the neural mechanisms beyond pain and its self-regulation and provides a mechanistic account for the role of both "traditional" pain-related regions and the NAc in pain regulation (Figure 2.3D). These results also highlight, that thr conceptual distinction between resting and task states might be - to a large degree - a false dichotomy. Rather, the brain is in a constant state of flux, which is only slightly perturbed by task states (even by so salient stimuli as pain) and the Hopfield projection can be used to visualize and quantify these dynamics.

To provide a comprehensive picture on how other tasks map onto the Hopfield projection, we obtained various task-based meta-analytic activation maps from Neurosynth (see Supplementary material X for details) and plotted them on the Hopfield projection (Figure 2.3E). This analysis revealed that the Hopfield projection can be used to visualize and quantify the dynamics of a wide range of cognitive processes, including sensory, motor, cognitive and social processes and reveals that the two principal axes of the projection map well to internal vs. external context and active inference vs. passive perception, respectively. In this coordinate system, visual processing is labeled "external-passive", sensory-motor processes "external-active", language, verbal cognition and working memory is labelled "internal-active" and long-term memory and autobiographic narratives fall into the "internal-passive" regime (Figure 2.3F). This analysis also revealed that the Hopfield projection can be used to visualize and quantify the dynamics of a wide range of cognitive processes, including sensory, motor, cognitive and social processes and reveals that the two principal axes of the projection map well to internal vs. external context and active inference vs. passive perception, respectively.

These results highlight a very powerful feature of the proposed generative framework, namely that it can be used to simulate and predict brain activity under different conditions. Predicting the effect of lower or higher level of activity in certain regions, or lower or higher connectivity among them, on global brain dynamics and responses to various tasks provides unprecedented opportunities for forecasting the effect of interventions, such as pharmacological or non-invasive brain stimulation, on brain function.

2.4 Clinical relevance

In our final analysis, we provide a brief outlook towards the potential clinical applications of CBH analysis. We analyzed three large public clinical databases as provided by the Autism Brain Imaging Data Exchange (Table **??**: ABIDE, **ref**), the Centers of Biomedical Research Excellence (Table **??**: COBRE, **ref**) and the Alzheimer's Disease Neuroimaging Initiative (Table **??**: ADNI, **ref**), analyzed resting state fMRI data of patients with autism spectrum disorder (ASD), schizophrenia (SCZ) and Alzheimer's disease (AD). Patients' data was contrasted to their respective control groups (typically developing controls for ASD, healthy control participants for SCH and individuals with mild cognitive impairment (MCI), respectively.

In all three datasets, we used the CBH model from study 1 and projected the fMRI timeseries of all involved participants onto the Hopfield projection. For each participant, we obtained the average activation of all time-frames belonging to the same attractor state (4 maps per participant) and compared these across groups with a permutation test, controlled for the family-wise error rate across brain regions and attractor states (122*4 comparisons).

We found several significant differences the mean attractor activation of patients as compared to the respective controls. In ASD, all four attractor activation maps showed significant differences (Figure 5A, **table**), characterized by altered activation in the *precuneus, posterior congulate, sensory-motor system, posterior insula, and cerebellum*.

In SCZ, the most prominent differences were found in the subsystem for internal context, with elevated activity of regions that are not typically active in this state, including the *thalamus, the striatum and several cortical regions* (Figure 5B, **table**). Additional activation increases in *visual and motor* areas were observed in the active inference subsystem.

In the AD vs. MCI comparison, we found significant differences in two of the four attractor activation maps (Figure 5C, **table**), indicating changes in the resting state activity of subsystems for passive inference and internal context (both of which together host long-term memory processes, see Figure 2.3F). At the regional level, differences are characterized by altered activation in the *dorsolateral prefrontal cortex (DLPFC) and the cerebellum*.

3 Discussion

Regions of the brain are in a continuous process of exchanging information, resulting in co-activations commonly referred to as functional connectivity. The amount of information exchanged is not constant across the brain but varies largely for various region pairs, spanning the complex network known as the functional connectome. Here we have proposed a simplistic yet powerful model of how activity flow through this complex network topology naturally restricts the system's dyanmics, and gives rise to distinct brain sates and characteristic dynamic responses to perturbations. In a series of experiments, we have shown that the proposed model can be used to accurately reconstruct and predict large-scale brain activity under different conditions, with unprecedented opportunities for forecasting the effect of interventions, such as pharmacological or non-invasive brain stimulation, on brain function.

The construct validity of our model is rooted in the activity flow principle, first introduced by . The activity flow principle states that functional connectivity between region A and B can be conceptualized as the degree to which activity is transferred form A to B. This principle has been shown to successfully predict held out brain activations by a weighted sum of the activations of all the regions where the weights are set to the functional connectivity of those regions to the held-out region (; ; ; ;)).

ToDo: latent FC-based modelling:

Our model was born from the intuition that the recurrent, iterative application of the activity flow equation results in a system showing close analogies with a type of recurrent artificial neural networks, know as Hopfiled networks (). Hopfield networks have previously been shown to exhibit a series of characteristics that are also highly relevant for brain function, including the ability to store and recall memories (**ref**), self-repair (**ref**), a staggering robustness to noisy or corrupted inputs (**ref**) and the tendency to produce multistable dynamics organized by the "gravitational pull" of a finite number of attractor states (**ref**).

The proposed link between activity flow and Hopfield networks has and important implication: network weights must be initialized with functional connectivity values, (specifcally, partial correlations, as recommend by), instead of applying an explicit training procedure (common in the "neuroconnectomist" approach

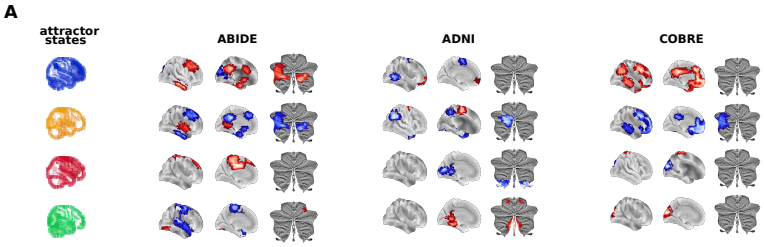


Figure 5: **Connectome-based Hopfield analysis as a sensitive tool for the study of clinical disorders.**
<\br>We quantified attractor state activations in three clincal datasets ((Table **??**) as the individual-level mean activation of all time-frames belonging to the same attractor state. CBH analysis of attractor state activations revealed significant differences in all three datasets. **A** Comparison of individuals with autism spectrum disorder (ASD) and typically developing controls (TD) is characterized by **todo**. **B** The most prominent Schizophrenia (SCZ)-related differences (as compared to healthy controls (HC) are related to the activity of the internalalization-related subsystem. **todo C** Alzheimer's disease (AD) is characterized by altered activation in **todo** the subsystems for passive inference and internal context (both of which together host long-term memory processes, see Figure 2.3F). All results are corrected for multiple comparisons across brain regions and attractor states (122*4 comparisons) with Bonferroni-correction. See Table **X** for detailed results.

(**ref**)) or using the structural connectome (a standard practice of conventional computational neuroscience (**ref**)).

Using functional conncetome-based Hopfield (CBH) model provides a simple yet powerful framework for the mechanistic understanding of brian dynamics. Its simplicity comes with an important advantages.

First, increasing model complexity results in an exponential explosion of the parameter space. Although complex, fine-grained computation models hold promise a full-blown understanding, they very easily overfit real data (**ref**). The basic CBH approach has only two hyperparameters (temperature and noise) and produce fairly consistent behavior on a wide range of parameter values. To demonstrate the power of simplicity, in the present work, we deliberately minimized fine-tuning of any free parameters. We fixed the temperature parameter at a value that robustly provides 4 attractor states and used a single noise level for all experiments (selected with a coarse optimization procedure to approximately mimic the distribution of real data).

Second, increasing complexity means increasing burden in terms of interpretability. The CBH model establishes a simple, direct link between two most popular measures of brain function: functional connectivity and brain activity. This link is not only conceptual, but also mathematical, and allows us to investigate and forecast changes of the system's dyanmics in response to perturbations of both activity and connectivity.

In this initial investigation, we further reduced complexity by restricting the analysis to a simplified 2-dimensional embedding of the state-space generated by the CBH approach, which we refer to as the Hopfield projection. This projection is a powerful tool for the visualization of the CBH model's dynamics, and allows for a direct comparison with the dynamics of the original brain activity.

However, the Hopfield projection only conveys a small proportion of the richness of the full state-space dynamics reconstructed by the CBH model. Investigating higher-dimensional dynamics, fine-tuned hyperparameters, the effect of different initializations and perturbations is an important direction for future work, with the potential to further improve the model's accuracy and usefulness.

Given these intentional simplifications, it is remarkable, if not surprising, how accurately the CBH model is able to reconstruct and predict brain dynamics under a wide range of conditions. Next to accurately reconstructing the distribution of, and the time spent in, different brain states during resting state, its superiority in explaining, and generalizing to, resting state brain activation patterns over principal components derived from the same data is particularly striking. The question arises, how can a relatively simple model, which is informed about empirical brain dynamics only through the functional connectome, be so powerful? A possible answer is that, while empirical data (and its principal components) are corrupted by noise and low sampling rate, the highly noise tolerant nature and the self-repair properties of the CBH architecture allow it to capture and reconstruct the basic principles of the underlying dynamics.

The noise-tolerance of the proposed architecture also explains the high replicability of CBH attractors across different datasets (study 2 and 3). The observed level of replicability allowed us to re-use the CBH model constructed with the connectome of study 1 for all subsequent studies, without any further fine-tuning or study-specific parameter optimization.

The connectome obtained form study 1 was also used to evaluate the model's ability to capture and forecast task-induced brain dynamics in study 4 and 5. In these analyses, was not only able to capture participant-level activity changes induced by pain and self-regulation (showing significant differences on the Hopfield projection and in terms of state energy) but also accurately predicted the non-linear changes in activity flow induced by activity changes characteristic.

Brain dynamics can not only be perturbed by task or other types of experimental or naturalistic interventions, but also by pathological alterations. In our analysis of clinical samples study 6-8 we found that mean attractor activations show characteristic alteration in autism spectrum disorder (ASD), Schizophrenia (SCH) and Alzheimer's disease (AD). These changes were also detectable on the Hopfield projection, and were accompanied by significant changes in the state energies. The Hopfiled projection also allowed us to visualize the effect of different types of perturbations on the brain's attractor landscape, providing a novel perspective on the pathophysiology of these disorders.

ToDo: more details on clinical outlook

ToDo: discuss: what are attractor states at all? Platonic idealizations of brain states, that are continously approximated by the brain?

Todo: for spontaneous and task-based brain dynamics. In the proposed framework, taks-based activity is not a mere response to external stimuli in certain brain locations but a perturbation of the brain's charcteristic dynamic trajectories. In this framework, conventional task-based fMRI analyses capture mean differences of the whole brain dyanmics, resulting in the widely reported focal "activation maps" thought to be specific to various taks and stimuli. In the CBH framework, the brain's characteristic trajectories are constrained by the underlying functional connectivity and only perturbed by external inpout, rather than predestinated.

ToDo: discuss: the CBH model is not a model of brain function, but a model of brain dynamics. It does not strice to explain various brain regions ability to perform certain computations, but the brain's characteristic trajectories, which are perturbed by tasks and other types of interventions.

Together, these results open up a series of exciting opportunities for the mechanistic understanding of brain function. By its generative nature, the CBH model could foster analyses that aim at disentangling causal relationships, which are extremely difficult to infer in case of systems as complex as the brain. It could, for instance, aid the differentiation of primary causes and secondary effects of particular activity or connectivity changes in various clinical conditions.

Moreover, the CBH approach might provide testable predictions about the effects of interventions, like pharmacological or non-invasive brain stimulation (e.g. transcranial magnetic or direct current stimulation, focused ultrasound) or neurofeedback, on brain function. For instance, in the context of pain, the CBH model might be used to predict the effect of various analgesic drugs (or other treatment strategies with known neural correlates) on the individual level (e.g. based on the individual functional connectome). Aidng the design of personalized medicine approaches is a particularly promising field of application for the proposed framework.

The generative nature of the proposed framework may be also used to generate synthetic brain activity data, which can be used to train and test machine learning algorithms, such as deep neural networks, for the prediction of brain activity from functional connectivity. This approach may be particularly useful in the context of clinical applications, where the amount of available data is often limited.

4 Conclusion

To conclude, here we have proposed a novel computational framework that accurately captures and predicts brain dynamics under a wide range of conditions. The framework models large-scale activity flow in the brain with a recurrent artificial neural network architecture that, instead of being trained to solve specific tasks or mimic certain dynamics, is simply initialized with the empirical functional connectome. The framework identifies biologically meaningful attractor states and provides a model for how these restrict brain dyanmics. The proposed framework, referred to as the connectome-based Hopfield (CBH) model, can accurately reconstruct and predict brain dynamics under a wide range of conditions, including resting state, task-induced activity changes, and pathological alterations. CBH analyses provide a simple, robust, and highly interpretable computational alternative to the conventional descriptive approach to investigating brain function and establish a link between connectivity and activity. The generative nature of the proposed model opens up a series of exciting opportunities for future research, including novel ways of assessing causality and mechanistic understanding and the possibility to predict the effects of various interventions, thereby paving the way for novel personalized medical approaches.

5 Methods

5.1 Hopfield network

The weights w_{ij} have to be symmetric and the diagonal elements are set to zero.

Todo

Todo

study	modality	analysis	n	age (mean±sd)	%female	references
study 1	resting state	discovery	41	26.1±3.9	37%	
study 2	resting state	replication	48	24.9±3.5	54%	
study 3	resting state	replication	29	24.8±3.1	53%	
study 4	task-based	pain self-regulation	33	todo	todo	
study 5 (Neurosynth)	task-based	coordinate-based meta-analyses	14371 studies in total	•	•	
study 6 (ABIDE)	resting state	Autism Spectrum Disorder	ASD: 98, NC: 74	todo	todo	todo
study 7 (ADNI)	resting state	Alzheimer's Disease vs. Mild Cognitive Impairment	AD:, MCI:	todo	todo	todo
study 8 (COBRE)	resting state	Schizophrenia	SCH: , HC:	todo	todo	todo

References

Xiao Liu and Jeff H. Duyn. Time-varying functional network information extracted from brief instances of spontaneous brain activity. *Proceedings of the National Academy of Sciences*, 110(11):4392–4397, feb 2013. doi:10.1073/pnas.1216856110. URL https://doi.org/10.1073%2Fpnas.1216856110.

Jonas Richiardi, Hamdi Eryilmaz, Sophie Schwartz, Patrik Vuilleumier, and Dimitri Van De Ville. Decoding brain states from fMRI connectivity graphs. *NeuroImage*, 56(2):616–626, may 2011. doi:10.1016/j.neuroimage.2010.05.081. URL https://doi.org/10.1016%2Fj.neuroimage.2010.05.081.

Diego Vidaurre, Stephen M. Smith, and Mark W. Woolrich. Brain network dynamics are hierarchically organized in time. *Proceedings of the National Academy of Sciences*, 114(48):12827–12832, oct 2017. doi:10.1073/pnas.1705120114. URL https://doi.org/10.1073%2Fpnas.1705120114.

Andreas Heinz, Graham K Murray, Florian Schlagenhauf, Philipp Sterzer, Anthony A Grace, and James A Waltz. Towards a unifying cognitive, neurophysiological, and computational neuroscience account of schizophrenia. *Schizophrenia Bulletin*, 45(5):1092–1100, nov 2018. doi:10.1093/schbul/sby154. URL https://doi.org/10.1093%2Fschbul%2Fsby154.

Adrien Doerig, Rowan P. Sommers, Katja Seeliger, Blake Richards, Jenann Ismael, Grace W. Lindsay, Konrad P. Kording, Talia Konkle, Marcel A. J. van Gerven, Nikolaus Kriegeskorte, and Tim C. Kietzmann. The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7):431–450, may 2023. doi:10.1038/s41583-023-00705-w. URL https://doi.org/10.1038%2Fs41583-023-00705-w.

Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsay, Kenneth D. Miller, Richard Naud, Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema, João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, oct 2019. doi:10.1038/s41593-019-0520-2. URL https://doi.org/10.1038%2Fs41593-019-0520-2.

Michael W Cole, Takuya Ito, Danielle S Bassett, and Douglas H Schultz. Activity flow over resting-state networks shapes cognitive task activations. *Nature Neuroscience*, 19(12):1718–1726, oct 2016. doi:10.1038/nn.4406. URL https://doi.org/10.1038%2Fnn.4406.

P. A Robinson, C. J Rennie, D. L Rowe, S. C O'Connor, and E Gordon. Multiscale brain modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1457):1043–1050, may 2005. doi:10.1098/rstb.2005.1638. URL https://doi.org/10.1098%2Frstb.2005.1638.

Claudia Cioli, Hervé Abdi, Derek Beaton, Yves Burnod, and Salma Mesmoudi. Differences in human cortical gene expression match the temporal properties of large-scale functional networks. *PLoS ONE*, 9(12): e115913, dec 2014. doi:10.1371/journal.pone.0115913. URL https://doi.org/10.1371%2Fjournal.pone.0115913.

Yulia Golland, Polina Golland, Shlomo Bentin, and Rafael Malach. Data-driven clustering reveals a fundamental subdivision of the human cortex into two global systems. *Neuropsychologia*, 46(2):540–553, 2008. doi:10.1016/j.neuropsychologia.2007.10.003. URL https://doi.org/10.1016%2Fj.neuropsychologia.2007.10.003.

Richard H. Chen, Takuya Ito, Kaustubh R. Kulkarni, and Michael W. Cole. The human brain traverses a common activation-pattern state space across task and rest. *Brain Connectivity*, 8(7):429–443, sep 2018. doi:10.1089/brain.2018.0586. URL https://doi.org/10.1089%2Fbrain.2018.0586.

Matthias Zunhammer, Tamás Spisák, Tor D. Wager, Ulrike Bingel, Lauren Atlas, Fabrizio Benedetti, Christian Büchel, Jae Chan Choi, Luana Colloca, Davide Duzzi, Falk Eippert, Dan-Mikael Ellingsen, Sigrid Elsenbruch, Stephan Geuter, Ted J. Kaptchuk, Simon S. Kessner, Irving Kirsch, Jian Kong, Claus Lamm, Siri Leknes, Fausta Lui, Alexa Müllner-Huber, Carlo A. Porro, Markus Rütgen, Lieven A. Schenk, Julia Schmid, Nina Theysohn, Irene Tracey, Nathalie Wrobel, and Fadel Zeidan and. Meta-analysis of neural systems underlying placebo analgesia from individual participant fMRI data. *Nature Communications*, 12(1), mar 2021. doi:10.1038/s41467-021-21179-3. URL https://doi.org/10.1038%2Fs41467-021-21179-3.

Choong-Wan Woo, Mathieu Roy, Jason T. Buhle, and Tor D. Wager. Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. *PLoS Biology*, 13(1):e1002036, jan 2015. doi:10.1371/journal.pbio.1002036. URL https://doi.org/10.1371%2Fjournal.pbio.1002036.