Interpretable Hierarchical RNNs for rs-fMRI: Promise and Limits of Individualized Brain Dynamics

Carlotta B.C Barkhau[1,2] , Keyvan Mahjoory[1] , Manuel Brenner[3], Elias Weber[4], Ramona Leenings[5], Clemens Pellengahr[1], Nils R. Winter[1], Maximilian Konowski[1], Tabea Straeten[2] , Susanne Meinert[1,6], Elisabeth J. Leehr[1,7], Kira Flinkenflügel[1], Tiana Borgers[1], Dominik Grotegerd[1], Hannah Meinert[1], Julia Hubbert[1], Christoph Jurischka[1], Judith Krieger[1], Wiebke Ringels[1], Frederike Stein[8], Florian Thomas-Odenthal[8], Paula Usemann[8], Lea Teutenberg[8], Igor Nenadic[8], Benjamin Straube[8], Nina Alexander[8], Andreas Jansen[8], Hamidreza Jamalabadi[8,9,10], Tilo Kircher[8], Markus Junghöfer[2], Udo Dannlowski[1,11], Tim Hahn[1]

[1] Institute for Translational Psychiatry, University of Münster, Germany
[2] Institute for Biomagnetism and Biosignalanalysis, University of Münster, Germany
[3] Ernst Strüngmann Institute for Neuroscience, Frankfurt, Germany
[4] Dept. of Theoretical Neuroscience, Central Institute of Mental Health (CIMH)
[5] Charité – Universitätsmedizin Berlin, Department of Psychiatry and Neuroscience , Berlin, Germany
[6] Institute for Translational Neuroscience, University of Münster, Germany
[7] Department of Clinical Psychology and Psychotherapy Georg-Elias-Müller-Institute of Psychology, Georg-August-University of Göttingen, Germany
[8] Department of Psychiatry and Psychotherapy, University of Marburg, Germany
[9] Center for Mind, Brain, and Behavior (CMBB), University of Marburg, Germany
[10] Faculty of Medicine, University of British Columbia, Canada
[11] Bielefeld University, Medical School and University Medical Center OWL, Protestant Hospital of the Bethel Foundation, Department of Psychiatry, Bielefeld, Germany

Corresponding author:
Carlotta B.C. Barkhau
Malmedyweg 15
48149 Münster, Germany
e-mail: carlotta.barkhau@uni-muenster.de

Abstract

Modeling individual brain dynamics from resting-state fMRI (rs-fMRI) remains challenging due to substantial inter-subject variability, measurement noise, and limited data length per subject. Here, we systematically evaluate a hierarchical dynamical systems framework based on shallow piecewise-linear recurrent neural networks (shPLRNNs) for individualized modeling of rs-fMRI data, with a particular focus on reproducing subject-specific functional connectivity (FC).

We applied the framework to 1,423 rs-fMRI samples from healthy participants of the Marburg-Münster Affective Disorders Cohort Study (MACS). Simulated rs-fMRI data robustly reproduced empirical FC patterns, with comparable reconstruction accuracy on training and independent validation sets. Generalization to unseen individuals was heterogeneous and strongly depended on how typical a subject's connectivity pattern was relative to the training cohort, with template similarity explaining 37% of variance in reconstruction accuracy.

Learned subject-specific parameters exhibited significant test-retest stability and higher within-subject than between-subject similarity on longitudinal data from two different timepoints, supporting their interpretation as individualized dynamical markers. Associations between individual parameters and demographic or cognitive variables were statistically significant but modest in effect size, and predictive performance remained below that obtained using empirical rs-fMRI features directly. Together, these results demonstrate that hierarchical shPLRNNs can extract meaningful and stable individual-specific dynamical structure from rs-fMRI data, while highlighting current limitations in capturing fine-grained individual differences. The findings delineate key trade-offs between model expressivity, generalization and subject specificity, and point to directions for future methodological refinement in individualized brain modeling.

1. Introduction

Understanding individual differences in brain function is a central aim in neuroscience, with important implications for personalized medicine and psychiatric research. Resting-state functional magnetic resonance imaging (rs-fMRI) has emerged as a widely used tool for probing intrinsic brain activity, offering non-invasive insight into functional organization and connectivity patterns. However, modeling individual brain dynamics from rs-fMRI remains challenging due to limited scan duration per subject, inter-subject variability and high noise levels inherent in the data (1,2).

Recent methodological advances, such as precision fMRI approaches (3), individual-specific parcellation techniques (4–6), whole-brain modeling (7–9) and machine learning frameworks tailored to noisy, high-dimensional data (10,11), have begun to address these limitations. Collectively, these approaches aim to improve the reliability and specificity of individual-level brain mapping, thereby supporting the development of more accurate biomarkers and therapeutic strategies in clinical neuroscience. Nevertheless, progress at the individual level has so far remained limited, underscoring the need for further methodological innovation.

Temporal Deep Learning architectures such as recurrent neural networks (RNNs) have emerged as a promising approach for modeling and simulating fMRI time series data (12–14). In particular, hierarchical RNN architectures have been shown to provide a flexible and interpretable framework for learning dynamical systems from time series data across subjects and experimental conditions (14). By leveraging shared information across a cohort, these models enable the extraction of both group-level and individual-specific dynamics, even when individual time series are relatively short (14,15). Importantly, the hierarchical structure yields low-dimensional, interpretable latent representations that facilitate clustering of subjects with similar dynamics and reveal linear relationships to underlying control parameters (14,16). In addition, these models exhibit strong generalization properties and support transfer learning to new parameter regimes, an essential feature for neuroscientific applications where per-subject data are often limited (13,14).

In neuroscience, data acquisition is frequently expensive and logistically challenging, particularly in clinical populations (17,18), resulting in relatively short and noisy time series per individual. This poses a key challenge for individualized modeling, as purely subject-specific approaches are often underpowered. Hierarchical RNNs address this limitation by leveraging shared structure across subjects while simultaneously encoding individual-specific dynamics in a low-dimensional form. This combination makes them particularly well suited for capturing the heterogeneous dynamics of rs-fMRI data and motivates their use in the present study (14–16).

Beyond their use as flexible sequence models, hierarchical RNNs can be understood as tools for dynamical system reconstruction. Rather than merely fitting observed time series, these models aim to infer an underlying dynamical process whose long-term behavior generates the observed data (13,14). In the context of rs-fMRI, this implies that subject-specific models are expected to preserve characteristic dynamical signatures, such as stable connectivity patterns and temporal dependencies, even when generating new data beyond the training window

(16,17). From this perspective, successful modeling is not defined by pointwise prediction accuracy alone, but by the ability to reproduce individual-specific functional connectivity profiles and other emergent properties of the system. This generative, dynamics-based viewpoint distinguishes dynamical systems (e.g. hierarchical RNNs) from purely discriminative machine learning approaches and motivates their use for individualized modeling of brain activity.

Despite their promise, challenges remain in accurately capturing subtle individual differences in brain activity in the presence of substantial noise and variability (19,20). The present study aims to systematically evaluate the efficacy of hierarchical RNNs in simulating individual rs-fMRI dynamics, with particular emphasis on the balance between learning shared group-level patterns and preserving individualized dynamics. By explicitly assessing both strengths and limitations, we seek to clarify the potential of these models for advancing our understanding of individual brain function.

Our approach employs an end-to-end optimization framework in which hierarchical RNNs are jointly trained on rs-fMRI time series from multiple subjects. This setup enables the model to learn global dynamical patterns characteristic of the cohort while simultaneously adapting to subject-specific nuances (14). End-to-end training ensures that individual-level parameters are optimized in concert with group-level dynamics, allowing the network to encode both shared and unique neural characteristics within a unified modeling framework (14).

The resulting individualized parameter sets provide compact, low-dimensional representations of subject-specific neural profiles. Consistent with prior work (6,11,21,22), we hypothesize that these personalized latent representations are associated with phenotypic attributes such as sex, age, body mass index (BMI), intelligence quotient (IQ) and years of schooling (YoS).

Accordingly, the aims of this study are twofold. First, we evaluate the ability of the hierarchical RNN to accurately simulate individual rs-fMRI dynamics and preserve subject-specific functional connectivity structure. Second, we investigate whether the individualized model parameters learned during training are systematically related to subject-specific attributes. To this end, we combine classical statistical analyses with machine learning approaches to explore potential associations. Establishing such links may provide important insights into how individual differences in brain function are reflected in the latent representations learned by the hierarchical RNN (22).

2. Methods

The data used in this work is part of the Marburg-Münster Affective Disorders Cohort Study (MACS) (23,24). Data were collected at two sites (Marburg and Münster, Germany) using identical study protocols and harmonized scanner settings. The study was approved by the ethics committees of the medical faculties at the University of Marburg and the University of Münster. All participants provided written informed consent and received financial compensation.

At the time of analysis, rs-fMRI data were available from 953 healthy control subjects, including second time-point measurements for a subset of 491 subjects, resulting in a total of 1,444 samples. From these, we selected samples with at least 200 time points, yielding a final dataset of 1,423 samples. All timeseries were truncated to a uniform length of 200 time points.

To prevent data leakage, samples from the same subject were assigned exclusively to either the training, test or validation set. The dataset was split into 945 training samples, 208 test samples and 191 validation samples (Figure 1).

2.1 Preprocessing and Functional Connectivity

Preprocessing of empirical rs-fMRI data was performed using the open-source Connectivity Analysis Toolbox (CATO) (25).

Preprocessing steps included slice timing correction (FSL sliceTimer), motion correction (FSL MCFLIRT) (26) and parcellation into cortical regions based on the Desikan-Killiany atlas using the anatomical T1-weighted image.

Motion-related metrics, including framewise displacement (FD) and DVARS, were computed following Power et al. (20). Covariates were regressed out from the rs-fMRI time series and a zero-phase Butterworth band-pass filter (0.01–0.1 Hz) was applied to remove scanner drift and physiological noise (27). Frames exhibiting excessive motion were removed using the following thresholds: maxFD = 0.25, maxDVARS = 1.5, minViolations = 2, backwardNeighbors = 1, forwardNeighbors = 0.

Functional connectivity (FC) matrices were computed by calculating Pearson correlation coefficients between the mean BOLD time series of each brain region.

2.2 Hierarchical RNN Model

We applied a hierarchical dynamical systems reconstruction (DSR) framework based on recurrent neural networks (RNNs) to model ROI-level rs-fMRI data across subjects (14). This approach promises to be particularly well-suited for scenarios with relatively short or noisy time series, as it leverages shared structure across the entire dataset while capturing subject-specific dynamics in a low-dimensional and interpretable form (14,21). Each subject's multivariate time series was modeled using a subject-specific generative RNN, implemented as a shallow piecewise-linear recurrent neural network (shPLRNN) (14).

Subject-specific RNN parameters were generated from low-dimensional feature vectors that are unique to each subject. These feature vectors are transformed into full model parameters via a shared set of group-level weights, which define a common mapping across all subjects. In this way, individual differences are captured by the feature vectors, while the shared weights encode population-level structure. This hierarchical design allows the model to combine shared dynamics with subject-specific variability, particularly in settings with short or noisy time series (14). See Figure 1 for a schematic overview.

2.3 Training and Hyperparameter Optimization

Model training was performed end-to-end using generalized teacher forcing to stabilize learning in the presence of potentially chaotic dynamics (28). The learned subject-specific feature representations can further serve as compact and dynamically meaningful biomarkers and may be used for downstream analyses such as behavioral prediction or clinical classification (11,22).
Two key architectural hyperparameters were optimized using Optuna (29): (i) the number of subject-specific latent parameters (20, 50, 100, 500) and (ii) the size of the shared hidden layer (250-1400), informed by scaling principles proposed by Brenner et al. (14). The teacher forcing ratio was gradually reduced from 0.5 to 0.15 during training (28). The learning rate was set to 0.0001 for shared parameters and 0.001 for individual parameters (14). A batch size of 2,000 (250 subjects per batch) was used, following recommendations in (14).

For each hyperparameter configuration, the model was first trained on the training set for 1,000 epochs. Then the group-level parameters were kept fixed, and we ran subject-specific fine-tuning on the test set for an additional 1,000 epochs, during which only individual parameters were updated, see Figure 1.

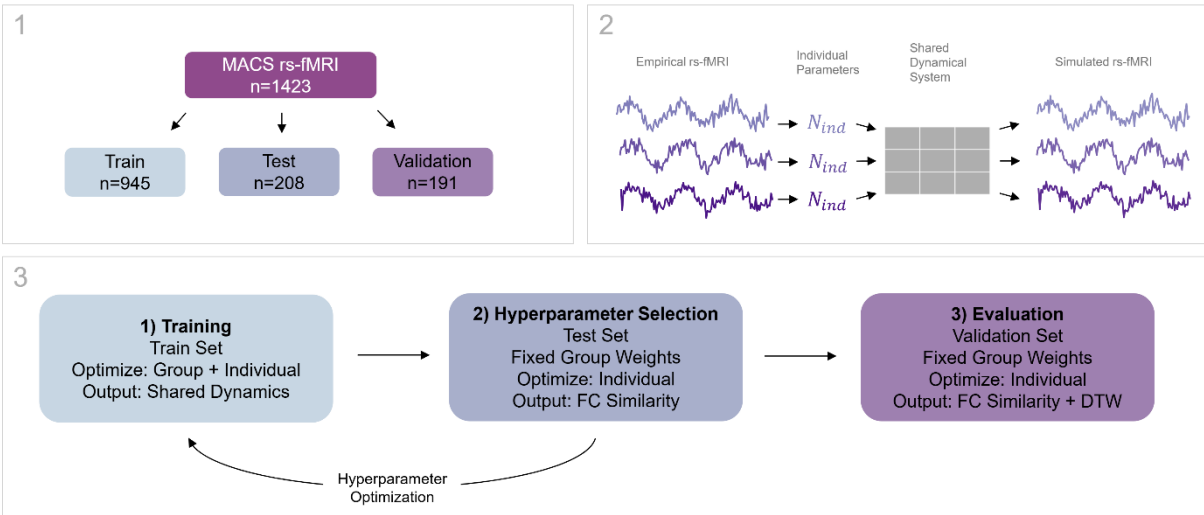

Figure 1: Study overview and modeling workflow. 1) Subject-wise split of the MACS resting-state fMRI dataset into training, test and independent validation sets to prevent data leakage. 2) Conceptual illustration of the hierarchical shPLRNN architecture with shared group-level recurrent dynamics and low-dimensional subject-specific individual parameters. 3) Training, hyperparameter selection and evaluation procedure. Group-level and individual parameters are jointly optimized on the training set. Hyperparameters are selected on a separate test set while keeping group-level parameters fixed. Final performance is assessed on an independent validation set after subject-specific fine-tuning.

2.4 Model Evaluation and Baseline Analyses

Model performance was quantified using the Pearson correlation between empirical and simulated FC matrices after fine-tuning. The hyperparameter configuration maximizing this criterion was selected for all subsequent analyses.

Generalization was assessed on the independent validation set, which was not used during optimization. Performance on training, test and validation sets was quantified using (i) Pearson correlation between the vectorized upper-triangle of the empirical and simulated FC

and (ii) dynamic time warping (DTW) distance between empirical and simulated time series. DTW computes the minimum-cost alignment between two sequences by allowing local stretching/compression of the time axis to match similar patterns despite timing shifts (30). DTW was included as a complementary temporal similarity measure, while recognizing that the primary modeling objective is the reproduction of emergent properties such as FC rather than exact temporal alignment.

To assess inter-individual differences in reconstruction quality within the validation set, we quantified how well each subject's empirical FC pattern aligned with the dominant group-level connectivity structure learned from the training data. Specifically, we computed a template similarity score for each validation subject. To this end, FC matrices were first averaged across all training subjects to obtain a group-mean FC template. For each validation subject, the Pearson correlation coefficient between its empirical FC matrix and the group-mean FC was then calculated. This template similarity score reflects the degree to which a subject's connectivity profile conforms to the canonical connectivity pattern of the training cohort. We examine how strongly this measure predicts individual reconstruction accuracy, thereby assessing whether subjects with more typical FC patterns are modeled more accurately by the hierarchical RNN.

To contextualize the performance of the hierarchical shPLRNN (hier-shPLRNN), we conducted several complementary baseline and robustness analyses. First, we trained a reduced hier-shPLRNN variant in which only a single subject-specific parameter was used per individual, thereby limiting the model's capacity to capture individual differences (Supplement S1). Second, to evaluate the dependence of model performance on training set size, we repeated the optimization procedure using progressively smaller subsets of the data (500, 100 and 50 subjects; Supplement S2). Finally, to assess the stability and robustness of the learned dynamics, we repeated the full hier-shPLRNN training procedure five times with independent random initializations and compared the resulting simulated FCs across runs (Supplement S3).

2.5 Subject Identification and Test-Retest Analysis

To evaluate the degree of individualization captured by the hier-shPLRNN-based simulation within the train set, we implemented a subject identification task based on FC patterns (11). For each subject, we calculated the Pearson correlation coefficient between their simulated FC matrix and all empirical FC matrices in the dataset. We then assessed the frequency with which the empirical FC of the correct subject had the highest correlation with its simulated counterpart (rank 1), was ranked among the top three or within the top five most correlated empirical FCs. This analysis quantifies the discriminability and subject-specific fidelity of the RNN-based representations (11,31).

To assess the test-retest stability and subject specificity of hier-shPLRNN parameters, we compared within-subject and between-subject similarity across repeated fMRI sessions within the train set. For each subject, model parameters were estimated independently for each available session. Subjects with two sessions were included in the analysis (n=308). Parameter similarity between two sessions was quantified using the Pearson correlation

coefficient computed over the full parameter vector (20 parameters). Prior to similarity computation, parameter values were z-standardized across subjects to account for scale differences between training sessions. As a reference baseline, an analogous analysis was performed on empirical FC. FC matrices were vectorized by extracting the upper triangular part of the matrix (excluding the diagonal). Test-retest FC similarity was defined as the Pearson correlation between the corresponding FC vectors of two sessions. For each subject, within-subject similarity was computed as the correlation between session pairs. Between-subject similarity was estimated using a matched procedure: for each within-subject comparison, a randomly selected session from a different subject was used, ensuring equal numbers of within- and between-subject comparisons. Statistical significance of within-subject similarity exceeding between-subject similarity was assessed using a one-sided permutation test (20,000 permutations). In addition, a one-sided Mann-Whitney U test was used as a non-parametric robustness check. To quantify the magnitude of the within-between separation, effect sizes were computed using both Cohen's d (standardized mean difference) and Cliff's delta (non-parametric effect size).

Finally, the relationship between data-level and model-level stability was assessed by correlating subject-wise FC test-retest similarity with parameter test-retest similarity using Pearson and Spearman correlation coefficients.

2.6 Statistical Analysis of Individual Parameters

One aim of this study is to evaluate whether the individual parameters possess sufficient depth to account for individual characteristics of the participants. Specifically, we examined sex, age, body mass index (BMI), years of schooling (YoS) and intelligence quotient (IQ) for each participant.

We employed an Analysis of Variance (ANOVA) model for sex and ordinary least squares (OLS) regression models for age, BMI, years of schooling (YoS) and IQ (continuous variables), predicting each target variable from the full set of model parameters, similar to the methods used in (32). All models included age, sex and scanning site as covariates (except that the outcome itself was not included as a covariate).

In the case of OLS, the general form of the regression was:

$$Y = \sum_{i=1}^{p} \beta_i X_i + \sum_{j=1}^{k} \gamma_j C_j + \epsilon,$$

where Y is the continuous target variable (e.g., age), X_i are the model parameters, C_j are covariates (e.g. sex, site) and ϵ is the residual error.

For the ANOVA on sex, we modeled the categorical group variable using the same covariates (age, site) and calculated effect sizes using partial eta-squared ($\eta^2_{partial}$). All models were estimated independently for each parameter, consistent with standard univariate mass-analysis procedures in neuroimaging. To account for multiple comparisons across parameters, we applied a false discovery rate (FDR) correction using the Benjamini-Yekutieli procedure with a threshold of $q < 0.05$ (33).

For each characteristic, the variable showing the strongest effect was selected and $\eta^2_{partial}$ (for categorical variables) or standardized β (for continuous variables) was calculated as a measure of effect size. Bootstrap confidence intervals were computed using the Bias-corrected and accelerated (BCa) bootstrap method, including group stratification.

To quantify the predictive potential of the variables showing the largest group effect for the categorical sex variable, a logistic regression was fitted on the confounded residuals of the linear models. Specifically, the covariates used in the analyses were regressed out of the data using the same linear models as described above, while excluding the group factor sex from the model first. The probabilities obtained from the logistic regression model were then used to plot a Receiver Operating Characteristic (ROC) curve and calculate the Area Under the Curve (AUROC) (34).

To quantify the predictive potential of the variables showing the largest effects for continuous targets, we calculated mean absolute error (MAE) between the target variables and the predictions of a regression model using the variable showing the largest effect as a single predictor. The target variable showing the largest association with the predictor in this regression model was deconfounded with respect to the defined covariates, thereby estimating the individual effect on the target variable.

2.7 Machine learning analysis of individual parameters

As ML algorithms are known for their ability to discover complex patterns across multiple parameters, we repeated the above tests using multivariate ML analysis instead of univariate statistical tests (21,35). To avoid data leakage, we only used the first time point for each subject, resulting in n=606 samples. The input to the ML algorithms consisted of the individual RNN feature vector and an algorithm was trained to predict each of the individual characteristics mentioned earlier.

Various ML models were trained, optimized and evaluated to predict the individual characteristics based on the model parameters. Each ML pipeline consisted of sequential preprocessing steps, including missing value imputation, feature normalization and either univariate feature selection (based on ANOVA F-values) or principal component analysis (PCA), followed by either a classification (for sex) or regression (for age, BMI, YoS, IQ) algorithm.

Imputation was done using the mean value of each feature within the training set (SimpleImputer, sklearn). Features were scaled using a RobustScaler to mitigate the influence of outliers. Feature selection was based on univariate F-tests with thresholds at 5%, 10% and 50% of retained features; alternatively, PCA was applied with full variance decomposition. The dimensionality reduction strategy was selected dynamically using a PHOTONAI Switch element during model optimization (36).

The classifiers/regressors included support vector machines (SVM), random forests, logistic regression (with L1, L2, or elastic net penalties), k-nearest neighbors (k = 5, 10, 15), Gaussian naïve Bayes and gradient boosting classifiers or regressors. Each algorithm underwent

hyperparameter tuning within a nested cross-validation framework (10 inner, 10 outer folds) (37). For example, SVMs were evaluated with linear, polynomial and RBF kernels across a wide range of C-values (10^{-8} to 10^{8}); logistic regression used C-values from 0.0001 to 10,000.

Model performance was evaluated using R^2 scores for regression tasks and balanced accuracy for classification. All ML analyses were performed using PHOTONAI (36), similar to the pipeline in (38).

2.8 Baseline Representations

As a reference approach to assess subject-specific characteristics independent of the hier-shPLRNN model, we additionally performed a PCA across all subjects. For this PCA, each subject's entire fMRI time series (all time points and ROIs) was vectorized into a single feature vector. We used the same number of PCA parameters as for the subject-specific parameters within the hier-shPLRNN. The principal components derived from this analysis provided subject-level representations within a common component space, capturing the dominant axes of inter-individual variance (39). The individual component scores served as a data-driven baseline, enabling comparison with the individual parameters estimated by the hier-shPLRNN and with demographic or behavioral variables such as age and sex. The resulting PCA parameters were used as input to the statistical and ML tests and compared their performance to the performance of the RNN parameters. Additionally, we also used the empirical rs-fMRI as input as another baseline comparison.

3. Results

3.1 Hyperparameter Optimization

Hyperparameter optimization identified a hier-shPLRNN configuration with 20 individual parameters and a shared hidden size of 1294 as optimal according to the optimization criterion, defined as the Pearson correlation between empirical and simulated FC after subject-specific fine-tuning. This configuration was used for all subsequent analyses of reconstruction performance and generalization. See Supplement Figure S4 for further details.

3.2 Reconstruction Performance across Train, Test and Validation Sets

Using the selected configuration, model performance was quantified on the training, test and independent validation sets using two metrics: (i) Pearson correlation between empirical and simulated FC matrices and (ii) normalized DTW distance between empirical and simulated multivariate time series.

On the training set (n = 945), the model achieved a mean FC correlation of r = 0.63 and a mean normalized DTW of 0.16. On the test set used during hyperparameter optimization (n = 208), mean performance after fine-tuning was r = 0.51 with a normalized DTW of 0.20. On the independent validation set (n = 191), performance was comparable, with a mean FC correlation of r = 0.51 and a normalized DTW of 0.16. These results are summarized in Figure 2. Examples of empirical and simulated time series and FC matrices for representative validation subjects are shown in Figures 3.

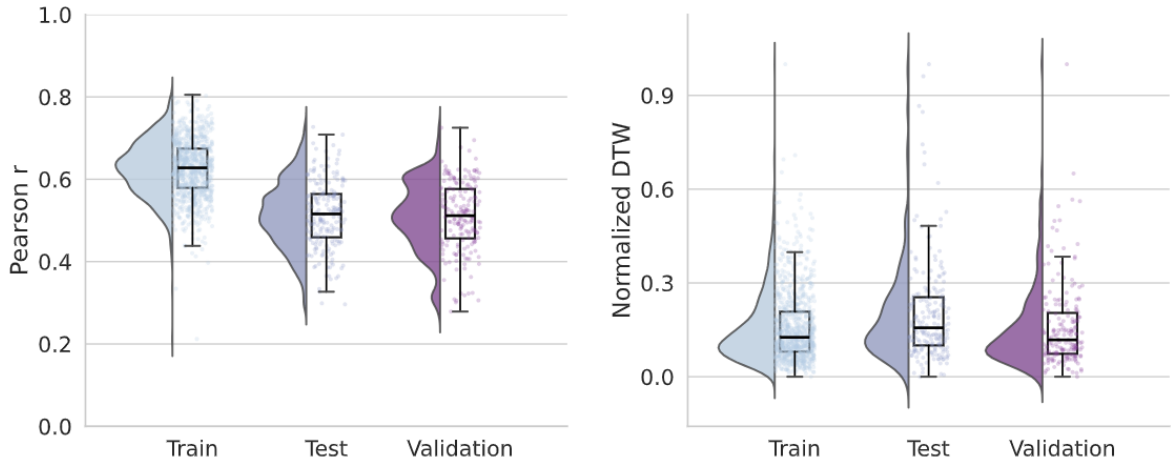

Figure 2: Reconstruction performance across datasets. Left: Pearson correlation between empirical and simulated FC matrices. Right: normalized DTW distance between empirical and simulated time series. Training subjects were optimized jointly (shared + individual parameters). For test and validation subjects, only individual parameters were optimized during subject-specific fine-tuning while shared parameters were fixed.

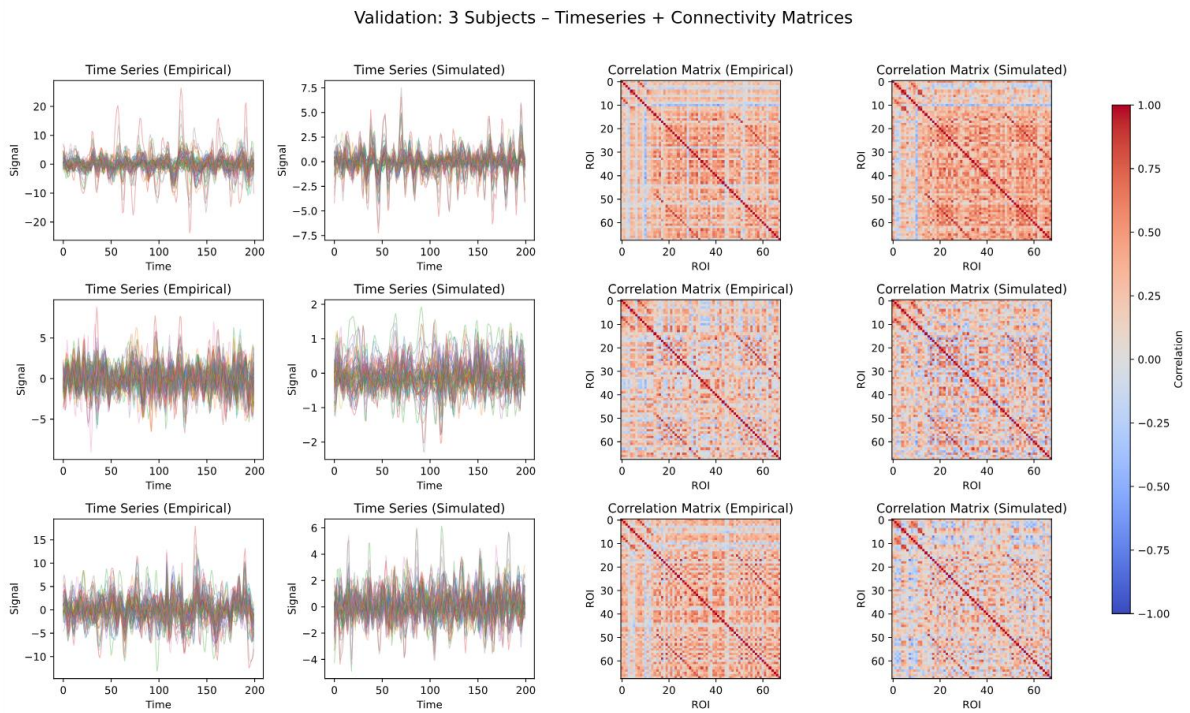

Figure 3: Simulated vs. empirical rs-fMRI time series for three exemplary validation subjects after fine-tuning. Left: Time series of rs-fMRI activity across 68 brain regions and 200 timepoints, comparing simulated (left column) and empirical (right column) data for three validation subjects. Right: Functional connectivity (FC) matrices based on pairwise Pearson correlations between 68 brain regions, shown for simulated (left column) and empirical (right column) rs-fMRI data. Each row corresponds to a different test subject.

3.3 Subject-wise Variability and Template Similarity

Although average reconstruction performance on the validation set was robust, reconstruction accuracy varied substantially across subjects. To assess whether reconstruction accuracy depends on how typical a subject's FC pattern is relative to the training cohort, we computed a template similarity score for each validation subject. Template similarity was defined as the

Pearson correlation between that subject's empirical FC and the group-mean FC template obtained by averaging empirical FC matrices across all training subjects.

Reconstruction accuracy increased with template similarity (Figure 4). A linear regression model with template similarity as the sole predictor explained $R^2 = 0.37$ of the variance in reconstruction accuracy across validation subjects.

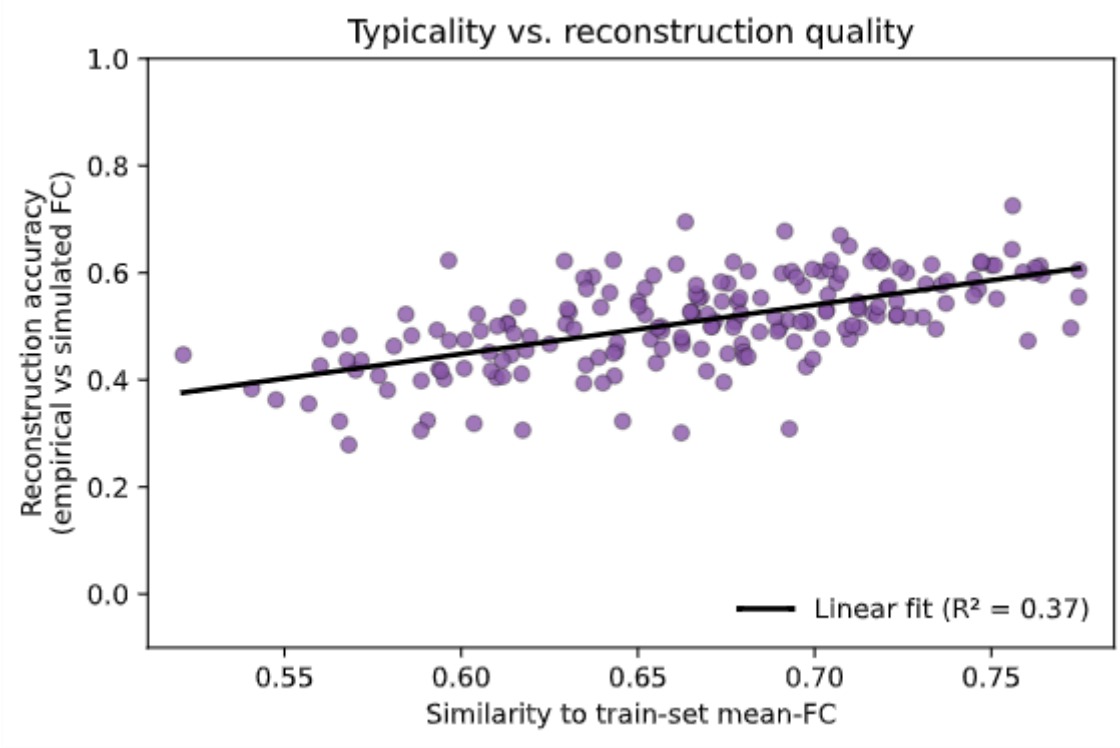

Figure 4. Typicality of individual functional connectivity predicts reconstruction accuracy. Each dot represents one validation subject. x-axis: template similarity (correlation with group-mean training FC). y-axis: reconstruction accuracy (correlation between empirical FC and simulated FC after fine-tuning). The fitted linear regression explains $R^2 = 0.37$ of variance in reconstruction accuracy.

3.4 Subject Identification based on FC

To evaluate the degree of individual fingerprints captured by the hier-shPLRNN simulation, we performed an FC-based subject identification analysis within the training set. For each subject, we correlated the simulated FC matrix with all empirical FC matrices and recorded the rank of the correct subject's empirical FC.

The correct subject was ranked Top-1 in 96/945 cases (10.16%), Top-3 in 178/945 cases (18.84%) and Top-5 in 226/945 cases (23.92%). Chance-level Top-1 identification is approximately 0.1% (1/945). The full correlation matrix underlying the FC-based subject identification analysis is shown in Supplementary Figure S5.

3.5 Test-retest Stability of Individual Parameters

To assess test-retest stability, we analyzed 308 subjects from the training set with two rs-fMRI sessions. Individual parameter vectors were estimated independently for each session and compared using Pearson correlation across the full parameter vector.

Hier-shPLRNN parameters showed higher within-subject similarity than between-subject similarity. Mean within-subject similarity was r = 0.286 ± 0.238, whereas mean between-subject similarity was r = 0.027 ± 0.240, yielding Δr = 0.259. This separation was significant (permutation test: $p = 5\times10^{-5}$; Mann-Whitney U: $p = 7.55\times10^{-33}$) and associated with large effect sizes (Cohen's d = 1.083; Cliff's δ = 0.553).

Empirical FC exhibited higher absolute test-retest similarity. Within-subject FC similarity was r = 0.613 ± 0.083, compared to between-subject similarity of r = 0.459 ± 0.078 (Δr = 0.154), with significant separation (permutation: $p = 5\times10^{-5}$; Mann-Whitney U: $p = 4.63\times10^{-73}$) and large effect sizes (Cohen's d = 1.914; Cliff's δ = 0.840).

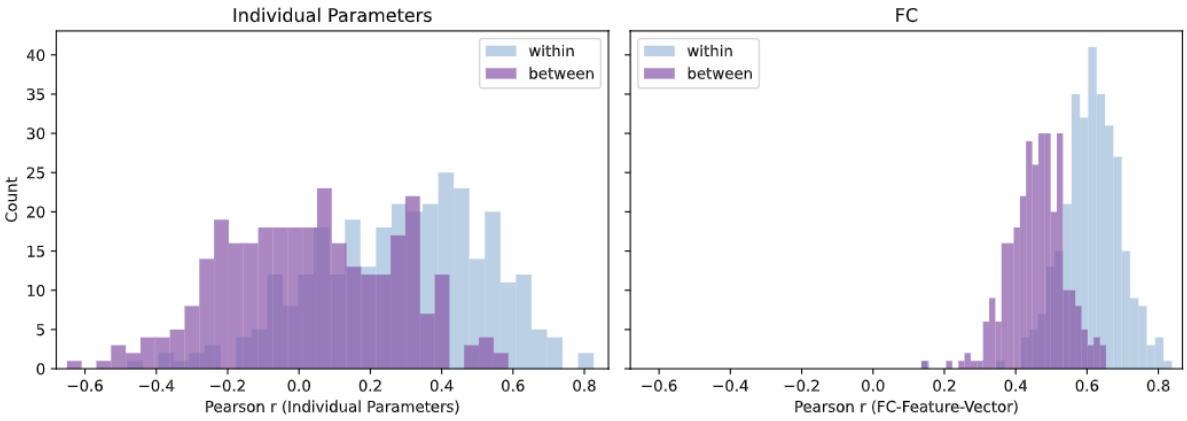

Figure 5. Test-retest similarity of individualized model parameters and functional connectivity. Left: Pearson correlation between subject-specific shPLRNN parameter vectors estimated from two rs-fMRI sessions, shown for within-subject and between-subject comparisons. Right: Corresponding test-retest similarity of empirical functional connectivity (FC), quantified as the Pearson correlation between vectorized FC matrices. Within-subject similarity exceeds between-subject similarity in both modalities, indicating subject-specific and reproducible representations.

Across subjects, parameter test-retest similarity was moderately correlated with FC test-retest similarity (Pearson r = 0.267, $p = 2.02\times10^{-6}$; Spearman r = 0.260, $p = 3.89\times10^{-6}$) (Figure 6).

3.6 Associations between Learned Parameters and Subject Characteristics

In univariate statistical analyses, hier-shPLRNN-derived parameters showed statistically significant associations (FDR-corrected $p < 0.05$) with sex, age and BMI, with modest effect sizes ($|\beta_{standardized}| \leq 0.37$; $\eta^2_{partial} = 0.021$). PCA-derived parameters yielded fewer significant results and smaller effect sizes. Using empirical rs-fMRI features directly resulted in stronger associations across variables (Table 1).

Characteristic	Model	p-uncorrected	p-corrected	$\eta^2_{partial}$	Overlap	BACC	AUROC
Sex	hier-shPLRNN	<0.001	<0.001	0.021	87 %	55.3 %	0.59
	PCA	0.016	1	0.006	88.7 %	54.6 %	0.55
	rs-fMRI	<0.001	<0.001	0.05	81 %	59.8 %	0.64

Characteristic	Model	p-uncorrected	p-corrected	$\beta_{standardized}$	MAE
Age	hier-shPLRNN	<0.001	<0.001	-0.372	34.59
	PCA	0.016	1	-0.081	34.36
	rs-fMRI	<0.001	<0.001	-0.34	34.21
BMI	hier-shPLRNN	<0.001	<0.001	-0.141	25.11
	PCA	0.001	0.072	0.116	24.79
	rs-fMRI	<0.001	<0.001	0.204	24.16
Years of Schooling	hier-shPLRNN	0.004	0.288	-0.095	15.56
	PCA	0.035	1	-0.071	14.38
	rs-fMRI	<0.001	<0.001	-0.143	14.05
IQ	hier-shPLRNN	0.136	1	-0.054	114.38
	PCA	0.015	1	-0.089	114.09
	rs-fMRI	<0.001	<0.001	-0.136	114.32

Table 1: Comparison of predictive performance across feature types for demographic and physiological variables. Results of statistical analyses predicting sex, age, BMI, years of schooling and IQ using three types of input features: hier-shPLRNN-derived parameters, PCA-derived parameters and empirical rs-fMRI data.

3.7 Machine Learning Prediction of Subject Characteristics

Across all machine learning models evaluated using hier-shPLRNN-derived parameters as input, the highest balanced accuracy for sex prediction was BACC = 0.7. For regression tasks, the best-performing models achieved MSE values of 76.89 (age), 12.4 (BMI), 6.04 (years of schooling) and 142.52 (IQ). Best results per target and feature type are summarized in Table 2. Overall, models using hier-shPLRNN-derived parameters outperformed PCA-derived parameters, while models trained on empirical rs-fMRI features performed best across almost all targets.

Characteristic	Model	BACC
Sex	hier-shPLRNN	0.7
	PCA	0.64
	rs-fMRI	0.78

Characteristic	Model	MSE	MAE	explained variance
age	hier-shPLRNN	76.89	6.76	0.49
	PCA	137.09	10.17	0.01
	rs-fMRI	59.41	6.23	0.6
BMI	hier-shPLRNN	12.4	2.82	0.18
	PCA	15.09	3.12	0.006
	rs-fMRI	9.04	2.4	0.39
Years of Schooling	hier-shPLRNN	6.04	2.16	0.05
	PCA	6.34	2.15	0.005
	rs-fMRI	6.03	2.16	0.05
IQ	hier-shPLRNN	142.52	10.03	0.04
	PCA	151.2	10.48	-0.01
	rs-fMRI	145.06	10.37	0.05

Table 2: Performance of machine learning models predicting individual characteristics using RNN-derived parameters. Summary of the best-performing machine learning algorithm for each target variable using hier-shPLRNN-derived parameters, PCA-derived parameters, and empirical rs-fMRI data as input features.

4. Discussion

In this study, we systematically evaluated a hier-shPLRNN-based dynamical system reconstruction framework for simulating individual rs-fMRI dynamics. The model was designed to disentangle shared population-level dynamics from compact, subject-specific representations while operating on relatively short and noisy time series. Our results provide insights into the model's reconstruction performance, generalization behavior and the extent to which individualized parameters capture stable subject-specific information.

A central methodological result concerns the outcome of the hyperparameter optimization. For the rs-fMRI data, the optimal configuration identified by Optuna closely followed the scaling principles previously proposed for low-dimensional chaotic systems (13,14), once appropriately adjusted for the dimensionality of the data. In particular, the optimal hidden layer size was consistent with the values suggested in (14) when scaled to the number of observed brain regions, indicating that these theoretical considerations generalize beyond synthetic benchmark systems to high-dimensional empirical neuroimaging data. This finding suggests that the shared recurrent backbone of the hier-shPLRNN requires sufficient expressive capacity to capture the dominant population-level dynamics present in rs-fMRI and that this capacity can be guided by principled scaling rules rather than ad hoc tuning.

In contrast, the optimization results for the number of individual parameters revealed a qualitatively different pattern. Rather than benefiting from a large number of subject-specific parameters, model performance peaked at a relatively small latent dimensionality. Increasing the number of individual parameters beyond this point did not improve reconstruction accuracy and in some configurations led to degraded performance. This finding indicates that, at least in the present setting, individualized dynamics are best captured by a compact representation and that simply increasing the number of subject-specific degrees of freedom does not necessarily enhance personalization. Instead, excessive individual parameterization may dilute the shared structure learned at the group level or increase sensitivity to noise, reinforcing the notion that more parameters do not automatically translate into better individual modeling.

Using the optimized configuration, the model achieved strong reconstruction performance on the training data and generalized to an independent validation set after subject-specific fine-tuning of only the individual parameters. While average reconstruction accuracy on unseen subjects was comparable to that observed during hyperparameter optimization, performance varied substantially across individuals. This variability was systematically related to how typical a subject's FC pattern was relative to the training cohort. Subjects whose empirical FC closely matched the group-mean FC were reconstructed with markedly higher accuracy, whereas subjects with more idiosyncratic connectivity profiles were modeled less well. Template similarity alone accounted for approximately 37% of the variance in reconstruction accuracy across validation subjects, highlighting the strong influence of population-level structure on individual-level generalization.

Conceptually, these results indicate that the hier-shPLRNN learns a population-level dynamical prior that captures a dominant connectivity manifold shared across subjects. Individual parameters then enable adaptation around this shared template. Generalization failures therefore arise primarily for subjects whose connectivity patterns fall outside the learned manifold, rather than from uniform overfitting to the training data. This distinction is important, as it suggests that improving robustness to atypical or out-of-distribution connectivity patterns, through stronger regularization, explicit modeling of deviations or domain adaptation, may be more impactful than further refinement of the fine-tuning procedure alone.

From a dynamical systems perspective, the hier-shPLRNN does not aim to reproduce empirical rs-fMRI time series point by point, but to reconstruct an underlying generative process whose long-term behavior gives rise to the observed data. In this view, successful modeling is reflected in the preservation of emergent properties such as subject-specific functional connectivity profiles, rather than in exact temporal alignment. The fact that the model robustly reproduces individual FC structure, despite variability in time-series reconstruction accuracy, is therefore consistent with its interpretation as a dynamical system reconstruction approach. At the same time, the dependence of generalization performance on template similarity suggests that the inferred dynamics primarily capture a dominant population-level attractor structure, around which individual dynamics are expressed as perturbations. Subjects whose empirical connectivity deviates strongly from this shared dynamical manifold are consequently more difficult to model, highlighting both the strengths and current limitations of this generative, dynamics-based framework.

Beyond reconstruction accuracy, we assessed whether the learned individual parameters capture stable and subject-specific information. Test-retest analyses demonstrated significantly higher within-subject than between-subject similarity of parameter vectors across repeated rs-fMRI sessions. Although empirical FC exhibited higher absolute test-retest similarity, the learned parameters showed a pronounced separation between within- and between-subject similarity, consistent with the idea that the model compresses high-dimensional connectivity information into a lower-dimensional representation that emphasizes subject-specific structure. The moderate correlation between FC stability and parameter stability further suggests that these parameters reflect both data reliability and additional aspects of brain dynamics not fully captured by static FC measures.

When relating the individualized parameters to demographic and psychometric variables, the observed associations were statistically significant but modest in effect size, consistent with prior work on rs-fMRI-based prediction (7). Subject identification based on simulated FC patterns showed above chance discriminability and predictive analyses using machine learning models achieved moderate performance, with the strongest results observed for sex and age.

Taken together, these results suggest that individualized brain dynamics can be captured by relatively low-dimensional representations, and that increasing the number of subject-specific parameters does not necessarily improve performance. In our setting, model accuracy peaked at a compact latent dimensionality, indicating that excessive parameterization may dilute shared structure or amplify noise rather than enhance individual specificity. While low-dimensional representations facilitate generalization and stability, they may still limit sensitivity to subtle individual differences, particularly when time series are short and noisy. Future work could therefore explore more flexible mechanisms for capturing individual variability, such as adaptive latent dimensionality or richer hierarchical structures.

In summary, hier-shPLRNNs provide a principled and interpretable framework for modeling individual brain dynamics from rs-fMRI data, particularly in settings where time series are short and noisy. Such limited data per subject pose a fundamental challenge for individualized modeling, as they constrain the reliable estimation of complex dynamical systems and

increase the risk of overfitting or noise-driven effects. Within this context, our results show that hierarchical modeling with compact subject-specific representations enables robust reconstruction and generalization, while excessive parameterization does not improve performance.

At the same time, several limitations remain. The model showed reduced performance for subjects with atypical connectivity patterns, indicating limited robustness to out-of-distribution dynamics. Moreover, associations between learned parameters and behavioral or demographic variables were modest, suggesting that the current representation captures only part of the relevant individual variability. Finally, the reliance on relatively short time series may inherently limit the recoverable complexity of individual dynamics.

Future work could address these limitations by incorporating longer or multimodal recordings, developing adaptive or more expressive hierarchical architectures, and improving robustness to atypical or heterogeneous connectivity patterns. Such extensions may help to better capture fine-grained individual differences while preserving the benefits of shared structure.

Acknowledgments

This work was funded in part by the consortia grants from the German Research Foundation (DFG) FOR 2107, SFB/TRR 393 (“Trajectories of Affective Disorders”, project grant no 521379614), and the Germany’s Excellence Strategy (EXC 3066/1 “The Adaptive Mind”, Project No. 533717223), as well as the DYNAMIC center, funded by the LOEWE program of the Hessian Ministry of Science and Arts (grant number: LOEWE1/16/519/03/09.001(0009)/98).

During the preparation of this work the authors used ChatGPT to improve language and readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Bibliography

1. Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE. Methods to detect, characterize, and remove motion artifact in resting state fMRI. NeuroImage. 2014 Jan;84:320–41. doi:10.1016/j.neuroimage.2013.08.048

2. Smith SM, Beckmann CF, Andersson J, Auerbach EJ, Bijsterbosch J, Douaud G, et al. Resting-state fMRI in the Human Connectome Project. NeuroImage. 2013 Oct;80:144–68. doi:10.1016/j.neuroimage.2013.05.039

3. Gordon EM, Laumann TO, Gilmore AW, Newbold DJ, Greene DJ, Berg JJ, et al. Precision Functional Mapping of Individual Human Brains. Neuron. 2017 Aug;95(4):791-807.e7. doi:10.1016/j.neuron.2017.07.011

4. Wang D, Buckner RL, Fox MD, Holt DJ, Holmes AJ, Stoecklein S, et al. Parcellating cortical functional networks in individuals. Nat Neurosci. 2015 Dec;18(12):1853–60. doi:10.1038/nn.4164

5. Kong R, Li J, Orban C, Sabuncu MR, Liu H, Schaefer A, et al. Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion. Cerebral Cortex. 2019 Jun 1;29(6):2533–51. doi:10.1093/cercor/bhy123

6. Kong R, Yang Q, Gordon E, Xue A, Yan X, Orban C, et al. Individual-Specific Areal-Level Parcellations Improve Functional Connectivity Prediction of Behavior. Cerebral Cortex. 2021 Aug 26;31(10):4477–500. doi:10.1093/cercor/bhab101

7. Barkhau CBC, Pellengahr C, Wang Z, Fisch L, Leenings R, Winter NR, et al. Toward Personalized Neuroscience: Evaluating Individual-Level Information in Neural Mass Models. Human Brain Mapping. 2025 Nov;46(16):e70413. doi:10.1002/hbm.70413

8. Griffiths JD, Wang Z, Ather SH, Momi D, Rich S, Diaconescu A, et al. Deep Learning-Based Parameter Estimation for Neurophysiological Models of Neuroimaging Data [Internet]. Neuroscience; 2022 [cited 2025 Mar 18]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.05.19.492664 doi:10.1101/2022.05.19.492664

9. Schirner M, Deco G, Ritter P. Learning how network structure shapes decision-making for bio-inspired computing. Nat Commun. 2023 May 23;14(1):2963. doi:10.1038/s41467-023-38626-y

10. Dadi K, Rahim M, Abraham A, Chyzhyk D, Milham M, Thirion B, et al. Benchmarking functional connectome-based predictive models for resting-state fMRI. NeuroImage. 2019 May;192:115–34. doi:10.1016/j.neuroimage.2019.02.062

11. Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat Neurosci. 2015 Nov;18(11):1664–71. doi:10.1038/nn.4135

12. Durstewitz D, Koppe G, Thurm MI. Reconstructing computational system dynamics from neural data with recurrent neural networks. Nat Rev Neurosci. 2023 Nov;24(11):693–710. doi:10.1038/s41583-023-00740-7

13. Koppe G, Toutounji H, Kirsch P, Lis S, Durstewitz D. Identifying nonlinear dynamical systems via generative recurrent neural networks with applications to fMRI. Isik L, editor. PLoS Comput Biol. 2019 Aug 21;15(8):e1007263. doi:10.1371/journal.pcbi.1007263

14. Brenner M, Weber E, Koppe G, Durstewitz D. LEARNING INTERPRETABLE HIERARCHICAL DYNAMI- CAL SYSTEMS MODELS FROM TIME SERIES DATA [Internet]. 2025;The Thirteenth International Conference on Learning Representations. Available from: https://openreview.net/forum?id=Vp2OAxMs2s

15. Farahibozorg SR, Bijsterbosch JD, Gong W, Jbabdi S, Smith SM, Harrison SJ, et al. Hierarchical modelling of functional brain networks in population and individuals from big fMRI data. NeuroImage. 2021 Nov;243:118513. doi:10.1016/j.neuroimage.2021.118513

16. Cui Y, Zhao S, Wang H, Xie L, Chen Y, Han J, et al. Identifying Brain Networks at Multiple Time Scales via Deep Recurrent Neural Network. IEEE J Biomed Health

Inform. 2019 Nov;23(6):2515–25. doi:10.1109/JBHI.2018.2882885

17. Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, et al. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci. 2017 Feb;18(2):115–26. doi:10.1038/nrn.2016.167

18. Van Horn JD, Toga AW. Human neuroimaging as a "Big Data" science. Brain Imaging and Behavior. 2014 Jun;8(2):323–31. doi:10.1007/s11682-013-9255-y

19. Lindquist M. Neuroimaging results altered by varying analysis pipelines. Nature. 2020 Jun 4;582(7810):36–7. doi:10.1038/d41586-020-01282-z

20. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. NeuroImage. 2012 Feb;59(3):2142–54. doi:10.1016/j.neuroimage.2011.10.018

21. Li H, Zhuo C, Cui Z, Cieslak M, Salo T, Gur RE, et al. Mapping individualized multi-scale hierarchical brain functional networks from fMRI by self-supervised deep learning [Internet]. Neuroscience; 2025 [cited 2025 May 30]. Available from: http://biorxiv.org/lookup/doi/10.1101/2025.04.07.647618 doi:10.1101/2025.04.07.647618

22. Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci. 2017 Mar;20(3):365–77. doi:10.1038/nn.4478

23. Kircher T, Wöhr M, Nenadic I, Schwarting R, Schratt G, Alferink J, et al. Neurobiology of the major psychoses: a translational perspective on brain structure and function—the FOR2107 consortium. Eur Arch Psychiatry Clin Neurosci. 2019 Dec;269(8):949–62. doi:10.1007/s00406-018-0943-x

24. Vogelbacher C, Möbius TWD, Sommer J, Schuster V, Dannlowski U, Kircher T, et al. The Marburg-Münster Affective Disorders Cohort Study (MACS): A quality assurance protocol for MR neuroimaging data. NeuroImage. 2018 May;172:450–60. doi:10.1016/j.neuroimage.2018.01.079

25. De Lange SC, Helwegen K, Van Den Heuvel MP. Structural and functional connectivity reconstruction with CATO - A Connectivity Analysis TOolbox. NeuroImage. 2023 Jun;273:120108. doi:10.1016/j.neuroimage.2023.120108

26. Jenkinson M, Bannister P, Brady M, Smith S. Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. NeuroImage. 2002 Oct;17(2):825–41. doi:10.1006/nimg.2002.1132

27. Boubela RN, Kalcher K, Huf W, Kronnerwetter C, Filzmoser P, Moser E. Beyond Noise: Using Temporal ICA to Extract Meaningful Information from High-Frequency fMRI Signal Fluctuations during Rest. Front Hum Neurosci. 2013;7. doi:10.3389/fnhum.2013.00168

28. Hess F, Monfared Z, Brenner M, Durstewitz D. Generalized Teacher Forcing for Learning Chaotic Dynamics [Internet]. 2023. doi:10.48550/ARXIV.2306.04406

29. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A Next-generation Hyperparameter Optimization Framework [Internet]. arXiv; 2019 [cited 2026 Jan 16].

Available from: https://arxiv.org/abs/1907.10902 doi:10.48550/ARXIV.1907.10902

30. Dynamic Time Warping. In: Information Retrieval for Music and Motion [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2007 [cited 2026 Feb 3]. p. 69–84. Available from: http://link.springer.com/10.1007/978-3-540-74048-3_4 doi:10.1007/978-3-540-74048-3_4

31. Amico E, Goñi J. The quest for identifiability in human functional connectomes. Sci Rep. 2018 May 29;8(1):8254. doi:10.1038/s41598-018-25089-1

32. Winter NR, Leenings R, Ernsting J, Sarink K, Fisch L, Emden D, et al. Quantifying Deviations of Brain Structure and Function in Major Depressive Disorder Across Neuroimaging Modalities. JAMA Psychiatry. 2022 Sep 1;79(9):879. doi:10.1001/jamapsychiatry.2022.1780

33. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Statist. 2001 Aug 1;29(4). doi:10.1214/aos/1013699998

34. Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006 Jun;27(8):861–74. doi:10.1016/j.patrec.2005.10.010

35. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. NeuroImage. 2017 Jan;145:137–65. doi:10.1016/j.neuroimage.2016.02.079

36. Leenings R, Winter NR, Plagwitz L, Holstein V, Ernsting J, Sarink K, et al. PHOTONAI—A Python API for rapid machine learning model development. Gadekallu TR, editor. PLoS ONE. 2021 Jul 21;16(7):e0254062. doi:10.1371/journal.pone.0254062

37. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. NeuroImage. 2017 Jan;145:166–79. doi:10.1016/j.neuroimage.2016.10.038

38. Winter NR, Blanke J, Leenings R, Ernsting J, Fisch L, Sarink K, et al. A Systematic Evaluation of Machine Learning–Based Biomarkers for Major Depressive Disorder. JAMA Psychiatry. 2024 Apr 1;81(4):386. doi:10.1001/jamapsychiatry.2023.5083

39. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Phil Trans R Soc A. 2016 Apr 13;374(2065):20150202. doi:10.1098/rsta.2015.0202