Latent Similarity Identifies Important Functional Connections for Phenotype Prediction

Anton Orlichenko, Gang Qu, Gemeng Zhang, Binish Patel, Tony W. Wilson, Julia M. Stephen, Vince D. Calhoun, *Fellow, IEEE,* and Yu-Ping Wang, *Senior Member, IEEE*

***Abstract*— *Objective*: Endophenotypes such as brain age and fluid intelligence are important biomarkers of disease status. However, brain imaging studies to identify these biomarkers often encounter limited numbers of subjects but high dimensional imaging features, hindering reproducibility. Therefore, we develop an interpretable, multivariate classification/regression algorithm, called Latent Similarity (LatSim), suitable for small sample size but high feature dimension datasets. *Methods*: LatSim combines metric learning with a kernel similarity function and softmax aggregation to identify task-related similarities between subjects. Inter-subject similarity is utilized to improve performance on three prediction tasks using multi-paradigm fMRI data. A greedy selection algorithm, made possible by LatSim's computational efficiency, is developed as an interpretability method. *Results*: LatSim achieved significantly higher predictive accuracy at small sample sizes on the Philadelphia Neurodevelopmental Cohort (PNC) dataset. Connections identified by LatSim gave superior discriminative power compared to those identified by other methods. We identified 4 functional brain networks enriched in connections for predicting brain age, sex, and intelligence. *Conclusion*: We find that most information for a predictive task comes from only a few (1-5) connections. Additionally, we find that the default mode network is overrepresented in the top connections of all predictive tasks. *Significance*: We propose a novel prediction algorithm for small sample, high feature dimension datasets and use it to identify connections in task fMRI data. Our work can lead to new insights in both algorithm design and neuroscience research.**

***Index Terms*— Default mode network, fMRI, functional connectivity, metric learning, PNC, small sample size**

I. Introduction

FUNCTIONAL magnetic resonance imaging (fMRI) provides a non-invasive estimate of brain activity by exploiting the blood oxygen level-dependent (BOLD) signal [1]. This high-acuity imaging data can be used to predict variables like age, sex, intelligence, and disease status [2] [3] [4] [5]. Interestingly, the gap between fMRI-predicted brain age and biological age can identify Alzheimer's disease patients prior to the onset of symptoms [6]. Prediction is hindered, however, by the combination of small sample size and very high feature number. This results in models that have poor reproducibility and generalizeability [7].

Studies with small sample size only have the power to detect very large effects. Many effects that are found in small studies may be due to noise. When identifying regions that are associated with in-scanner tasks, it was found that the average minimum cohort size needed to reproducibly identify the same region 50% of the time in independent samples was N=36 [8]. In contrast, models deployed clinically use thousands of subjects for training and validation [9]. In 2017 and 2018, the median cohort sizes for published experimental and clinical MRI studies were 23 and 24 subjects, respectively, and less than 1% of the 272 papers surveyed reported cohort sizes greater than 100 [10]. This may be attributed to both cost, at $500-$1000 per subject, and the difficulty of collecting the data, stemming from long scan times, subject discomfort in the scanner, and experimental design [10].

Additionally, for fMRI-based predictions to be useful clinically, they must be interpretable. There is a large literature on the interpretability of machine learning in medical imaging [11] [12]; however, there is often a tradeoff between model accuracy and interpretability. This raises questions about robustness in the clinical setting [13]. For example, *Zhang et al.* show that different processing methods can yield similar accuracy in a sex prediction task, but with different discriminative features identified by each method [14]. Identifying a minimal set of valid functional connections may increase model robustness, and make inroads into causal analysis of brain networks [15].

Finally, many recent studies in the deep learning field shift their focus to integrate data from multiple omics [16], or multiple omics and imaging [17]. This is done for two purposes: to improve prediction accuracy and to learn novel interactions between different modalities. CCA-based models have been proposed that use response variable-guided feature

Manuscript received on September 14, 2022. This work was supported in part by NIH grants (P20GM109068, P20GM144641, R01MH121101, R01MH104680, R01MH107354, R01MH103220, R01EB020407, R56MH124925) and NSF grant (#1539067). *(Corresponding author: Yu-Ping Wang.)*

Anton Orlichenko, Gang Qu, Gemeng Zhang, Binish Patel, and Yu-Ping Wang are with the Department of Biomedical Engineering, Tulane University, New Orleans, LA 70118. (e-mail: wyp@tulane.edu).

Julia M. Stephen is with the Mind Research Network, Albuquerque, NM 87106. (e-mail: jstephen@mrn.org).

Tony W. Wilson is with the Institute for Human Neuroscience, Boys Town National Research Hospital, Boys Town, NE 68010. (e-mail: tony.wilson@boystown.org).

Vince D. Calhoun is with the Tri-Institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS) (Georgia State University, Georgia Institute of Technology, Emory University), Atlanta, GA 30303. (e-mail: vcalhoun@gsu.edu).

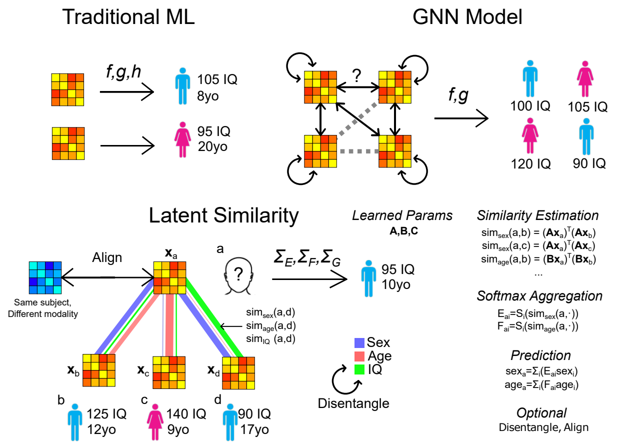


Fig. 1. An overview of the Latent Similarity model. In traditional ML, estimation of response variables is decoupled from inter-subject similarity, whereas GNN models require additional degrees of freedom to estimate edges between subjects. Our model calculates similarity between subjects based on a set of response variables and incorporates multi-modal feature alignment (in addition to ensembling) as well as sparsity and feature disentanglement.

alignment [18] [19]. However, these models do not consider inter-subject relationships and cannot control disentanglement between different predictive tasks.

In this paper, we introduce LatSim (Figure 1), a model in the spirit of metric learning [20], that is both robust and interpretable. Traditional machine learning (ML) models in fMRI, which work directly on functional connectivity (FC) [21], are vulnerable to noise or random confounders like scanner drift or head motion [22]. Graph neural networks (GNN) use inter-subject information as an adjunct to calculations performed directly on FC [23]. However, graph edges may be ambiguous or non-binary, requiring additional degrees of freedom for their estimation [24] [25]. In contrast, LatSim learns an inter-subject similarity metric, $d(\mathbf{x}_a, \mathbf{x}_b)$, and uses the inter-subject similarity, without a self-loop, to make predictions.

The contribution of our work is three-fold. First, we propose a novel metric learning-based model, LatSim, which is robust, interpretable, computationally efficient, multi-view, and multi-task. Second, we use LatSim and a greedy selection algorithm to identify the most discriminitive connections for age, sex, and intelligence prediction among adolescents in the Philadelphia Neurodevelopmental Cohort (PNC) dataset [26]. We show that the such connections are superior to those identified by saliency maps. Third, we give a justification why LatSim performs better than traditional ML models with low sample sizes and high feature dimensionality.

The rest of this paper is organized as follows. Section II gives the mathematical foundations of LatSim and its relationship to other models. Section III provides simulation and experimental results. Section IV discusses significant brain networks and reasons why LatSim performs better in the low sample-size, high-dimensionality regime. Section V concludes with a recapitulation of the work. We make the code publicly available at the link in the footnote.[1]

TABLE I
COMMONLY USED NOTATION.

Notation	Description
$\mathbf{X} \in \mathbb{R}^{N \times d}$	A matrix of dimension N by d
X_{ij}	The (i,j)-th entry of matrix $\mathbf{X}$
$\mathbf{X}_{i,:}$	The i-th row of matrix $\mathbf{X}$
$\mathbf{X}_i, \mathbf{X}^{(i)}$	The i-th matrix in a set of matrices
$\mathbf{X}^{\text{T}}$	The transpose of matrix $\mathbf{X}$
A, B	Random variables
F_i	The i-th element of a set
y_i	The i-th entry of vector y
$\odot$	The Hadamard product
$\mathbf{1}$	A matrix of ones
$\text{diag}(\mathbf{a})$	A square matrix with the elements of $\mathbf{a}$ on the main diagonal, 0s elsewhere
Σ_{abc}	Summation over indices a, b, c
$\mathbb{E}[\cdot]$	Expectation
$\text{Var}[\cdot], \text{Cov}[\cdot]$	Variance, covariance
μ, σ^2, ρ	Mean, variance, correlation
$\lvert C \rvert$	Cardinality of set C
$\lvert\lvert \cdot \rvert\rvert_1$	The l_1 norm
$\lvert\lvert \cdot \rvert\rvert_2$	The l_2 norm

II. METHODS

A. Kernel CCA

To compute similarity between subjects, we utilize ideas from canonical correlation analysis (CCA) [27] [28]. Conventional CCA seeks to find relationships between the features of two different views of a dataset. It aligns the two views, $\mathbf{X}_1 \in \mathbb{R}^{N \times d_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{N \times d_2}$, by finding canonical variables $\mathbf{w}_1$ and $\mathbf{w}_2$ that maximize the correlation between $\mathbf{X}_1\mathbf{w}_1$ and $\mathbf{X}_2\mathbf{w}_2$:

$$\begin{aligned} \underset{\mathbf{w}}{\text{maximize}} \quad & \mathbf{w}_1^{\text{T}}\mathbf{X}_1^{\text{T}}\mathbf{X}_2\mathbf{w}_2 \\ \text{s.t.} \quad & \mathbf{w}_1^{\text{T}}\mathbf{X}_1^{\text{T}}\mathbf{X}_1\mathbf{w}_1 = 1, \\ & \mathbf{w}_2^{\text{T}}\mathbf{X}_2^{\text{T}}\mathbf{X}_2\mathbf{w}_2 = 1 \end{aligned} \tag{1}$$

where N is the number of subjects and $d_1 = d_2 = d$ is the feature dimension. Kernel CCA (kCCA) [29] [30] transforms features into a reproducing kernel Hilbert space (RKHS), and finds the alignment between the transformed features $\mathbf{K}_1$ and $\mathbf{K}_2$. The similarity in the RKHS is $k(\mathbf{X}_{i,:}, \mathbf{X}_{j,:}) = \phi(\mathbf{X}_{i,:})^T\phi(\mathbf{X}_{j,:})$, where $\phi : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ is the feature transformation. LatSim learns a linear kernel $\mathbf{A} \in \mathbb{R}^{d \times d'}$; however, this still allows detection of nonlinear relationships.

The main idea behind CCA and kCCA is to maximize the similarity between two or more signals after some constrained transformation. This constrained transformation moves the data to a latent space, which may be of lower dimension. The limitation of CCA and kCCA is that they are unsupervised learning techniques that must account for every similarity between the signals, not just those relevant for a particular application, although recent work is tackling this problem [19].

B. Latent similarity

In contrast to unsupervised learning, LatSim maximizes similarity of subjects relative to a response variable of interest, such as age, sex or intelligence. First, similarities are computed as the inner product of the low-dimensional projections of subject features, based on a learned kernel:

[1] https://github.com/aorliche/LatentSimilarity/.

$$\begin{aligned}\text{sim}(a,b) &= \langle \phi(\mathbf{x}_a), \phi(\mathbf{x}_b) \rangle \\ &= \mathbf{x}_a \mathbf{A}\mathbf{A}^{\mathrm{T}} \mathbf{x}_b^{\mathrm{T}},\end{aligned} \tag{2}$$

where $\mathbf{A} \in \mathbb{R}^{d \times d'}$ is the kernel matrix and $\mathbf{x}_a, \mathbf{x}_b \in \mathbb{R}^d$ are feature vectors for subjects a and b, respectively. These similarities are then adjusted by passing them through a softmax activation function while masking each subject's self-similarity. The entire model for a single predictive task and a single fMRI paradigm is as follows:

$$\begin{aligned}\mathbf{M} &= \text{diag}(\infty), \\ \mathbf{E} &= S_{Row}((\mathbf{1} - \mathbf{M}) \odot \mathbf{X}\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}), \\ S(\mathbf{z})_i &= \frac{e^{z_i/\tau}}{\Sigma_{j=0}^{N} e^{z_j/\tau}},\end{aligned} \tag{3}$$

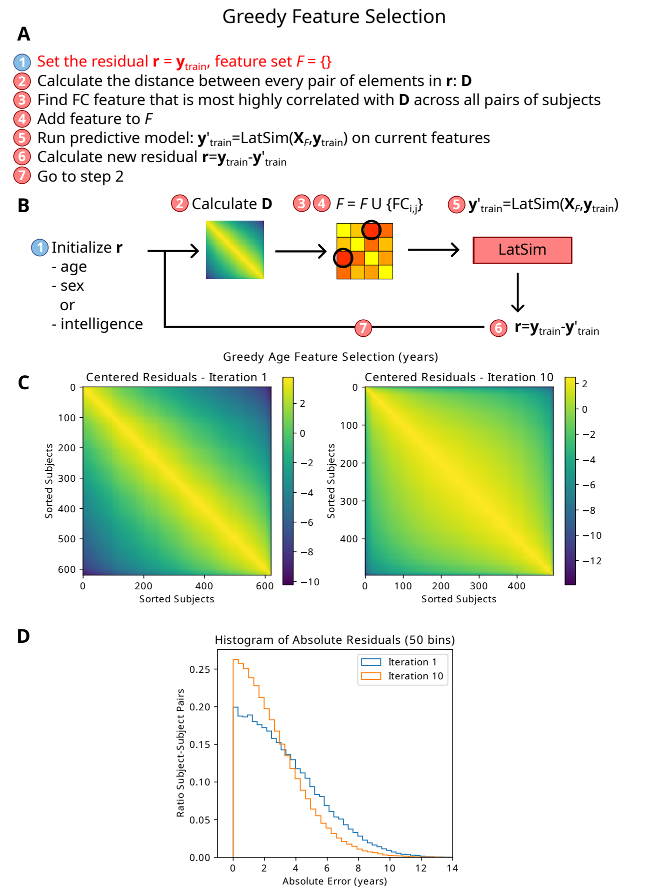
where $\mathbf{E} \in \mathbb{R}^{N \times N}$ is the final similarity matrix, $\mathbf{M} \in \mathbb{R}^{N \times N}$ is a mask to remove self-loops in predictions, $\infty \in \mathbb{R}^N$ is a vector of infinite-valued elements, $\mathbf{1} \in \mathbb{R}^{N \times N}$ is a matrix of ones, $\mathbf{X} \in \mathbb{R}^{N \times d}$ is the feature matrix, $\mathbf{A} \in \mathbb{R}^{d \times d'}$ is the kernel taking connectivity features to a lower latent dimension, N is the number of subjects, d is the number of features (FCs), $S(\mathbf{z})_i$ is the softmax function with temperature τ, and $S_{Row}(\mathbf{Z})$ is a function applying softmax to each row of the input matrix. High or low temperature τ determines whether the subject-subject similarity matrix $\mathbf{E}$ is more dense or sparse, respectively. The final similarity matrix of training and test set subjects is multiplied by the training set response variable to yield the prediction:

$$\hat{\mathbf{y}} = \mathbf{E}\mathbf{y}_{train} \tag{4}$$

The model is trained, using gradient descent, by minimizing the following objective function. Here we assume for brevity the existence of two fMRI feature matrices $\mathbf{X}_a$ and $\mathbf{X}_b$, and two predictive tasks, one regression (1) and one classification (2), for which we identify four kernel matrices $\mathbf{A}_{1a}$, $\mathbf{A}_{1b}$, $\mathbf{A}_{2a}$ and $\mathbf{A}_{2b}$:

$$\begin{aligned}\underset{\mathbf{A}_{1a}, \mathbf{A}_{1b}, \mathbf{A}_{2a}, \mathbf{A}_{2b}}{\text{minimize}} \quad & \\ & \frac{1}{N}(\mathbf{y}^{(1)} - \mathbf{E}^{(1a)}\mathbf{y}^{(1)})^2 + \\ & \frac{1}{N}(\mathbf{y}^{(1)} - \mathbf{E}^{(1b)}\mathbf{y}^{(1)})^2 + \\ & \gamma_1 \frac{1}{N} \Sigma_{n=1}^{N} \Sigma_{c=1}^{C} \mathbf{Y}^{(2)}_{:,c} \cdot \log(\mathbf{E}^{(2a)}\mathbf{Y}^{(2)})_{:,c} + \\ & \gamma_2 \frac{1}{N} \Sigma_{n=1}^{N} \Sigma_{c=1}^{C} \mathbf{Y}^{(2)}_{:,c} \cdot \log(\mathbf{E}^{(2b)}\mathbf{Y}^{(2)})_{:,c} + \\ & \lambda_1 ||\mathbf{A}_{1a}||_1 + \lambda_2 ||\mathbf{A}_{1b}||_1 + \\ & \lambda_3 ||\mathbf{A}_{2a}||_1 + \lambda_4 ||\mathbf{A}_{2b}||_1 + \\ & \alpha_1 ||\mathbf{A}_{1a} \odot \mathbf{A}_{2a}||_1 + \\ & \alpha_2 ||\mathbf{A}_{1b} \odot \mathbf{A}_{2b}||_1 + \\ & \beta_1 ||\mathbf{X}_a\mathbf{A}_{1a} - \mathbf{X}_b\mathbf{A}_{1b}||_2 + \\ & \beta_2 ||\mathbf{X}_a\mathbf{A}_{2a} - \mathbf{X}_b\mathbf{A}_{2b}||_2,\end{aligned} \tag{5}$$

where $\mathbf{E}^{(1a)} \in \mathbb{R}^{N \times N}$, for example, is the similarity matrix for task 1 and fMRI paradigm a, $\mathbf{y}^{(1)} \in \mathbb{R}^N$ (numeric) and $\mathbf{Y}^{(2)} \in \mathbb{R}^{N \times C}$ (one-hot categorical) are the stacked response variables for tasks 1 and 2, respectively, N is the number of subjects, C is the number of classes in task 2, γ_i is a task importance weight, λ_i is a sparsity-inducing hyperparameter, α_i is a hyperparameter promoting feature disentanglement, and β_i is a hyperparameter promoting alignment between fMRI paradigms. Note that our experiments on the PNC dataset in Section III-B.1 used precomputed vectorized functional connectivity matrices as the input, e.g., $\mathbf{X}$ is a matrix where each row is the vectorized FC of one subject.



1 Set the residual **r** = $\mathbf{y}_{train}$, feature set F = {}
2 Calculate the distance between every pair of elements in **r**: **D**
3 Find FC feature that is most highly correlated with **D** across all pairs of subjects
4 Add feature to F
5 Run predictive model: $\mathbf{y}'_{train}$=LatSim($\mathbf{X}_F$,$\mathbf{y}_{train}$) on current features
6 Calculate new residual **r**=$\mathbf{y}_{train}$-$\mathbf{y}'_{train}$
7 Go to step 2

Fig. 2. The greedy feature selection algorithm. **A.** A summary of the algorithm. **B.** A flowchart representation. **C.** Visualization of the residual distance matrices used to choose an FC feature at each iteration, at iterations 1 and 10. **D.** Histogram of absolute residual distance matrix values at iterations 1 and 10. Since LatSim works on subject pairs, our objective is to fit distances between residuals.

In the conventional image domain, *Zheng et al.* have proposed a similar metric learning approach using softmax aggregation for image classification [31]. However, their work makes use of a pre-trained backbone, is semi-supervised, and does not provide all of the possibilities for feature selection, disentanglement, and alignment as does LatSim (see Equation 5).

C. Greedy selection algorithm and model interpretability

A greedy selection algorithm was developed to compare with other interpretability methods [32]. The algorithm selects connections one at a time by ranking their ability to separate dissimilar subjects, i.e., their ability to minimize similarity between subjects that are "far apart" with regards to the current residual:

$$\begin{aligned} \mathbf{r}^{(i)} &= \text{LatSim}(\mathbf{X}_{F_{i-1}}, \mathbf{y}) - \mathbf{y}, \\ D_{ab} &= (r_a^{(i)} - r_b^{(i)})^2, \\ \mathbf{D} &= \mathbf{D} - \frac{1}{N^2}\Sigma_{ab} D_{ab}, \\ F_i &= F_{i-1} \cup \{\underset{j}{\text{argmin}}\ \Sigma_{ab}\ (D_{ab} X_{aj} X_{bj})\}, \end{aligned} \tag{6}$$

where LatSim : $\mathbb{R}^{N \times |F_{i-1}|} \to \mathbb{R}^N$ is the predictive model, $r_a^{(i)}$ is the residual at iteration i for subject a, $\mathbf{D} \in \mathbb{R}^{N \times N}$ is a centered matrix of differences between residuals, $F_i = \{0, \ldots, i\}$ is the set of selected connections at iteration i, $\mathbf{X} \in \mathbb{R}^{N \times d}$ is the vectorized FC matrix for all subjects, and $\mathbf{y} \in \mathbb{R}^N$ is the response variable. A summary of the algorithm is presented in Figure 2. We describe feature selection results in Section III-B.3.

The greedy algorithm can select the several dozen most relevant features given a single predictive task. To select discriminative features using the fully trained model, we find the correlation between subject similarities and residual distances, as in Equation 6 above, except the FCs are multiplied by the learned model weights:

$$F = \underset{j}{\text{argsort}}\ \Sigma_{abd}\ (D_{ab} A_{dj}^2 X_{aj} X_{bj}), \tag{7}$$

where the residual is set to the response variable, $\mathbf{D}$ is calculated as before, $\mathbf{A} \in \mathbb{R}^{d \times d'}$ is the set of model weights, and F is the resulting set of ranked features.

Except for greedy feature selection, we optimized prediction of all three response variables (age, sex, and intelligence) at the same time in the same LatSim model. Greedy selection required optimizing a single task at once, as the best feature for age prediction may not be the best feature for sex or intelligence prediction. LatSim was trained using PyTorch on an NVIDIA Titan Xp with CUDA support.

D. Spurious correlation

We hypothesize that overfitting occurs due to feature noise or confounds, such as scanner motion, whose effects are more severe for smaller size cohorts. These confounds may create spurious correlations in a subset of the cohort.

We define a spuriously correlated feature X to be one that appears to be highly correlated with response variable Y for only a subset of subjects:

$$|\rho_S| \begin{cases} > 0 & s \in S \\ \approx 0 & s \in C \setminus S \end{cases} \tag{8}$$

where ρ_S is the value of the spurious correlation, C is the study cohort, and $S \subseteq C$ is a subset of the cohort such that $C \setminus S$ is maximized.

Note that spurious correlation may actually be true correlation identifying subgroups, but we hypothesize that a spurious correlation is more likely to be false as $|S|$ decreases. We conduct simulation experiments in Section III-A that suggest LatSim is more robust against spurious correlation than traditional feature-based models. When $|S|$ is close to $|C|$, and the effect is systematic, we cannot tell whether the correlation is true or false.

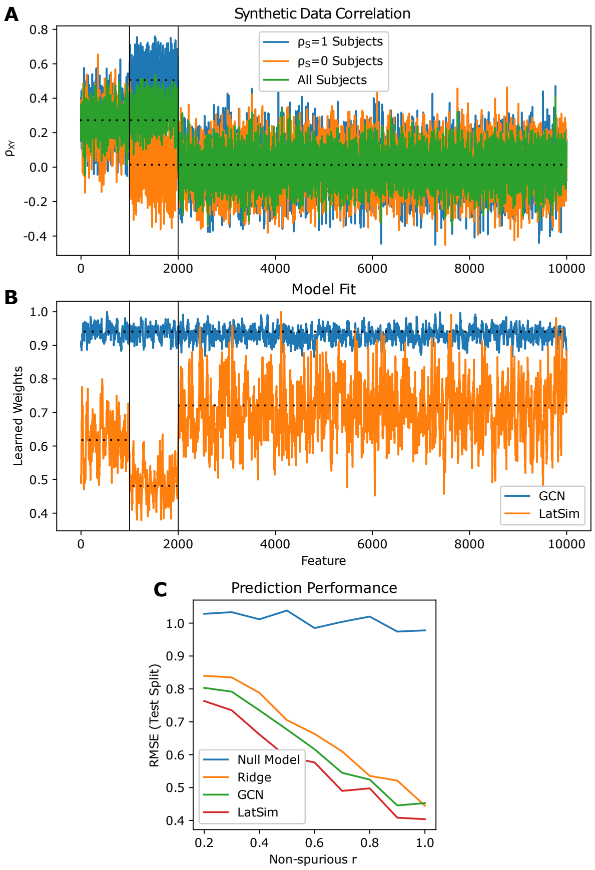


Fig. 3. Results of simulations on synthetic data with spurious correlation. **A.** Data generated with non-spurious $\rho = 1/(2\sqrt{4.25}) \approx 0.25$, present in all subjects, and spurious $\rho_S \approx 1/\sqrt{5} \approx 0.5$, present in half of subjects. Correlation of response variable with feature for the training set is shown. Only the first two thousand features have any information relevant for prediction. **B.** Absolute value of learned model weights for the GCN and LatSim models, averaged over the first hidden layer (GCN) or latent dimension (LatSim). Weights are smoothed by a convolution kernel of size 20 to aid visualization. **C.** Average predictive performance (RMSE between ground truth $\boldsymbol{y_i}$ and predicted $\boldsymbol{\hat{y}_i}$) over 6 independent train/validation/test splits, evaluated on the test split.

III. Results

We first demonstrate the superior performance of LatSim in a simulation study, then apply it to brain development fMRI data consisting of children and adolescents. We use both full-model and greedy feature selection to identify important functional connections for age, sex, and intelligence prediction.

A. Simulation experiment

We performed a simulation experiment to test LatSim in the presence of a ground truth dataset. A set of $N_{train} = 40$, $N_{val} = 120$, and $N_{test} = 120$ subjects with 10,000 normally-distributed features x_{ni} was generated, where n and i refer to subject and feature, respectively. Each subject was also associated with a response variable y_n. The data generation process for each subject was as follows:

$$\begin{aligned} y_n &\sim \mathcal{N}(0,1), \qquad z_{ni} \sim \mathcal{N}(0,4) \\ x_{ni} &= \begin{cases} z_{ni} + y_n r, & \text{if } i \leq 1000 \\ z_{ni} + y_n r_S, & \text{if } 1000 < i \leq 2000 \text{ and } n \text{ even} \\ z_{ni}, & \text{otherwise} \end{cases} \end{aligned} \tag{9}$$

where r and r_S are correlation-generating parameters for non-spurious and spurious correlations, respectively. In other words, the first 1,000 features were correlated with the response variable at level ρ, the next 1,000 features of half of the subjects were correlated at level ρ_S (and had 0 correlation for the other half of subjects), and the remaining 8,000 features were left uncorrelated. We varied r from 0.2 to 1 while keeping $r_S = 1$. It can be seen that final feature to response variable correlation is $\rho = r/\sqrt{r^2+4}$ for correlated features for all subjects, and $\rho_S \approx r_S/\sqrt{r_S^2+4}$ for spuriously correlated features for half of subjects.

The simulation showed that LatSim performs better than both a GCN [33] and Ridge Regression model in the presence of the spurious correlation ρ_S (see Figure 3). Additionally, LatSim was the only model identifying the three types of features: correlated, spuriously correlated, and uncorrelated. All results are on the test split. We believe insensitivity to spurious correlation is one of the reasons that LatSim performs well in the low-sample, high-dimensionality regime (see Section IV-C). A multi-layer perceptron (MLP) with L1-regularization performed as well as Ridge Regression (not shown). The GCN model was not interpretable via either weight magnitude or gradient-based saliency. The MLP model identified only sparse features and selected features in the non-informative range. In contrast, LatSim was able to consistently identify the full range of informative features.

Notably, the weights are smaller for correlated features than for non-correlated features. This is an artifact of taking the absolute value of weights in order to average them across latent dimensions. Conversely, the spuriously-correlated weights are, on average, smaller than the constantly-correlated weights. To explain, suppose there are 2 sets of features, A and B, which are correlated and non-correlated, respectively. The similarity between two subjects will be:

$$\begin{aligned} &\mathbb{E}[(w_A A_1 + w_B B_1)(w_A A_2 + w_B B_2)] \\ &\qquad = w_A^2 \mathbb{E}[A_1 A_2] + w_B^2 \mathbb{E}[B_1 B_2] \\ &\qquad\quad + w_A w_B \mathbb{E}[A_1 B_2] + w_A w_B \mathbb{E}[A_2 B_1] \\ &\qquad = w_A^2 \mathbb{E}[A_1 A_2] > 0, \end{aligned} \tag{10}$$

hence it doesn't matter what magnitude the weights B have, because the expectation of the non-$A_1 A_2$ terms is zero due to independence and the standard normal distribution of features. Conversely, if there is a subset of features A that are spuriously-correlated, it is beneficial to reduce the spurious weights compared to the non-spurious ones.

B. Brain development study

TABLE II

DEMOGRAPHIC INFORMATION FOR THE SUBSET OF THE PNC DATASET USED IN OUR EXPERIMENTS. WRAT SCORE HAS BEEN ADJUSTED FROM ITS RAW VALUE BY REGRESSING OUT THE EFFECTS OF AGE.

	Number of Subjects
Males	286
Females	334
Total	620

	Min	Mean	Max
Age (months)	103	180±39	271
Age (years)	8.6	15±3.3	22.6
WRAT score	70	102±15.7	145

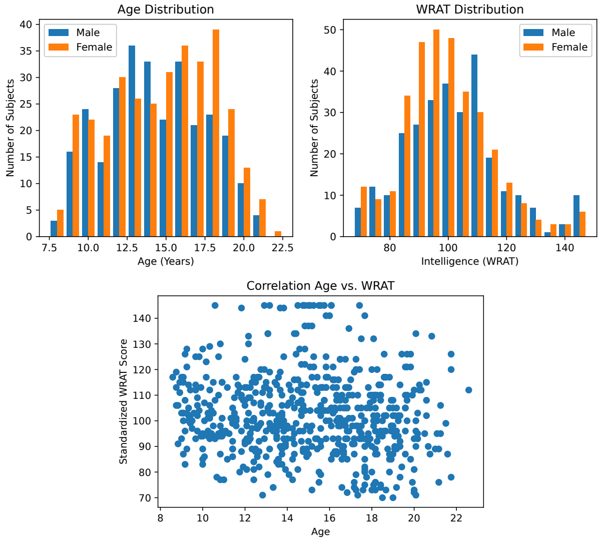


Fig. 4. Demographics of the 620-subject subset of the PNC study used in our experiments. WRAT score has been adjusted from its raw value by regressing out the effects of age.

1) Dataset: We trained and validated our model on the publicly available PNC dataset. The PNC dataset contains multi-paradigm fMRI data, neurocognitive assessments, and psychiatric evaluations for 1,445 healthy adolescents ages 8-23. We chose 620 subjects from the cohort who had both working memory paradigm (nback) and emotion identification paradigm (emoid) fMRI scans, along with results from the 1-hour Wide Range Achievement Test (WRAT) [34] to measure general intelligence.

fMRI was performed using a 3T Siemens TIM Trio whole-body scanner with a single-shot, interleaved multi-slice, gradient-echo, echo-planar imaging sequence. The resolution was set to be 3x3x3 mm with 46 slices. The imaging parameters were TR = 3000 ms, TE = 32 ms, and flip angle = 90 degrees. Gradient magnitude was 45 mT/m, having a maximum slew rate of 200 T/m/s. The duration of the nback scan was 11.6 minutes (231 TR), during which time subjects were asked to conduct the n-back memory task, which is related to working memory and lexical processing [35]. The duration of the emoid scan was 10.5 minutes (210 TR), during which time subjects viewed faces displaying different emotions and gave an indication of what emotion was displayed. The demographics of our study cohort are given in Table II and the distribution is visualized in Figure 4.

Data was pre-processed with SPM12[2]. This included using multiple regression for motion correction, as well as spatial normalization and smoothing by a 3mm Gaussian kernel [36]. Pre-processing was similar to [37]. The Power template [38] was used to parcellate BOLD signal among 264 regions of interest, from which a 264×264 symmetric connectivity matrix was constructed using Pearson correlation. The unique $d = 34,716$ entries in the upper right triangle, excluding the main diagonal, were vectorized and taken as the FC features for each subject.

The goal of the experiment was to predict subject age, sex, and intelligence as measured by WRAT score. Prediction performance was measured by root mean squared error (RMSE) for age and intelligence prediction, and accuracy for sex prediction, respectively. LatSim was compared against simple linear models (Least Squares and Logistic Regression), a Graph Convolutional Network (GCN), a Multi-Layer Perceptron (MLP), and a Multimodal Graph Convolutional Network (M-GCN). M-GCN is a recent deep learning model for functional connectome analysis [39] based on the CNN [40] architecture.

The inputs to all models were nback, emoid, and the arithmetic sum of nback and emoid task based vectorized FC matrices, from which separate predictions were made and averaged as part of an ensemble. The sum of nback and emoid FC was used to increase ensemble size. Standardization (Z-score normalization) was performed on the vectorized FC matrices using statistics from the training dataset applied to training, validation, and test datasets. Z-score normalization was performed only for the LatSim model, since the other models sometimes did not converge for Z-score normalized data. All predictive and feature selection experiments were carried out using 10-fold cross validation (CV), with an 80% training, 10% validation, and 10% test split. Hyperparameters were selected using random grid search (see Table III for LatSim hyperparameters and Table IV for those of other models). The search grid was initialized to be a 5-decade window around prior assumptions of optimal hyperparameters, with search points occurring at decade intervals for all models. A total of 100 grid points were evaluated with three repetitions. The only exceptions were dropout, which was sampled at 0.1 intervals, latent/hidden dimension, which was set heuristically, and number of training epochs, which was set to just past the maximum best validation epoch for each model individually. Hyperparameters were estimated for the largest training set size ($N = 496$) and subsequently used for all training set sizes, with the belief that over-optimization would give a distorted view of models and reduce reproducibility.

2) Prediction: LatSim achieved superior predictive performance on the PNC dataset in all three predictive tasks, especially at low sample sizes. The result of the entire experiment is given in Figure 5, and the low and high sample size results are given in Table V.

At N=30, close to the previously reported threshold of N=36 for modestly reproducible fMRI results, we see that LatSim is the only model not to overfit. It surpassed the other models by a significant margin in two of three predictive tasks. Interestingly, LatSim performed much better at small sample sizes than the simple linear models, which we attribute to use of $\mathcal{O}(n^2)$ inter-subject connections rather than the n subjects themselves. LatSim remains the best performing model until about N=100, at which point it is only slightly better than the other best predictive model, GCN. We note that the GCN model performs almost as well as LatSim, except at low sample sizes. We also note that with a categorical response variable such as sex, the performance of both LatSim and GCN is reduced. We believe the advantage of both LatSim and the GCN model lies in utilizing inter-subject similarities and differences. This is hindered by a lack of granularity in the response variable.

TABLE III
HYPERPARAMETERS FOR PNC EXPERIMENTS (LATSIM).

Predictive Tasks	Age, Sex, Intelligence
fMRI Paradigms	nback, emoid, nback+emoid
Classification Multiplier	$\gamma = 1000$
Sparsity Parameter	$\lambda = 10$
Disentanglement Parameter	$\alpha = 100$
Feature Alignment Parameter	$\beta = 0.1$
Original Dimension	$d = 34,716$
Latent Dimension	$d' = 2$
Temperature	$\tau = 1$
Feature Dropout Rate	0.5
Edge Matrix Dropout Rate	0.1
Number of Training Epochs	200
Optimizer	Adam
Learning Rate	1e-4
L2 Regularization Parameter	1e-4
Weight Initialization	1e-4$\cdot\mathcal{N}(0,1)$

TABLE IV
HYPERPARAMETERS FOR PNC EXPERIMENTS (COMPARISON MODELS).

Predictive Tasks	Age, Sex, Intelligence
fMRI Paradigms	nback, emoid, nback+emoid
Model	*Least Squares Regression*
Implementation	PyTorch[3]
Model	*Logistic Regression*
Implementation	scikit-learn[4]
Regularization Parameter	C=1
Model	*All Deep Models*
Optimizer	Adam
Weight Initialization	PyTorch Default
Learning Rate	1e-4
Model	*MLP*
Layers	34,716 x 100 (hidden)
Number of Training Epochs	10,000
L1 Regularization Parameter	1e-2
L2 Regularization Parameter	1e-3
Model	*M-GCN*
Implementation	Github Repository[5]
Number of Training Epochs	5,000
L2 Regularization Parameter	1e-4
Model	*GCN*
Layers	34,716 x 100 (hidden)
Number of Training Epochs	10,000
Graph Type	Complete
Neighbor Weight (Total)	0.5
Node Self-loop Weight	0.5
L2 Regularization Parameter	1e-4

Based on the prediction results, LatSim can fit a dataset

[2] http://www.fil.ion.ucl.ac.UK/spm/software/spm12/

[3] https://pytorch.org/

[4] https://scikit-learn.org/stable/

[5] https://github.com/Niharika-SD/M-GCN

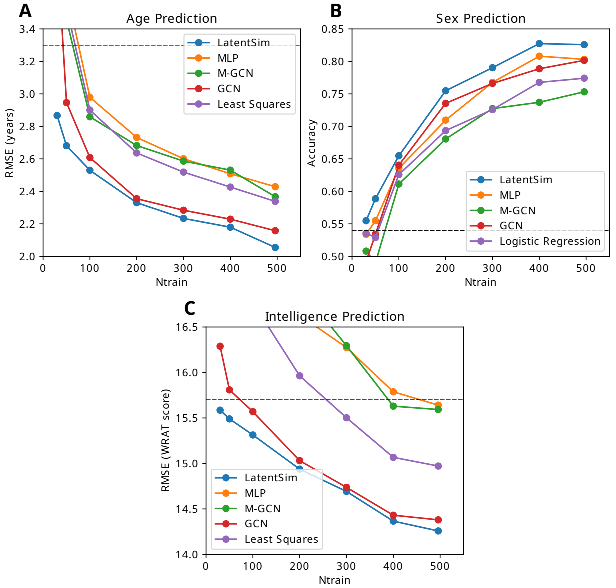


Fig. 5. Results of age (**A**), sex (**B**), and intelligence (**C**) prediction experiments on our subset of the PNC dataset. Dashed black lines represent the null model. All models except LatSim performed worse than chance at the N=30 training set size for all tasks.

TABLE V
RESULTS OF PNC EXPERIMENTS.

	Age		Sex		Intelligence	
	(RMSE, years)		(Accuracy)		(RMSE, WRAT score)	
Model	N=30	N=496	N=30	N=496	N=30	N=496
Null	3.3		0.54		15.7	
M-GCN	4.47	2.37	0.51	0.75	23.27	15.59
MLP	4.52	2.43	0.53	0.8	21.17	15.64
GCN	3.89	2.16	0.49	0.8	16.29	14.38
Linear	4.36	2.34	0.54	0.77	19.8	14.97
LatSim	**2.86**	**2.05**	**0.55**	**0.82**	**15.59**	**14.26**
p-value	**2.2e-6**	**5e-3**	0.32	0.11	**0.02**	0.30

in orders of magnitude less time compared to other models (see Table VI). This makes it possible to perform large-scale bootstrapping, mixture of experts, and ensembling that is not possible with traditional ML models. It also allows for the use of greedy selection.

3) Significant FCs in prediction: The most important FCs for all prediction tasks are given in Table VII. All connections are given with Automated Anatomical Labeling (AAL) region names [41] and with Montreal Neurological Institute (MNI) region coordinates. For age prediction, the most important connections were Insula_R to Putamen_R and Temporal_Inf_R to Frontal_Med_Orb_R, being present in the top 10 connections for both the nback and emoid paradigms. For sex prediction, the Precentral_L to Temporal_Pole_Mid_R FC was found in the top 10 connections for the emoid paradigm. For intelligence prediction, the Postcentral_L to Postcentral_R FC was found

TABLE VI
TRAINING TIME FOR ALL 10 FOLDS OF 10-FOLD CROSS VALIDATION.

Model	LatSim	Lstsq	Logistic	GCN	MLP	M-GCN
Epochs	200	-	100	1e4	1e4	5e3
Training Time	**4.3s**	**<1s**	63.4s	406s	364s	5912s

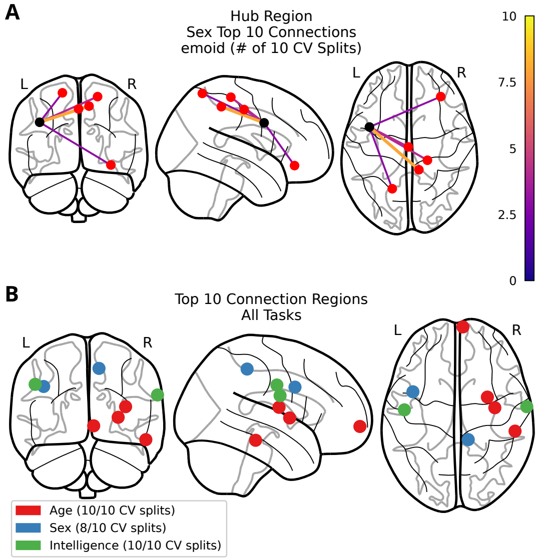


Fig. 6. **A.** Identification of an interesting ”hub” region found by emoid paradigm sex prediction that was included in 5 separate connections from among the top 10 connections across all CV splits. **B.** Visualization of regions found in the top 10 connections of more than 8 CV splits using the greedy selection algorithm.

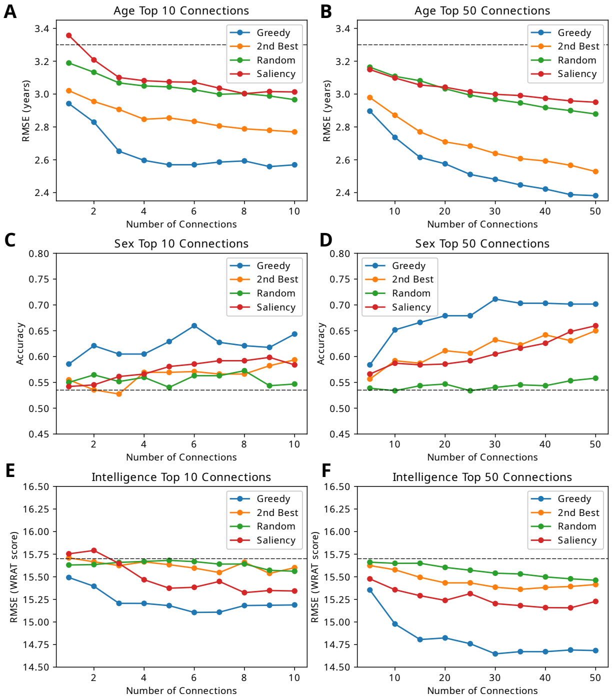


Fig. 7. Comparison of four connection selection strategies. Dashed black lines represent the null model. Selection up to 10 connections (**A**, **C**, **E**) was done without dropout, whereas selection up to 50 connections (**B**, **D**, **F**) was done with 0.5 dropout.

TABLE VII

MOST IMPORTANT CONNECTIONS FOR DISCRIMINATING AGE, SEX, AND INTELLIGENCE AMONG HEALTHY ADOLESCENTS. THE # CV SPLITS COLUMN SHOWS THE NUMBER OF CV SPLITS FOR WHICH THE CONNECTION APPEARED IN THE TOP 10 CONNECTIONS OF THE GREEDY SELECTION ALGORITHM.

Region 1	MNI Coords	Network	Region 2	MNI Coords	Network	# CV Splits	Paradigm	Prediction Task
Insula_R	(36,-9,14)	SMT	Putamen_R	(29,1,4)	SUB	10/10	Both	Age
Temporal_Inf_R	(55,-31,-17)	UNK	Frontal_Med_Orb_R	(6,67,-4)	DMN	10/10	Both	Age
Frontal_Mid_L	(-34,55,4)	FRNT	Frontal_Mid_Orb_L	(-42,45,-2)	FRNT	9/10	nback	Age
Thalamus_R	(6,-24,0)	SUB	Left Brainstem	(-5,-28,-4)	SUB	9/10	emoid	Age
Precentral_L	(-41,6,33)	FRNT	Temporal_Pole_Mid_R	(11,-39,50)	SAL	8/10	emoid	Sex
Insula_R	(27,16,17)	UNK	Frontal_Inf_Orb_R	(49,35,-12)	DMN	3/10	emoid	Sex
Temporal_Pole_Mid_R	(46,16,-30)	DMN	Temporal_Pole_Mid_R	(52,7,-30)	DMN	3/10	nback	Sex
Frontal_Sup_Orb_R	(24,32,-18)	UNK	Fusiform_R	(27,-37,-13)	DMN	3/10	nback	Sex
Postcentral_L	(-49,-11,35)	SMT	Postcentral_R	(66,-8,25)	SMT	10/10	Both	Intelligence
Temporal_Mid_R	(52,-2,-16)	DMN	Precuneus_R	(10,-62,61)	DRSL	6/10	nback	Intelligence
Cerebelum_6_L	(-16,-65,-20)	CB	Postcentral_R	(66,-8,25)	SMT	5/10	emoid	Intelligence
Precentral_R	(44,-8,57)	SMT	Temporal_Inf_L	(-42,-60,-9)	DRSL	5/10	emoid	Intelligence

SMT=Sensory/Somatomotor, CNG=Cingulo-opercular Task Control, AUD=Auditory, DMN=Default Mode, MEM=Memory Retrieval, VIS=Visual, FRNT=Fronto-parietal Task Control, SAL=Salience, SUB=Subcortical, VTRL=Ventral Attention, DRSL=Dorsal Attention, CB=Cerebellum, UNK=Uncertain

in the top 10 connections for both the nback and emoid paradigms. In addition, for sex prediction, we identified the Left Inferior Frontal Gyrus (Precentral_L) as a region making multiple top 10 connections, as shown in Figure 6.

Using only the first few connections gives half of the predictive power of using the full set of $d = 34,716$ connections. In particular, Figure 7 shows that the first 3 connections, if properly chosen, can contain more information than the next 50 connections, chosen in the same manner. Specifically, 10 FCs can explain 21% of variance for age, 50 FCs can explain 27%, whereas with the full set of FCs the GCN model can explain 35% and LatSim can explain 38%. The selected connections were chosen using the greedy feature selection algorithm. Figure 7 shows that the FCs chosen by greedy selection are superior to those chosen by gradient-based saliency, as well as to random FCs. Additionally, we compared the FCs chosen by greedy selection to the next-best FCs that would be chosen by it. We believe this helps validate the significance of our identified connections, since, for small numbers of connections, we could not find a minimal combination of FCs that performed as well as that found by greedy selection.

Selecting connections with the fully trained LatSim model corroborated the trend found by greedy selection. As seen in Figure 8, we identified a very few "core" connections that were disproportionately important to the prediction task. The rest of the connections were interchangeable in terms of discriminative ability. Note, for instance, the rapid increase in accuracy for the 3 best FCs and the subsequent plateau in Figure 7. Likewise, almost all of the connections found in the top 50 connections by greedy selection were also found in the top 50 connections of the full model.

IV. DISCUSSION

A. Significant functional networks

The top connections identified by this study contain regions that fall into the default mode (DMN), subcortical (SUB), fronto-parietal task control (FRNT), and sensory/somatomotor (SMT) brain functional networks (FNs). Abbreviations are given as a footnote to Table VII. Regions that belong to the same FN (within-module) tend to be more synchronized than regions from different FNs (between-module) [42]. In Figure 8C, blocks on the main diagonal of the FC matrices represent connections within-module, while blocks off the main diagonal represent connections between-module. Recently, *Jiang et al.* found that, in an older population, connections between the DMN, SMT, and SUB networks were highly predictive for age [43]. They also found that a DMN-SUB connection was correlated with high cognitive performance.

The DMN was overrepresented in the top 10 connections for all predictive tasks; 36% of regions identified were part of the DMN, whereas DMN regions constitute 22% of the Power atlas. Robust developmental changes have been identified in the DMN, and DMN connectivity has been positively correlated with high cognitive performance [44]. *Fan et al.* found that DMN connectivity increases from childhood until young adulthood [45]. *Pan et al.* identified FCs which included DMN regions to be more important in predicting intelligence than FCs which didn't [46].

The SMT network was overrepresented in top 10 connection regions for intelligence prediction. In that task, 43% of top 10 connection regions belonged to the SMT network, whereas SMT regions constitute 13% of the Power atlas. It is known that dysfunction in the SMT network is correlated with depression [47]. However, FC represents synchronization between brain regions, and the cause of altered FC may not lie in the region itself. Table VII shows that the top SMT connections involve the CB network, leading to the idea that complex motor control is related to intelligence.

Many of the most important connections we identified for each predictive task are not recognized as part of an FN, and are classified as unknown-network (UNK). 24% of regions identified in top 10 connections are labeled UNK, whereas UNK regions constitute 10% of ROIs in the Power atlas. These connections include cerebellar regions; some cerebellar regions are not included in the CB network because they contribute to functions other than motor function [48], including social thinking and emotion [48]. *Zhang et al.* recently found disrupted effective connectivity in UNK cerebellar regions in

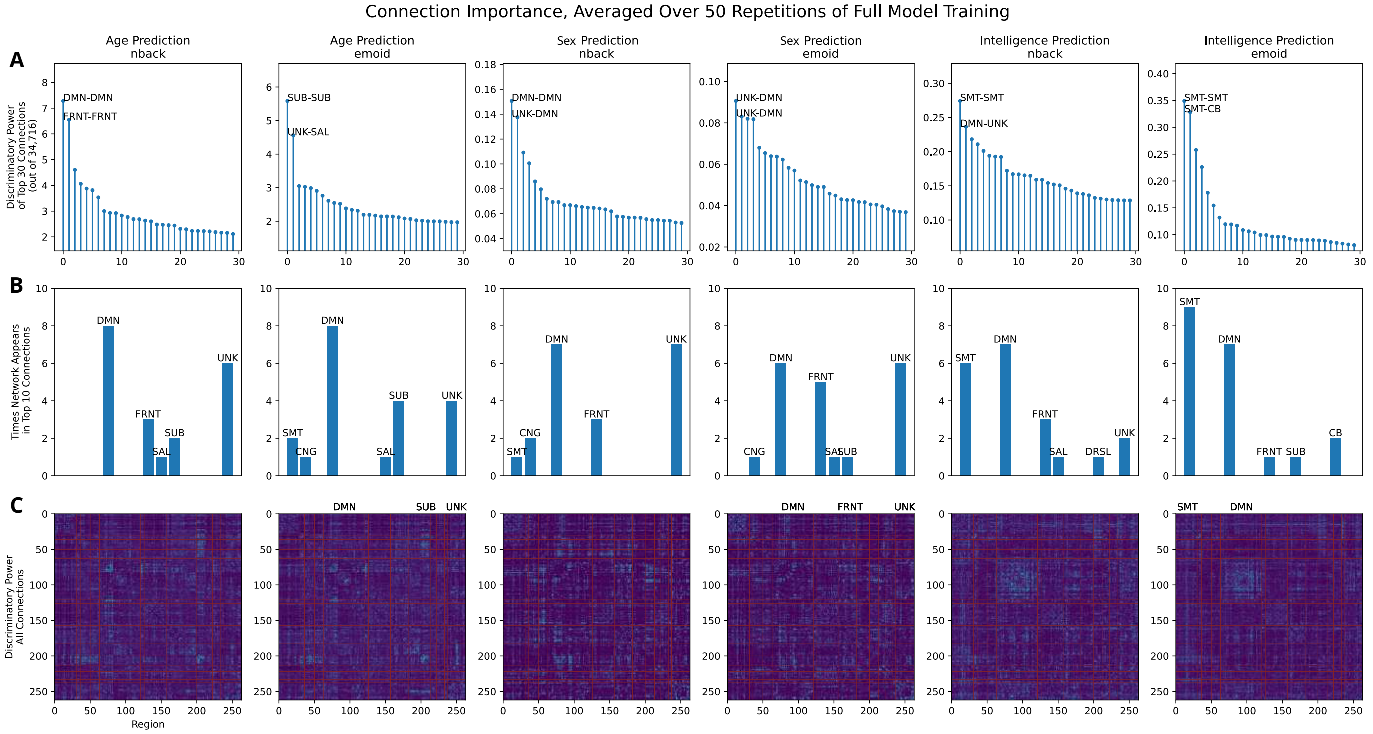


Fig. 8. Important connections identified by running the full model with the entire $d = 34,716$ set of connections as inputs. As with greedy selection, we show that the first several connections are far more important than the remaining ones (**A**). Notably, the DMN is highly represented in the top 10 connections for all predictive tasks and modalities (**B**). The DMN as a whole seems to be important for intelligence prediction (**C**). Importance was averaged over 50 repetitions of an 80-10-10 train/validation/test split. Discriminative power was calculated as in Equation 7. Correlation was greater than zero for all connections. See Table VII for definitions of abbreviations.

individuals with schizophrenia, relative to controls [49].

B. Significant FCs

Greedy selection identified 4 FCs present in more than 8 out of 10 CV splits for one of the predictive tasks:

- **Insula_R to Putamen_R (Age)**. The Insula_R has many functions in humans dealing with low-level sensation, emotion, and high-level cognition [50]. *Mazzola et al.* hypothesized that the Insula_R participates in the social brain and found increased activation when participants watched scenes of joyful or angry actors [51]. Increased Putamen_R volume has been linked to autism spectrum disorder [52], and reduced amygdala-Putamen_R FC has been linked to ADHD [53].
- **Temporal_Inf_R to Frontal_Med_Orb_R (Age)**. The Temporal_Inf_R region is associated with language processing [54]. Temporal_Inf_R FC was found to be decreased in adolescent schizophrenia patients [55]. The Frontal_Med_Orb_R region is part of the prefrontal cortex and is associated with dysfunctional connectivity in major depressive disorder [56].
- **Precentral_L to Temporal_Pole_Mid_R (Gender)**. The Precentral_L region is associated with reading and language processing [57]. *Delvecchio et al.* found morphological differences in this region between sexes [58]. The Temporal_Pole_Mid_R region is linked to social contracts, precautions, and strategies [59].
- **Postcentral_L to Postcentral_R (Intelligence)**. This is a connection between regions symmetric about the body mid-line. The postcentral gyrus is involved in proprioception and contains the primary somatosensory cortex. Lesions in these regions may cause speech dysfunction [60] [61]. *Sander et al.* found inter-hemispheric connectivity to play a role in the ability to learn new languages [62] .

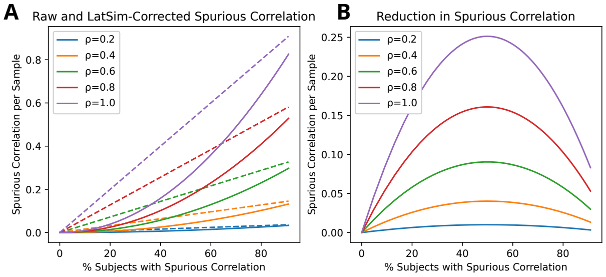


Fig. 9. **A.** Spurious correlation per sample in a traditional ML model (dashed lines) versus LatSim (solid lines). **B.** The absolute reduction in spurious correlation as a function of frequency in the sample.

Notably, AAL regions extend over a large area, and Power atlas ROIs do not correspond exactly to AAL regions.

C. Robustness to spurious correlation

In this section, we argue that LatSim is robust to spurious correlation because it identifies features based on $\mathcal{O}(n^2)$ inter-subject connections, rather than the number of subjects in the cohort.

Assume feature X is spuriously correlated with response variable Y on a subset of the cohort $S \subseteq C$, $s = |S|$, $n = |C|$, and $X, Y \sim \mathcal{N}(0, 1)$. That is, for each subject u:

$$|\rho_S| \begin{cases} > 0 & u \in S \\ \approx 0 & u \in C \setminus S \end{cases} \tag{11}$$

LatSim uses weighted inner product similarity wX_1wX_2 between the features of two subjects as input, where w is a learned weight. The correlation between wX_1wX_2 and $D = (Y_1 - Y_2)^2$ determines how well this feature pair predicts the response variable:

$$\begin{aligned}
\rho_{XX,D} &= \frac{\sigma^2_{XX,D}}{\sqrt{\sigma^2_{XX}\sigma^2_D}} \\
&= \frac{\text{Cov}[wX_1wX_2, (Y_1 - Y_2)^2]}{\sqrt{\text{Var}[wX_1wX_2]\text{Var}[(Y_1 - Y_2)^2]}} \\
&= \frac{\mathbb{E}[X_1X_2(Y_1 - Y_2)^2 - \mu_{XX}\mu_D]}{\sqrt{\mathbb{E}[X_1^2X_2^2 - \mu^2_{XX}]\mathbb{E}[(Y_1 - Y_2)^4 - \mu^2_D]}} \\
&= \frac{\mathbb{E}[X_1X_2(Y_1 - Y_2)^2 - 0 \cdot 2]}{\sqrt{\mathbb{E}[X_1^2X_2^2 - 0^2]\mathbb{E}[(Y_1 - Y_2)^4 - 2^2]}} \\
&= \frac{\mathbb{E}[X_1X_2Y_1^2 - 2X_1X_2Y_1Y_2 + X_1X_2Y_2^2]}{\sqrt{(1 \cdot 1)(12 - 4)}} \\
&= \frac{0 - 2\rho^2_{XY} + 0}{\sqrt{8}} \\
&= \begin{cases} -\frac{\rho_S^2}{\sqrt{2}} & 1, 2 \in S \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{12}$$

Since expectation is a linear operator, we can find the average value over the entire cohort:

$$\begin{aligned}
\rho_{XX,D} &= -\frac{s(s-1)}{n(n-1)}\frac{\rho_S^2}{\sqrt{2}} \approx -\frac{s^2}{n^2}\frac{\rho_S^2}{\sqrt{2}} \\
&\approx k\frac{s^2}{n^2}\rho_S^2
\end{aligned} \tag{13}$$

Conversely, in a traditional model, feature X is correlated with response variable Y as the size of the subset S:

$$\rho_{X,Y} = \frac{s}{n}\rho_S \tag{14}$$

A plot of the functions in Equations 13 and 14 is given in Figure 9. The maximum reduction in spurious correlation occurs at $s/n = 0.5$ and is about 1/4 of the value of the spurious correlation. The relative reduction is linear and maximal when $s = 0$, i.e., there are no subjects with spurious correlation (not shown). As s/n increases, the reduction in spurious correlation is diminished. This suggests that large model capacity is not the only reason complicated models falter at low sample sizes. We see in our experiments, e.g., in Table V, that the linear models perform worse than both LatSim and some other deep learning models.

Like LatSim, a k-layer GNN model also works on interactions between subjects, but as an adjunct to the prediction from the node self-loop. It also requires either additional degrees of freedom to estimate edge weights, or an arbitrary choice of a distance function and/or threshold. We believe the reason that a GCN model did so well in our experiments is that we made it incredibly simple: only 2 layers were used, and edge weights were uniform and equal in sum to the self-weights. It was found that expanding the GCN to 3 or 4 layers hurt performance. We believe the performance benefit comes from having a good prior and feature selection, not additional model capacity. Due to the very weak relationships between features and response variables in our data, we believe the advantage of the GCN was in averaging. This strategy breaks down at low sample sizes, where spurious feature correlation still causes large errors to be present at the node self-loop.

V. Conclusion

This paper proposes a novel model, LatSim, in the vein of metric learning, that is robust against overfitting at small sample sizes. It is interpretable, computationally efficient, multi-task and multi-view capable, and able to enforce feature disentanglement. First, we showed that LatSim is superior in the small sample size, high dimensionality regime, through both simulation and experiments on real datasets. Second, we identified specific connections within and between the sensory/somatomotor, default mode, fronto-parietal task control, and subcortical networks that are highly discriminative for age, sex, and intelligence in healthy adolescents. Third, we quantified the number of features required to attain a given prediction accuracy. Fourth, we showed that there are several core connections that are more discriminative for each predictive task than other connections. Finally, we found that connections identified by greedy selection were superior compared to those found by saliency methods. Our model may spur new research into algorithm development and, in turn, lead to new insights into the mechanisms underlying human cognition.

References

[1] J. W. Belliveau *et al.*, "Functional mapping of the human visual cortex by magnetic resonance imaging." *Science*, vol. 254 5032, pp. 716–9, 1991.

[2] A. Orlichenko, G. Qu, and Y.-P. Wang, "Phenotype guided interpretable graph convolutional network analysis of fMRI data reveals changing brain connectivity during adolescence," in *Medical Imaging 2022: Biomedical Applications in Molecular, Structural, and Functional Imaging*, B. S. Gimi and A. Krol, Eds., vol. 12036, International Society for Optics and Photonics. SPIE, 2022, pp. 294 – 303. [Online]. Available: https://doi.org/10.1117/12.2613172

[3] S. İçer, İrem Acer, and A. Baş, "Gender-based functional connectivity differences in brain networks in childhood," *Computer Methods and Programs in Biomedicine*, vol. 192, p. 105444, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260719310685

[4] G. Qu *et al.*, "Ensemble manifold regularized multi-modal graph convolutional network for cognitive ability prediction," *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 12, pp. 3564–3573, 2021.

[5] Y. Du, Z. Fu, and V. D. Calhoun, "Classification and prediction of brain disorders using functional connectivity: Promising but challenging," *Frontiers in Neuroscience*, vol. 12, 2018.

[6] P. R. Millar *et al.*, "Predicting brain age from functional connectivity in symptomatic and preclinical alzheimer disease," *Neuroimage*, vol. 256, no. 119228, p. 119228, Aug. 2022.

[7] V. Berisha *et al.*, "Digital medicine and the curse of dimensionality," *NPJ Digit. Med.*, vol. 4, no. 1, p. 153, Oct. 2021.

[8] B. O. Turner *et al.*, "Small sample sizes reduce the replicability of task-based fMRI studies," *Commun. Biol.*, vol. 1, no. 1, p. 62, Jun. 2018.

[9] H. Salehinejad *et al.*, “A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography,” *Scientific Reports*, vol. 11, 2021.

[10] D. Szucs and J. P. Ioannidis, “Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990–2012) and of latest practices (2017–2018) in high-impact journals,” *NeuroImage*, vol. 221, p. 117164, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811920306509

[11] A. S. Lundervold and A. Lundervold, “An overview of deep learning in medical imaging focusing on mri,” *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019, special Issue: Deep Learning in Medical Physics. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0939388918301181

[12] Z. Salahuddin *et al.*, “Transparency of deep neural networks for medical image analysis: A review of interpretability methods,” *Computers in Biology and Medicine*, vol. 140, p. 105111, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521009057

[13] G. Varoquaux and V. Cheplygina, “Machine learning for medical imaging: methodological failures and recommendations for the future,” *NPJ Digit. Med.*, vol. 5, no. 1, p. 48, Apr. 2022.

[14] X. Zhang *et al.*, “Gender differences are encoded differently in the structure and function of the human brain revealed by multimodal mri,” *Frontiers in Human Neuroscience*, vol. 14, 2020.

[15] N. Z. Bielczyk *et al.*, “Disentangling causal webs in the brain using functional magnetic resonance imaging: A review of current approaches,” *Netw Neurosci*, vol. 3, no. 2, pp. 237–273, 2019.

[16] I. Subramanian *et al.*, “Multi-omics data integration, interpretation, and its application,” *Bioinform. Biol. Insights*, vol. 14, p. 1177932219899051, Jan. 2020.

[17] W. Hu *et al.*, “Interpretable multimodal fusion networks reveal mechanisms of brain cognition,” *IEEE Transactions on Medical Imaging*, vol. 40, pp. 1474–1483, 2021.

[18] S. M. Gross and R. Tibshirani, “Collaborative regression,” *Biostatistics*, vol. 16, no. 2, pp. 326–338, Apr. 2015.

[19] X. Song *et al.*, “Joint sparse collaborative regression on imaging genetics study of schizophrenia,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2022.

[20] M. Kaya and H. S. Bilge, “Deep metric learning: A survey,” *Symmetry*, vol. 11, p. 1066, 2019.

[21] M. P. Van Den Heuvel and H. E. H. Pol, “Exploring the brain network: a review on resting-state fmri functional connectivity,” *European neuropsychopharmacology*, vol. 20, no. 8, pp. 519–534, 2010.

[22] S. Kato *et al.*, “Effects of head motion on the evaluation of age-related brain network changes using resting state functional MRI,” *Magn Reson Med Sci*, vol. 20, no. 4, pp. 338–346, Oct. 2020.

[23] J. Zhou *et al.*, “Graph neural networks: A review of methods and applications,” *ArXiv*, vol. abs/1812.08434, 2020.

[24] P. Velickovic *et al.*, “Graph attention networks,” *ArXiv*, vol. abs/1710.10903, 2018.

[25] M. Zhang and Y. Chen, “Link prediction based on graph neural networks,” in *Advances in Neural Information Processing Systems*, S. Bengio *et al.*, Eds., vol. 31. Curran Associates, Inc., 2018.

[26] T. D. Satterthwaite *et al.*, “Neuroimaging of the philadelphia neurodevelopmental cohort,” *NeuroImage*, vol. 86, pp. 544–553, 2014.

[27] H. Hotelling, “RELATIONS BETWEEN TWO SETS OF VARIATES*,” *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 12 1936. [Online]. Available: https://doi.org/10.1093/biomet/28.3-4.321

[28] G. Li *et al.*, “Application of deep canonically correlated sparse autoencoder for the classification of schizophrenia,” *Computer Methods and Programs in Biomedicine*, vol. 183, p. 105073, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169260719307655

[29] S. Akaho, “A kernel method for canonical correlation analysis,” *CoRR*, vol. abs/cs/0609071, 2006. [Online]. Available: http://arxiv.org/abs/cs/0609071

[30] O. Richfield *et al.*, “Learning schizophrenia imaging genetics data via multiple kernel canonical correlation analysis,” 2016, pp. 507–511.

[31] M. Zheng *et al.*, “Simmatch: Semi-supervised learning with similarity matching,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 14 471–14 481.

[32] Y. Atzmon, U. Shalit, and G. Chechik, “Learning sparse metrics, one feature at a time,” in *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, ser. Proceedings of Machine Learning Research, D. Storcheus, A. Rostamizadeh, and S. Kumar, Eds., vol. 44. Montreal, Canada: PMLR, 11 Dec 2015, pp. 30–48. [Online]. Available: https://proceedings.mlr.press/v44/atzmon2015.html

[33] T. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” *ArXiv*, vol. abs/1609.02907, 2017.

[34] P. Sayegh *et al.*, “Quality of education predicts performance on the wide range achievement test-4th edition word reading subtest.” *Archives of clinical neuropsychology : the official journal of the National Academy of Neuropsychologists*, vol. 29 8, pp. 731–6, 2014.

[35] J. D. Ragland *et al.*, “Working memory for complex figures: an fmri comparison of letter and fractal n-back tasks.” *Neuropsychology*, vol. 16 3, pp. 370–9, 2002.

[36] K. J. Friston *et al.*, “Characterizing dynamic brain responses with fmri: A multivariate approach,” *NeuroImage*, vol. 2, pp. 166–172, 1995.

[37] J. Fang *et al.*, “Fast and accurate detection of complex imaging genetics associations based on greedy projected distance correlation,” *IEEE Trans. Med. Imaging*, vol. 37, no. 4, pp. 860–870, Apr. 2018.

[38] J. D. Power *et al.*, “Functional network organization of the human brain,” *Neuron*, vol. 72, pp. 665–678, 2011.

[39] N. S. Dsouza *et al.*, “M-GCN: A multimodal graph convolutional network to integrate functional and structural connectomics data to predict multidimensional phenotypic characterizations,” in *Medical Imaging with Deep Learning*, 2021. [Online]. Available: https://openreview.net/forum?id=ud-iBiED9zb

[40] Y. LeCun *et al.*, “Gradient-based learning applied to document recognition,” in *Proceedings of the IEEE*, vol. 86, no. 11, 1998, pp. 2278–2324. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.42.7665

[41] E. T. Rolls *et al.*, “Automated anatomical labelling atlas 3,” *NeuroImage*, vol. 206, p. 116189, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811919307803

[42] T. D. Satterthwaite *et al.*, “Linked Sex Differences in Cognition and Functional Connectivity in Youth,” *Cerebral Cortex*, vol. 25, no. 9, pp. 2383–2394, 03 2014. [Online]. Available: https://doi.org/10.1093/cercor/bhu036

[43] R. Jiang *et al.*, “A neuroimaging signature of cognitive aging from whole-brain functional connectivity,” *Advanced Science*, vol. n/a, no. n/a, p. 2201621. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/advs.202201621

[44] J. Persson *et al.*, “Longitudinal assessment of default-mode brain function in aging,” *Neurobiology of Aging*, vol. 35, no. 9, pp. 2107–2117, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0197458014002668

[45] F. Fan *et al.*, “Development of the default-mode network during childhood and adolescence: A longitudinal resting-state fmri study,” *NeuroImage*, vol. 226, p. 117581, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811920310661

[46] G. Pan *et al.*, “Multiview diffusion map improves prediction of fluid intelligence with two paradigms of fmri analysis,” *IEEE Transactions on Biomedical Engineering*, vol. 68, no. 8, pp. 2529–2539, 2021.

[47] L. Zhang *et al.*, “Sensory, somatomotor and internal mentation networks emerge dynamically in the resting brain with internal mentation predominating in older age,” *NeuroImage*, vol. 237, p. 118188, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1053811921004651

[48] F. Van Overwalle, Q. Ma, and E. Heleven, “The posterior crus II cerebellum is specialized for social mentalizing and emotional self-experiences: a meta-analysis,” *Soc Cogn Affect Neurosci*, vol. 15, no. 9, pp. 905–928, 11 2020.

[49] G. Zhang *et al.*, “Detecting abnormal connectivity in schizophrenia via a joint directed acyclic graph estimation model,” *Neuroimage*, vol. 260, p. 119451, Jul 2022.

[50] L. Q. Uddin *et al.*, “Structure and Function of the Human Insula,” *J Clin Neurophysiol*, vol. 34, no. 4, pp. 300–306, Jul 2017.

[51] V. Mazzola *et al.*, “What Impact does An Angry Context have Upon Us? The Effect of Anger on Functional Connectivity of the Right Insula and Superior Temporal Gyri,” *Front Behav Neurosci*, vol. 10, p. 109, 2016.

[52] W. Sato *et al.*, “Increased putamen volume in adults with autism spectrum disorder,” *Frontiers in Human Neuroscience*, vol. 8, 2014. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2014.00957

[53] K. R. McLeod *et al.*, “Functional connectivity of neural motor networks is disrupted in children with developmental coordination disorder and attention-deficit/hyperactivity disorder,” *NeuroImage: Clinical*, vol. 4, pp. 566–575, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213158214000400

[54] J. Xu *et al.*, “Language in context: emergent features of word, sentence, and narrative comprehension,” *Neuroimage*, vol. 25, no. 3, pp. 1002–1015, Apr 2005.

[55] J. Zhao *et al.*, “Abnormal global-brain functional connectivity and its relationship with cognitive deficits in drug-naive first-episode adolescent-onset schizophrenia,” *Brain Imaging Behav*, vol. 16, no. 3, pp. 1303–1313, Jun 2022.

[56] Z. He *et al.*, “Functional dysconnectivity within the emotion-regulating system is associated with affective symptoms in major depressive disorder: A resting-state fMRI study,” *Aust N Z J Psychiatry*, vol. 53, no. 6, pp. 528–539, 06 2019.

[57] J. Liu *et al.*, “A dynamic causal modeling analysis of the effective connectivities underlying top-down letter processing,” *Neuropsychologia*, vol. 49, no. 5, pp. 1177–1186, Apr 2011.

[58] G. Delvecchio *et al.*, “Sexual Regional Dimorphism of Post-Adolescent and Middle Age Brain Maturation. A Multi-center 3T MRI Study,” *Front Aging Neurosci*, vol. 13, p. 622054, 2021.

[59] T. Bereczkei *et al.*, “Neural correlates of Machiavellian strategies in a social dilemma task,” *Brain Cogn*, vol. 82, no. 1, pp. 108–116, Jun 2013.

[60] B. Tomasino *et al.*, “Foreign accent syndrome: a multimodal mapping study,” *Cortex*, vol. 49, no. 1, pp. 18–39, Jan 2013.

[61] J. DiGuiseppi and P. Tadi, *Neuroanatomy, Postcentral Gyrus*. Treasure Island, FL: StatPearls Publishing, Jul 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK549825/

[62] K. Sander *et al.*, “Interhemispheric functional brain connectivity predicts new language learning success in adults,” *Cerebral Cortex*, 03 2022, bhac131. [Online]. Available: https://doi.org/10.1093/cercor/bhac131