Regularized partial correlation provides reliable functional connectivity estimates while correcting for widespread confounding

Kirsten L. Peterson[1,2,*], Ruben Sanchez-Romero[1], Ravi D. Mill[1], Michael W. Cole[1]

[1]Center for Molecular and Behavioral Neuroscience, Rutgers University, Newark, NJ, 07102
[2]Behavioral and Neural Sciences PhD Program, Rutgers University, Newark, NJ, 07102
*Corresponding author. Email address: kirsten.peterson@rutgers.edu

Keywords: Network neuroscience, fMRI, Regularization, Structural connectivity, Diffusion MRI, Individual differences

Abstract

Functional connectivity (FC) has been invaluable for understanding the brain's communication network, with strong potential for enhanced FC approaches to yield additional insights. Unlike with the fMRI field-standard method of pairwise correlation, theory suggests that partial correlation can estimate FC without confounded and indirect connections. However, partial correlation FC can also display low repeat reliability, impairing the accuracy of individual estimates. We hypothesized that reliability would be increased by adding regularization, which can reduce overfitting to noise in regression-based approaches like partial correlation. We therefore tested several regularized alternatives – graphical lasso, graphical ridge, and principal component regression – against unregularized partial and pairwise correlation, applying them to empirical resting-state fMRI and simulated data. As hypothesized, regularization vastly improved reliability, quantified using between-session similarity and intraclass correlation. This enhanced reliability then granted substantially more accurate individual FC estimates when validated against structural connectivity (empirical data) and ground truth networks (simulations). Graphical lasso showed especially high accuracy among regularized approaches, seemingly by maintaining more valid underlying network structures. We additionally found graphical lasso to be robust to noise levels, data quantity, and subject motion – common fMRI error sources. Lastly, we demonstrated that resting-state graphical lasso FC can effectively predict fMRI task activations and individual differences in behavior, further establishing its reliability, external validity, and ability to characterize task-related functionality. We recommend graphical lasso or similar regularized methods for calculating FC, as they can yield more valid estimates of unconfounded connectivity than field-standard pairwise correlation, while overcoming the poor reliability of unregularized partial correlation.

Introduction

The brain is a complex system, and to fully understand it we must understand how its components interact. Interactions between brain regions are typically investigated using functional/effective connectivity (FC) methods, which measure statistical relationships between regions' neural activities. The most common FC method used with functional magnetic resonance imaging (fMRI) has been, by far,

pairwise Pearson correlation (Biswal et al., 1995; Zalesky et al., 2012), with similar pairwise measures being used with other neuroimaging approaches (e.g., coherence with electroencephalography; (Srinivasan et al., 2007). Although pairwise correlation is easy to interpret and compute, numerous studies have demonstrated that it severely overestimates FC (for review see Friston, 2011; Reid et al., 2019), detecting indirect and false (confounded) connections in addition to true direct connections. These problems can be reduced by using multivariate FC methods such as partial correlation and multiple regression (Cole et al., 2016; Marrelec et al., 2006; Reid et al., 2019; Smith et al., 2011). Such methods improve upon pairwise correlation by conditioning on the time series of all measured regions (Figure 1A), letting them resolve confounding and indirect influences to estimate connectivity more validly.

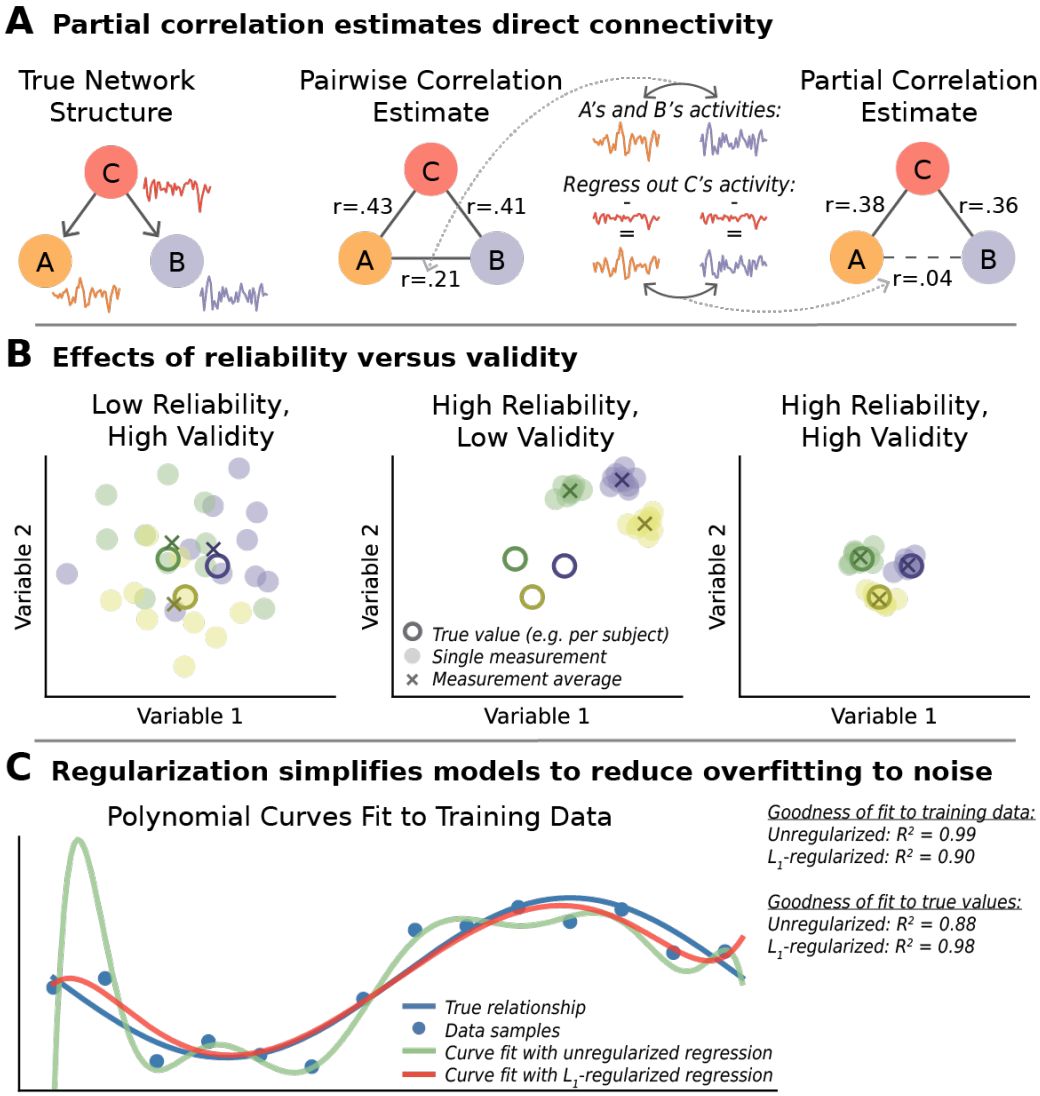

Figure 1 – Partial correlation corrects for confounds but suffers from low reliability, likely due to overfitting, which can be reduced with regularization. **A)** Partial correlation can estimate true connectivity more validly than pairwise correlation (the field standard for fMRI) in the presence of confounders (node C) (Reid et al., 2019). Here, regions A and B show a strong pairwise correlation despite not interacting, due to the third region C influencing (i.e. confounding) both A and B. Partial correlation takes region C into account to properly estimate the relationship between A and B. Empirical tests suggest confounder connectivity patterns are widespread throughout the brain, leading to thousands of false connections in many fMRI FC studies (Sanchez-Romero and Cole, 2021). **B)** An FC method must be both reliable and valid for its individual measurements to be consistently accurate (right panel). Recent studies have suggested that pairwise correlation is often reliable yet invalid (center) while partial correlation has low reliability but high validity (left; Mahadevan et al. 2021; Fiecas et al. 2013). Note that colors represent distinct entities such as subjects or task states, measured across multiple sessions. **C)** We hypothesized that partial correlation and related multivariate FC methods may show low reliability due to overfitting to noise, since estimated model coefficients that are more readily influenced by noise will be less stable with repeated sampling. We illustrate the issue of overfitting by fitting polynomial curves to simulated data (blue dots; true curve plus noise), where differing regression models were used to estimate the coefficients of polynomial terms. Note that multiple regression is the statistical basis for computing partial correlation. The curve produced by unregularized multiple regression (green line) is overly complex and specific to the noise in the training data. It differs substantially from the true underlying relationship (blue line) and would also differ from new curves fit to new data. L1-regularization constrains the flexibility of the model and results in a simpler but more accurate curve (red line).

While partial correlation and related FC methods have a clear theoretical advantage over pairwise correlation, they have also been shown to have worse repeat reliability (Fiecas et al., 2013; Mahadevan et al., 2021). Reliability gauges the stability of measurements and is essential to the accuracy of individual estimates. However, validity is an even more important criterion, assessing how well a method measures what it intends to measure. In this study, we define validity as being independent of reliability, evaluating the systematic correctness of a method, or how close the average of many repeated estimates is to the true value. Validity and reliability as used here can therefore be considered the analogues of bias and variance, concepts in statistics and machine learning. We then define individual measurement accuracy (or just "accuracy") as reflecting the closeness of individual estimates to the true value, a function of both validity and reliability that measures overall correctness. If a method is valid but unreliable (Figure 1B, left panel), as has been shown for partial correlation FC (Fiecas et al., 2013; Mahadevan et al., 2021), then its individual measurements would be frequently dissimilar from the true values due to their high variability, although they may approximate the truth in aggregate (e.g., after averaging over many sessions from a single subject, or many subjects in a population). Alternatively, a method can be reliable but not valid (Figure 1B, center), which more resembles pairwise correlation FC (Fiecas et al., 2013; Mahadevan et al., 2021; Sanchez-Romero & Cole, 2021). Such resulting measurements would be close to each other but far from the truth, meaning that none of the individual measurements were accurate representations. An ideal method would be reliable and valid (Figure 1B, right), producing stable measurements that are each close to the truth.

Given its otherwise high validity, partial correlation has the potential to be an extremely useful FC method if its low reliability were overcome (Fiecas et al., 2013; Mahadevan et al., 2021). We hypothesized that the reported instability of partial correlation and related multiple regression-based FC methods occurs from overfitting to

noise and could be ameliorated by regularization techniques. Overfitting to noise often results from a model's excessive complexity, and it can be exacerbated by factors such as low quantity and poor quality of the data being fit (Blum et al., 2020; Hastie et al., 2009; Ying, 2019). Such complexity arises in partial correlation and multiple regression FC with fMRI as the models fit more variables to include all measured nodes (brain regions or voxels). Increasing the complexity of a model allows it to better fit the specific training data by accounting for more variance, including noise. By capturing arbitrary patterns in the training data, such an overfit model will not generalize to independent data or accurately estimate coefficients (Blum et al., 2020; Hastie et al., 2009; Lever et al., 2016; Ying, 2019). We illustrate this problem in Figure 1C by fitting a polynomial regression model to noisy data, demonstrating that a more complex model (i.e., a model with more variables) overfits to the noise, reducing model reliability as it will fit to unique noise with every new data sample. In the case of FC estimation, overfit regression models will adjust connectivity coefficients to incorporate chance similarities in regions' activities, resulting in unstable coefficient weights that seldom reflect true connectivity values.

Regularization is a common strategy for reducing model complexity, and therefore reducing overfitting to noise, in both statistics and machine learning (Blum et al., 2020; Hastie et al., 2009; Ying, 2019). A variety of regularization techniques have been developed, such as explicitly penalizing complexity during model fitting. Two common methods are L_1 (lasso) and L_2 (ridge) regularization, which simplify models by penalizing them by the summed absolute values and squares of their coefficients, respectively (Hoerl & Kennard, 1970; Tibshirani, 1996). Another regularization method – principle component (PC) regression – works by fitting the model to a subset of PCs, reducing the number of variables (and hence model complexity) but keeping much of the presumed signal (Jolliffe, 1982).

While applying regularized multivariate methods to FC estimation is not completely novel, the practice remains substantially under-used relative to pairwise correlation. Indeed, several studies have tested and recommended regularized multivariate FC (Brier et al., 2015; Duff et al., 2013; Pervaiz et al., 2020; Smith et al., 2013), with some even introducing new implementations (Mejia et al., 2018; Nie et al., 2017; Ryali et al., 2012; Varoquaux et al., 2010). Still, other results did not clearly recommend the extra step of regularization (Fiecas et al., 2013; Mahadevan et al., 2021; Smith et al., 2011), and some analyses that used FC for prediction did not demonstrate an unequivocal benefit of regularized multivariate methods over pairwise correlation (Duff et al., 2013; Sala-Llonch et al., 2019). In addition, while there are many regularized methods available, few studies have tested the differences between them and even then typically to a limited extent (Brier et al., 2015; Mejia et al., 2018; Nie et al., 2017; Pervaiz et al., 2020; Ryali et al., 2012; Varoquaux et al., 2010). This lack of comparison between methods can leave researchers uncertain of which regularization approach to implement. Such ambiguities, as well as a lack of understanding of these regularized multivariate methods, may prevent researchers from adopting methods that would otherwise benefit their analyses.

The purpose of this study is to test the suitability of regularized partial correlation and similar multivariate methods for estimating causally valid FC. To do this, we compared the performances of three established regularized methods – graphical lasso,

graphical ridge, and PC regression – with unregularized partial correlation and the field-standard pairwise correlation. The use of three regularized methods allowed us to generalize utility of the fundamental concept of regularization across techniques, as well as test for differences between them.

We began by separately testing the reliability, validity, and individual measurement accuracy of the FC methods, where we define validity as being independent of reliability while accuracy is not (see above). In empirical data, reliability was quantified as between-session similarity and intraclass correlation. We quantified validity and individual measurement accuracy as the convergence of FC estimates (group-averaged and individual, respectively) with the corresponding subjects' structural connectivity, computed from diffusion MRI data. To confirm our findings, we also assessed the reliability, validity, and individual measurement accuracy of FC in simulated data, comparing estimated FC matrices to ground truth networks. We then extended our assessment of FC measurement accuracy and validity by examining the resilience of the FC methods to practical pitfalls of fMRI data: short scan lengths, scanner noise (Blum et al., 2020; Hastie et al., 2009; Ying, 2019), and subject head movement artifacts (Power et al., 2015). Lastly, we further broadened our assessment of FC accuracy to task-related functionality, using regularized FC to generate held-out task activations with activity flow modelling (Cole et al., 2016) and using FC to predict individual differences in subject age and intelligence. Together, these tests of FC method reliability, group-level validity, and overall accuracy demonstrate the effectiveness of regularized partial correlation FC for improving upon the field-standard pairwise correlation FC approach.

Methods

Functional connectivity estimation methods

This study compared the performance of five FC methods: pairwise Pearson correlation, partial correlation, graphical lasso, graphical ridge, and PC regression. All methods estimated FC from the same empirical resting state or simulated timeseries, which were always z-scored (Hastie et al., 2009). Each FC matrix was calculated independently from a single session of data, coming from a single subject or simulated network.

Pairwise correlation. Pairwise correlation FC was computed as the Pearson correlation between each pair of timeseries.

Partial correlation. Partial correlation can be calculated in two ways with near identical results. The more intuitive approach involves regressing the timeseries of all other nodes (the conditioning variables) from the two target nodes' timeseries, and then computing the Pearson correlation between the two residuals (Figure 1A). This study instead used the inverse covariance approach, which is advantageous because it is less computationally expensive. First the covariance matrix of all nodes is inverted, giving the precision matrix P. Then the partial correlation coefficients are calculated as:

$$r_{AB|\mathrm{C}} = -\mathrm{P}_{AB}/\sqrt{\mathrm{P}_{AA}\mathrm{P}_{BB}},$$

for nodes *A* and *B* conditioned on set **C**.

Graphical lasso. The first of our regularized methods, graphical lasso ("glasso") implements partial correlation with an L_1 penalty to limit model complexity (Friedman et al., 2008; Tibshirani, 1996). The penalty is applied when computing the precision matrix,

and this regularized precision matrix is then transformed into the partial correlation matrix. L_1 regularization works by adding to the model cost function the term:

$$\lambda_1 \sum_{j \neq k} |\mathrm{P}_{jk}|,$$

which is proportional to the summed absolute values of entries in the estimated precision matrix (P), not including the diagonal. The amount of regularization is scaled by the hyperparameter λ_1. L_1 regularization tends to drive less informative coefficients to exactly zero, producing a sparse result. In this way, L_1 regularization can also perform feature selection. We implemented graphical lasso using the Python package GGLasso (Schaipp et al., 2021).

Graphical ridge. Graphical ridge applies L_2 regularization (also called Tikhonov regularization) to partial correlation (Hoerl & Kennard, 1970). It applies the L_2 penalty term:

$$\lambda_2 \sum_{j \neq k} \mathrm{P}_{jk}^2,$$

which is the summed square values of entries in the estimated precision matrix (P) multiplied by the hyperparameter λ_2. While L_2 regularization also encourages model simplification (i.e., by bringing coefficients closer to zero), it does not cause sparsity to the extent that L_1 regularization does (shrinking less meaningful coefficients to exactly zero). Because of the square term, L_2 regularization exerts uneven pressure on edges based on their coefficient weights, with low coefficients contributing disproportionately small amounts to the penalty term. As coefficients are shrunk, their penalties become increasingly negligible so that they are seldom brought to exactly zero as occurs with L_1 regularization. Meanwhile, high coefficients produce comparatively much larger penalties, causing the model to be biased against higher weights. In this way, L_2 regularization encourages a narrower range of weights that are shared more evenly across coefficients (see Figure 2). We implemented graphical ridge using the R package rags2ridges (Peeters et al., 2022).

Principal components regression. PC regression combines multiple regression with principal component analysis (PCA) to induce regularization (Hastie et al., 2009; Jolliffe, 1982). To construct an FC matrix using any multiple regression method, the regression model is fit once for each node, with that target node's timeseries being predicted by the timeseries of all other nodes. The row of the FC matrix that corresponds with the target node is filled in with the beta coefficients for each predictor node. To perform PC regression, PCA is first applied to the predictor variables, and only a subset of the PCs is given to the regression model. Number of PCs is a hyperparameter for this method, usually selected as the *n* PCs accounting for the highest variance. This reduces overfitting to noise by reducing the number of variables in the regression model and presumably discarding noisier or less relevant dimensions of data. After fitting the PCs to the original to-be-predicted target timeseries, the PC coefficients are transformed to reflect the predictor nodes' contributions, which become entries in the FC matrix. We used Python Scikit-learn's LinearRegression and PCA functions for this method. Using multiple regression to calculate FC in this way creates an asymmetrical matrix, but this does not necessarily represent directionality of connections as FC asymmetry usually implies. To make them more comprehensible, we symmetrized each PC regression FC matrix by averaging the original with its transpose.

Hyperparameter selection. Regularization often requires the selection of hyperparameters, the choice of which can greatly impact the resulting FC estimates. We

therefore had to optimize the hyperparameters within each regularized method before making comparisons between methods. In the methods we tested, the hyperparameters are λ_1 for graphical lasso, λ_2 for graphical ridge, and number of PCs for PC regression. For graphical lasso and graphical ridge, lambda values of zero would coincide with unregularized partial correlation and higher values would lead to a greater degree of regularization. For PC regression, using all components would yield unregularized multiple regression while using fewer components would produce more regularization. If the hyperparameter of any method does not induce enough regularization, then (according to our hypothesis) the model will remain overfit, making the FC estimates unreliable (i.e., having high variance). If the hyperparameter induces too much regularization, however, the model will be underfit, discarding relevant information such that the FC estimates are less valid (i.e., having high bias; Hastie et al., 2009; Lever et al., 2016). The optimal value would balance these two tendencies.

We sought a measure of model fit that could be applied in the same way for all FC methods that we tested, to limit the possibility of model fit metrics biasing FC method performance. Our solution was to determine optimal hyperparameter values as those whose models can best predict held-out timeseries data. This form of cross-validation is a standard machine learning approach for testing a regression model's accuracy in the context of fitting time series (Pardoe, 2020). Resting-state fMRI time series were used for both model fitting and testing with held-out data. For each row of an FC matrix (representing all connections to a single node), the connectivity weights were treated as beta coefficients in a regression model to predict that single node's held-out activity from the concurrent activities of all other nodes. Prediction accuracy was then calculated as the coefficient of determination (R^2) between predicted and actual activities. This was applied with 10-fold cross-validation within each individual session timeseries, where each fold served as held-out, to-be-predicted data for one iteration while all other folds were used to compute the FC matrix. We made sure to use a similar number of timepoints to produce these cross-validation matrices (90% of timepoints) as we would for the final matrices, as we found that models fit with less data tend to prefer a greater level of regularization (see Results). This was repeated for all tested hyperparameter values, and the optimal hyperparameter for each session was that which produced the highest R^2 value averaged over all folds.

After initial exploration, we tested the following ranges of hyperparameters for empirical data: for graphical lasso, λ_1 = 0-0.1 (increments of 0.001 for 0-0.005, then 0.005); for graphical ridge, λ_2 = 0-2 (increments of 0.02 for 0-0.1, then 0.1); for PC regression, number of PCs = 10-359 (most possible components; increments of 10). For simulated data, we tested the ranges: for graphical lasso, λ_1 = 0-0.25 (increments of 0.002 for 0-0.01, then 0.01; upper limit extended to 0.3 when varying noise and data quantity); for graphical ridge, λ_2 = 0-2.5 (increments of 0.02 for 0-0.1, then 0.1; upper limit extended to 6 when varying noise and data quantity); for PC regression, number of PCs = 5-99 (most possible components; increments of 5).

Empirical MRI data and processing

Our empirical analyses used the Human Connectome Project in Aging (HCP-A) dataset (Bookheimer et al., 2019; Harms et al., 2018), which is publicly available through the NIMH Data Archive. This extensive dataset includes behavioral measures

and high-quality multimodal MRI for 1200+ participants sampled across the adult lifespan (36-100+), with collection of additional measures (diffusion data, multiple behavioral measures) used in the present study. Participants were recruited from the areas surrounding the four acquisition sites (Washington University St. Louis, University of Minnesota, Massachusetts General Hospital, and University of California, Los Angeles). All participants gave informed consent through the institutional review board associated with each recruitment site. We excluded from our analyses any participants who were noted to have quality control issues or who were missing any of the MRI scans or behavioral measures that we utilized in this study. This left us with 472 subjects, divided evenly between discovery (n = 236 subjects, 141 females; mean age = 56.9 years, SD = 13.95) and replication (n = 236 subjects, 134 females; mean age = 58.2 years, SD = 14.35) datasets.

Our analyses utilized resting-state fMRI, task fMRI, and diffusion MRI from HCP-A (Harms et al., 2018). Across the four sites, data was collected using a Siemens 3T Prisma scanner. fMRI scans were acquired with TR = 800 ms, 72 slices, and 2.0 mm isotropic voxels. The resting-state fMRI data were collected in four runs over two days, each run containing 488 volumes and lasting 6 min 41 s. For task fMRI data we analyzed the Go/No-go task, which contained 300 volumes and lasted 4 min 11 s. We used the publicly available minimally preprocessed fMRI data, processed using the HCP minimal preprocessing pipeline (Glasser et al., 2013). The preprocessed cortical surface data were parcellated into 360 brain regions according to the multimodal Glasser parcellation (Glasser et al., 2016). We did not apply ICA-FIX denoising in favor of our own nuisance regression procedures, previously described by Ito et al. (2020). We first removed the first 5 frames from each run and demeaned and linearly detrended the timeseries. We then performed nuisance regression as described by Ciric et al. (2017) with 24 motion regressors and 40 physiological noise regressors, the latter being modeled from white matter and ventricle timeseries components using aCompCor (Behzadi et al., 2007). Note that – as in Cole et al. (2021) – aCompCor was used in place of global signal regression, given evidence that it has similar benefits as global signal regression for removing artifacts (Power et al., 2018) but without regressing gray matter signals (mixed with other gray matter signals) from themselves, which may result in false correlations (Murphy et al., 2009; Power et al., 2017). The cleaned resting state data was then partitioned into two sessions of 956 TRs (12 min 45 s), the concatenations of runs 1-2 and runs 3-4. Those two sessions of resting-state data were used to produce two FC matrices per subject for all FC methods (see *Functional connectivity estimation methods*).

We next estimated task-evoked activations from task fMRI timeseries. We used the Go/No go task, in which 92 stimuli (simple geometric shapes) were presented. We estimated the mean activation values separately for "hit", "correct rejection", "false alarm", and "miss" events by fitting a general linear model with 4 separate regressors (canonical hemodynamic response function convolved with event onsets). Our analyses only included the "hit" and "correct rejection" regressors, as these were the events where subjects responded correctly. Our task regression was performed concurrently with the nuisance regression described above.

Diffusion MRI is used to estimate white matter tracts, or structural connectivity. It measures the diffusion of water molecules in each voxel in different directions. Where

the molecules are only able to move freely in certain directions, this typically corresponds with a white matter tract. HCP-A acquired diffusion MRI with 1.5 mm isotropic voxels, sampling 92-93 directions in each of two shells (b = 1500 and 3000 s/mm^2) and repeating the entire acquisition in both AP and PA phase encoding directions (Harms et al., 2018). This large quantity of data aids in making accurate estimations of structural connectivity (SC). We preprocessed the data using the HCP diffusion pipeline (Sotiropoulos et al., 2013). Tractography was performed using DSI Studio (Yeh et al., 2013), as it was shown to be among the most accurate pipelines tested by Maier-Hein et al. (2017). We first estimated local fiber orientations using generalized Q-sampling imaging. We then transformed each subject's individualized cortical parcellation into volume space from surface space to use with DSI Studio. Tractography was performed with a deterministic algorithm, by placing thousands of seeds throughout the white matter volume and letting them spread as streamlines from the initial point, directed through 3-dimensional space by the local fiber orientation vector field. Structural connection weights between each pair of regions were determined as the normalized count of streamlines that overlap with voxels in both regions. These streamlines estimate the presence and size of white matter tracts physically linking the brain areas.

Out-of-scanner behavioral measures were used to gauge individual differences in intelligence. We used factor analysis to estimate psychometric *g*, or the general intelligence factor, for each subject (Johnson et al., 2008; McCormick et al., 2022). We analyzed data from all available measures of fluid intelligence (Bookheimer et al., 2019), which consisted of the Picture Sequence Memory Test, Dimensional Change Card Sort Test, Flanker Task Control and Attention Test, Pattern Completion Processing Speed Test, and List Sorting Working Memory Test from the NIH Toolbox (Weintraub et al., 2013), as well as the Rey Auditory Verbal Learning Test (Rey, 1941) and Trail Making Test B (Bowie & Harvey, 2006). All scores were unadjusted for age. Factor analysis was applied to all subjects' test scores (separately for discovery and replication datasets) to estimate the unitary factor loadings underlying all cognitive measures. The *g* scores were then calculated for each subject, reflecting general intelligence in a single value. We implemented factor analysis using the R package psych (Revelle, 2017).

Simulated networks and timeseries

Our simulation analyses involved creating random network connections, simulating activity timeseries for all nodes using linear modelling, estimating FC on those simulated data using each FC method, and then comparing the resulting FC matrices. The generated networks were directed, as this was required to simulate activity. They were designed to be modular (organized into communities), small-world (showing high clustering but efficient paths between all nodes), roughly scale-free (containing densely connected hub nodes), and rich-club (having hubs highly connected to each other), because the brain network has been shown to have these characteristics (Bullmore & Sporns, 2009; van den Heuvel & Sporns, 2011).

Networks were comprised of 100 nodes organized into 5 modules of 20 nodes, with edges added based on the Barabási-Albert model (Albert & Barabási, 2002). For each module, we started with 2 connected nodes and added one at a time until reaching the full size of 20. Each node came with 2 attached edges (binary, undirected), which

attached to existing nodes with preference for those with high degree. That is, the probability of the new edge attaching to any node *i* over each other possible node *j* was:

$$p_i = k_i / \sum_j k_j,$$

where k is node degree, or the number of edges connected to that node. We then added extra-modular edges, adding them one at a time until there were 10% as many edges outside of modules as within modules. The nodes to be connected were again chosen with higher preference given to nodes with high degree, with each of the two nodes having the above probability of being attached. This resulted in all networks having 408 directed (or 204 bidirectional) edges (4.12% of all possible edges; 370 or 19.47% of intra-modular edges; 38 or 0.48% of extra-modular edges). After establishing this binary backbone, we randomly assigned weights to each directed edge from a uniform distribution ranging from 0 to 0.3, not inclusive. 100 networks were generated for the main analyses and an additional 100 networks were used when testing the impact of the amount of data and noise level.

Activity timeseries were simulated from linear models as described by Sanchez-Romero and Cole (2021) and Sanchez-Romero et al. (2023). The relationships of variables in the network were described by the linear model:

$$X = WX + E,$$

where X is a dataset of p nodes with n datapoints, E is a dataset of p normally distributed, independent intrinsic noise terms with n datapoints, and W is the matrix of connectivity weights between all p nodes, which must have its diagonal equal to zero. The activities in X were calculated by expressing the linear model as:

$$X = (I - W)^{-1}E,$$

where I is the identity matrix. Additional measurement noise, scaled by intrinsic noise level, was then added to each timeseries. For the main analyses, we simulated 250 datapoints per session and 50 sessions for each of the 100 networks. Each had a noise level of 0.5, meaning that the added measurement noise had roughly half the amplitude of the original signal. When simulating data with varying amounts of data and noise levels, we simulated one session for each combination of 100, 200, 300, 400, 500, 1000, and 10000 datapoints and 0.25, 0.50, and 1.00 noise levels for each of the 100 additional networks. All FC methods were applied to each session of simulated data (see Functional connectivity estimation methods).

Analyses

Between-session similarity. Our primary measure of repeat reliability was between-session similarity, which we calculated as the Pearson correlation of all edge weights (vectorized upper triangle) between the FC matrices of two sessions. For empirical data, this was calculated between the two session FC matrices for each subject (n = 236 within-subject pairs), and for simulated data, this was calculated between the first pair of session FC matrices for each network (n = 100 within-network pairs). This controls for subject differences in empirical FC. While there may still be some legitimate differences in FC between sessions, we expect this contribution to be relatively small. FC estimated from simulated data should show no between-session differences, as they were generated from the exact same network. A strong correlation indicates a small degree of variability (high reliability), while a weak correlation indicates that noise is obscuring the underlying edge weights.

Intraclass correlation. The intraclass correlation coefficient (ICC) has previously been used in FC studies to quantify repeat reliability (Fiecas et al., 2013; Mahadevan et al., 2021; Mejia et al., 2018). Unlike the other metrics presented here, it is calculated for each edge rather than for each subject. We calculated for each edge one-way random-effects ICC, or ICC(1,1), according to Shrout and Fleiss (1979):

$$ICC = \frac{BMS - WMS}{BMS + (k-1)WMS},$$

where BMS is the mean squares between subjects (variability in subject-averaged edge weights across the group), WMS is the mean squares within subjects (variability in session edge weights within each subject), and k is the number of sessions (k = 2 for this study). BMS should include true variability across subjects plus noise variability, while WMS should primarily contain noise variability.

While intraclass correlation is a generally valid metric, we observed that it is biased against sparse FC methods if applied to all edges. As can be seen from the above formula, ICC is zero when WMS and BMS are equal, which happens when there is only noise and no variability specific to individual differences. When estimating connectivity in the brain, many FC methods yield some null edges that systematically approximate zero for all subjects and only vary from zero due to noise. These null edges would accordingly have ICCs distributed around zero, and sparse FC methods would then have ICCs around zero for the majority of their edges. To avoid this automatic penalty for null edges, we therefore only analyzed edges which were systematically assigned nonzero weights, leaving the potential for between-subject variability. To accommodate a high degree of sparsity, we chose these edges as those which had group-averaged weights in the 98th percentile for every FC method being tested, giving us 539 edges.

Structural-functional similarity. Like partial-correlation FC and multiple-regression FC, diffusion weighted MRI is used to measure direct connections between brain regions, but it does so by mapping white matter tracts. Because structural connectivity is calculated from an entirely different MRI modality from FC, it has highly distinct biases and error sources. Structural connectivity is not a perfect ground truth, as it has been shown to report many false positive and false negative connections (Maier-Hein et al., 2017; Rheault et al., 2020; Sotiropoulos & Zalesky, 2019), and (unlike FC) it cannot reflect the aggregate effects of microanatomy (e.g., synaptic strengths). However, given that a direct functional connection can only exist if a structural connection exists as well, and given that diffusion-weighted MRI reflects structural connectivity above chance (Donahue et al., 2016), this is still an appropriate validation benchmark.

We calculated structural-functional similarity as the Pearson correlation between functional and structural connectivity weights. It was computed from both group-averaged and individual FC to assess validity and individual measurement accuracy, respectively. For validity, the group-averaged FC matrix was compared with the group-averaged structural matrix. For individual measurement accuracy, we compared each session 1 FC matrix and the SC matrix from the same subject (n = 236 individual matrices). Averaging over many estimates can nullify much of the measurement noise, leaving only the underlying connectivity structures. The group-averaged structural-functional similarity should therefore reflect the validity of each FC method without the influence of reliability, while the individual similarities depend on both.

Ground truth similarity. For simulated data, we computed ground truth similarity as the Pearson correlation between each FC matrix and the corresponding ground truth weights matrix. Since all FC methods that we implemented estimate undirected connectivity, we calculated correlations with symmetrized versions of the ground truth matrices. As we did with structural-functional similarity, we computed ground truth similarity using group-averaged as well as individual FC matrices, as averaging reduces the impact of poor reliability to indicate just group-level validity. For the main simulation analyses, we calculated ground truth similarity of group-averaged matrices using FC matrices averaged over 50 sessions for each network (n = 100 group-averaged matrices) and individual matrices using only the session 1 FC matrix from each network (n = 100 individual matrices). When analyzing the effect of amount of data and noise level, we calculated ground truth similarity of individual FC matrices for each network (n = 100 individual matrices) for all conditions.

Correlation with subject motion. Head motion artifacts in fMRI timeseries can lead to spurious, systematic changes in FC estimates, often dependent on the distance between regions (Power et al., 2015). We quantified the influence of subject motion on FC weights using quality control-functional connectivity (QC-FC) correlation, as described by Ciric et al. (2017). QC-FC correlation was calculated for each edge as the partial correlation between subjects' estimated edge weights and mean relative root mean square (RMS) displacement, conditioned on subject age and sex. We calculated median QC-FC correlation and percentage of edges with significant correlations ($p < .05$ following false discovery rate correction; Benjamini & Hochberg, 1995) for each FC method. We additionally calculated QC-FC correlation while controlling for FC sparsity. This was done by re-calculating each QC-FC correlation while only including subjects with nonzero weights ($|w| > .01$) for the given edge. Only edges where at least 80% of subjects had nonzero weights were analyzed.

Task activation prediction. Task activation prediction measures how accurately the FC matrices relate to independent functional activity. For this we used the task-evoked activations for "hit" and "correct rejection" events in the Go/No go task. Predictions were generated using activity flow modeling, which simulates the generation of activity in a target node from the activity flows of all other nodes over their connections with the target (Cole et al., 2016). The predicted activity of each target node was calculated as:

$$\hat{A}_j = \sum_{i \neq j} A_i C_{ij},$$

where A_i is the activity of source node *i* and C_{ij} is the FC coefficient between the target *j* and source node *i*. Then, the similarity of the predicted activity ($\hat{A}_j$) to actual activity (A_j) was quantified using the coefficient of determination (R^2) or Pearson correlation.

Age and intelligence prediction. Lastly, we tested how well the FC weights from different methods could predict subject age and intelligence (psychometric *g* score). As fluid intelligence often declines with age (Harada et al., 2013; Salthouse, 2010), we first regressed subject age from the *g* scores so that we could examine their relationship with FC independently. We predicted age and age-adjusted *g* scores separately by fitting ridge regression models within 10-fold cross-validation. 9 folds of subjects were used to train each regression model, with FC edge weights (session 1 only) as predictor variables. We used ridge regression because the number of variables (n = 64620 edges) is much larger than the number of observations (n = 212 or 213 subjects in the 9

folds). The hyperparameter values were selected through nested cross-validation by Scikit-learn's RidgeCV function. We then applied each model to the held-out fold of subjects to calculate their predicted *g* scores and ages. The predicted values were pooled over all folds, and we quantified model accuracy for each FC method as the Pearson correlation between predicted and actual values.

Results

Visualizing empirical fMRI functional connectivity

Before comparing the different FC methods based on purely quantitative measures, we found it informative to first visualize the resulting FC matrices. We present both group-averaged and individual subject matrices in Figure 2, with the nodes ordered according to the Cole-Anticevic Brain-wide Network Partition to reveal the network architecture (Ji et al., 2019; Figure 2A). For the regularized methods, we measured the mean optimal hyperparameters for both sessions (n = 236 subjects) to be λ_1 = 0.034 for graphical lasso (SD = 0.006), λ_2 = 0.434 for graphical ridge (SD = 0.163), and 54.3 PCs for PC regression (SD = 12.9). The group-averaged matrices facilitate comparing the network structures generated by each FC method. Meanwhile, individual subject matrices can expose low measurement reliability if the edge weights vary substantially from the group-averaged weights.

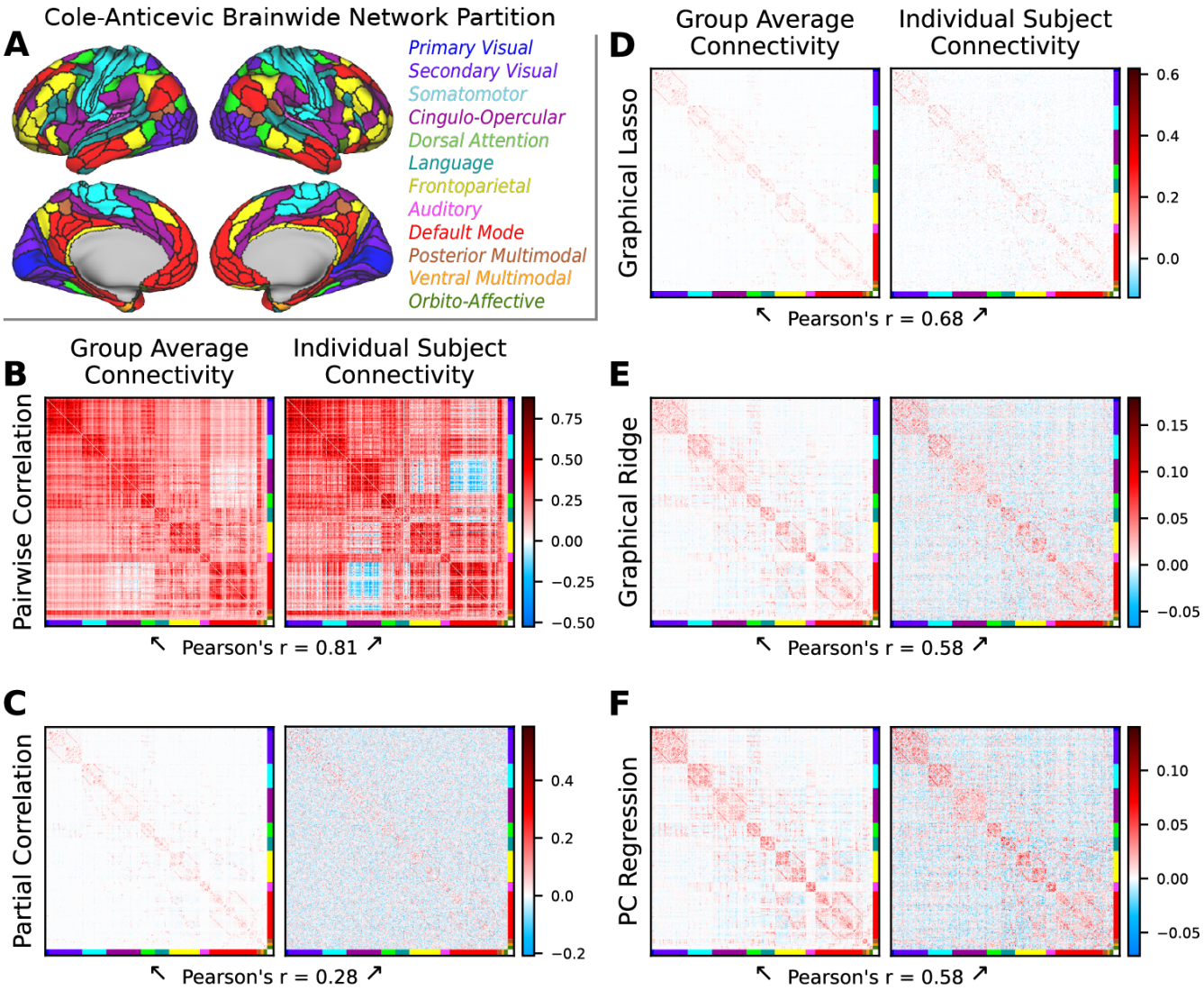

Figure 2 – Empirical FC matrices produced by different methods. The nodes were ordered according to the Cole-Anticevic Brain-wide Network Partition **(A)** to visualize the network structure (Ji et al., 2019). Group-averaged connectivity matrices (N=236; left) are shown beside a random subject's individual matrices (right). The Pearson correlation between the individual FC matrix and the group average is printed below each set of matrices. **B)** Pairwise correlation (the field standard), **C)** partial correlation, **D).**graphical lasso, **E)** graphical ridge, and **F)** PC regression.

Two contrasting FC methods are pairwise correlation and partial correlation. Pairwise correlation (the field standard) can be seen to produce dense connectivity, as it reflects any linear time series similarity between nodes (Figure 2B). Partial correlation instead creates a sparse network, which is theoretically limited to direct and unconfounded connections (Figure 2C). However, this sparse network structure is obscured by edge weights that vary widely from the group average, possibly reflecting noise. We will investigate this possibility quantitatively in subsequent sections.

We tested several different regularization methods, and displaying their resulting FC matrices illustrates some clear differences between graphical lasso, graphical ridge, and PC regression (Figure 2D-E). Graphical lasso can be seen to produce a similar

sparse network structure to partial correlation at the group level. Looking at its individual subject matrix, however, suggests that graphical lasso may produce less edge weight variability. The networks resulting from graphical ridge and PC regression appear similar to each other in that their connectivity graphs are more dense than that of graphical lasso, with lower weights spread over more edges. Their structures appear to diverge somewhat from that of partial correlation, which may bode poorly for the methods if partial correlation FC is in fact a valid representation of direct brain connectivity at the group level. Like graphical lasso, graphical ridge and PC regression show less deviation between individual and group-averaged FC.

Reliability of FC with empirical fMRI data

Low repeat reliability has been observed with partial correlation FC (Fiecas et al., 2013; Mahadevan et al., 2021), which may mean that this otherwise effective method is inaccurate at the level of individual measurements. We hypothesized, however, that regularization could solve this problem. We tested the reliability of FC methods in empirical data using between-session similarity and intraclass correlation.

Replicating previously published results, we found partial correlation FC had low between-session similarity (mean r = 0.103, SD = 0.015; Figure 3A) and low ICCs (mean ICC = .300, SD = .116; Figure 3B). This is especially poor when compared with pairwise correlation FC, which scores significantly higher for both between-session similarity (mean r = 0.797, SD = 0.049; $p < .00001$) and ICCs (mean ICC = 0.704, SD = 0.063; $p < .00001$). Our comparison statistics (see Table 1, row 1) confirm that pairwise correlation FC has significantly higher reliability than partial correlation FC.

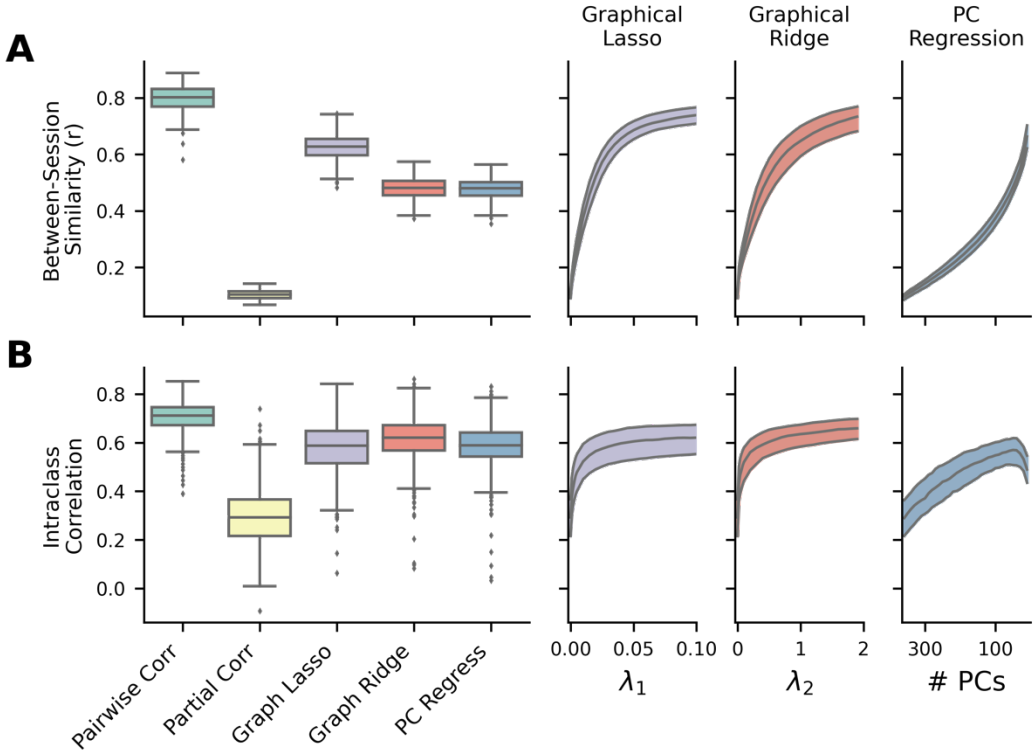

Figure 3 – Reliability of FC methods with empirical fMRI data. The boxplots show results where the regularization hyperparameters have been optimized for each FC matrix, while the right plots show the medians and IQRs across different hyperparameter values for the regularized methods. For PC regression, number of PCs is plotted in descending order because fewer PCs correspond with more regularization. **A)** Between-session similarity, the Pearson correlation between each subject's session 1 and session 2 FC matrices (n = 236). **B)** Intraclass correlation (Shrout & Fleiss, 1979), calculated for each edge in a conservative subset of nonzero edges (n = 539).

Table 1 – Statistical tests comparing FC methods on reliability and structural-functional similarity using empirical data. Between-session similarity, intraclass correlation, and individual structural-functional similarity were compared using two-tailed, dependent-sample t-tests. The Pearson's r values for between-session similarity and individual structural-functional similarity were normalized using Fisher's z transformation (arctanh) before statistical comparisons. Group-averaged structural-functional similarity (single Pearson's r per method) was compared using the method described by Meng et al. (1992), which tests for a significant difference between "correlated correlations" (correlations with a shared variable – structural connectivity weights in this case). For all reported correlation differences, correlations were transformed to Fisher's z, subtracted, and then transformed back to Pearson's r. Alpha levels were adjusted to .005 (.05/10) following Bonferroni correction to account for the multiple comparisons within each column.

Methods		Between-session similarity (Pearson's r)	Intraclass correlation (ICC)	Structural-functional similarity (Pearson's r)	
				Group-averaged	Individual
1	Pairwise correlation vs. partial correlation	mean Δr = 0.762, T(235) = 118, p < .00001	mean ΔICC = 0.407, T(538) = 80.8, p < .00001	Δr = -0.377, z = -5.42, p < .00001	mean Δr = 0.029, T(235) = -29.0, p < .00001
2	Pairwise correlation vs. graphical lasso	mean Δr = 0.352, T(235) = 50.2, p < .00001	mean ΔICC = 0.127, T(538) = 28.7, p < .00001	Δr = -0.337, z = -4.98, p < .00001	mean Δr = -0.163, T(235) = -123, p < .00001
3	Pairwise correlation vs. graphical ridge	mean Δr = 0.525, T(235) = 77.8, p < .00001	mean ΔICC = 0.090, T(538) = 24.8, p < .00001	Δr = -0.212, z = -3.80, p = .0001	mean Δr = -0.059, T(235) = -59.3, p < .00001
4	Pairwise correlation vs. PC regression	mean Δr = 0.526, T(235) = 79.3, p < .00001	mean ΔICC = 0.120, T(538) = 33.4, p < .00001	Δr = -0.163, z = -2.95, p = .003	mean Δr = -0.033, T(235) = -34.1, p < .00001
5	Partial correlation vs. graphical lasso	mean Δr = -0.560, T(235) = -143, p < .00001	mean ΔICC = -0.281, T(538) = -76.9, p < .00001	Δr = 0.046, z = 1.87, p = .061	mean Δr = -0.191, T(235) = -166, p < .00001
6	Partial correlation vs. graphical ridge	mean Δr = -0.395, T(235) = -157, p < .00001	mean ΔICC = -0.317, T(538) = -68.6, p < .00001	Δr = 0.179, z = 4.47, p < .00001	mean Δr = -0.088, T(235) = -124, p < .00001
7	Partial correlation vs. PC regression	mean Δr = -0.394, T(235) = -126, p < .00001	mean ΔICC = -0.287, T(538) = -56.9, p < .00001	Δr = 0.228, z = 4.94, p < .00001	mean Δr = -0.062, T(235) = -90.2, p < .00001
8	Graphical lasso vs. graphical ridge	mean Δr = 0.212, T(235) = 73.6, p < .00001	mean ΔICC = -0.036, T(538) = -11.6, p < .00001	Δr = 0.134, z = 3.43, p = .0006	mean Δr = 0.105, T(235) = 122, p < .00001
9	Graphical lasso vs. PC regression	mean Δr = 0.214, T(235) = 57.3, p < .00001	mean ΔICC = -0.007, T(538) = -1.77, p = .077	Δr = 0.184, z = 4.14, p = .00003	mean Δr = 0.130, T(235) = 138, p < .00001
10	Graphical ridge vs. PC regression	mean Δr = 0.002, T(235) = .801, p = .424	mean ΔICC = 0.030, T(538) = 21.4, p < .00001	Δr = 0.051, z = 5.64, p < .00001	mean Δr = 0.025, T(235) = 112, p < .00001

In accordance with our hypothesis, all three regularized methods significantly improved on the reliability displayed by partial correlation, shown with both between-session similarity (graphical lasso: mean r = 0.625, SD = 0.047; graphical ridge: mean r = 0.478, SD = 0.037; PC regression: mean r = 0.478, SD = 0.039; all p < .00001, Table 1, rows 5-7) and edge ICCs (graphical lasso: mean ICC = 0.577, SD = 0.106; graphical ridge: mean ICC = 0.614, SD = 0.096; PC regression: mean ICC = 0.584, SD = 0.096; all p < .00001, Table 1, rows 5-7). Among the regularized methods, graphical lasso produced the highest between-session similarity (significantly higher than graphical

ridge and PC regression; both $p < .00001$, Table 1, rows 8 and 9), and graphical ridge and PC regression showed no significant difference ($p = .424$, Table 1, last row). Their ICCs were very close, with graphical ridge showing a slight lead (significantly higher than graphical lasso and PC regression; both $p < .00001$, Table 1, rows 8 and 10) and graphical lasso and PC regression showing no significant difference ($p = .077$, Table 1, row 9). While the regularized methods did not reach the high scores of pairwise correlation in either between-session similarity or edge ICCs (all $p < .00001$, Table 1, rows 2-4), they did come much nearer than unregularized partial correlation. Each regularized method could have been made more reliable by implementing a greater degree of regularization, as demonstrated in the right panels of Figure 3A-B. However, we would not recommend optimizing hyperparameters for reliability, as we have found this to trade off against the validity of the resulting FC matrices (see below). All results were also replicated in the replication dataset (Supplementary Figure S1 and Supplementary Table S1).

Similarity of FC to structural connectivity

While it is important for measurement tools to be reliable, it is imperative that they be valid (Figure 1B), meaning that they represent what they are intended to. We approximated the validity and individual measurement accuracy of empirical FC by comparing it to an independent estimate of direct brain connections – structural connectivity. Structural connectivity represents the presence of white matter tracts between regions, and since functional interactions rely on white matter infrastructure, it is expected that FC should largely mirror structural connectivity (van den Heuvel et al., 2009). We calculated structural-functional similarity between group-averaged matrices to estimate validity (independent of reliability) and between individual subject matrices to indicate individual measurement accuracy (combined validity and reliability).

We again start by examining pairwise correlation and partial correlation FC. The pairwise correlation FC matrices were not expected to show a high correlation with the SC matrices because pairwise correlation does not distinguish direct from indirect connectivity, producing matrices that are much less sparse. Indeed, pairwise correlation displayed low structural-functional similarity from both its group-averaged ($r = 0.244$; Figure 4A) and individual matrices (mean $r = 0.165$, SD = 0.017; Figure 4B). Partial correlation should be expected to show greater similarity because, like structural connectivity, it estimates direct connections between brain regions. As expected, the group-averaged partial correlation matrix produced a relatively high degree of structural-functional similarity ($r = 0.568$) that was significantly greater than that produced by pairwise correlation ($p < .00001$; Table 1, row 1). However, the individual partial correlation matrices displayed low structural-functional similarity (mean $r = 0.137$, SD = 0.012), even lower than those from pairwise correlation ($p < .00001$, Table 1, row 1). This supports the high group-level validity of partial correlation FC for measuring direct connections, but it also demonstrates how its low reliability can severely undermine the accuracy of individual measurements.

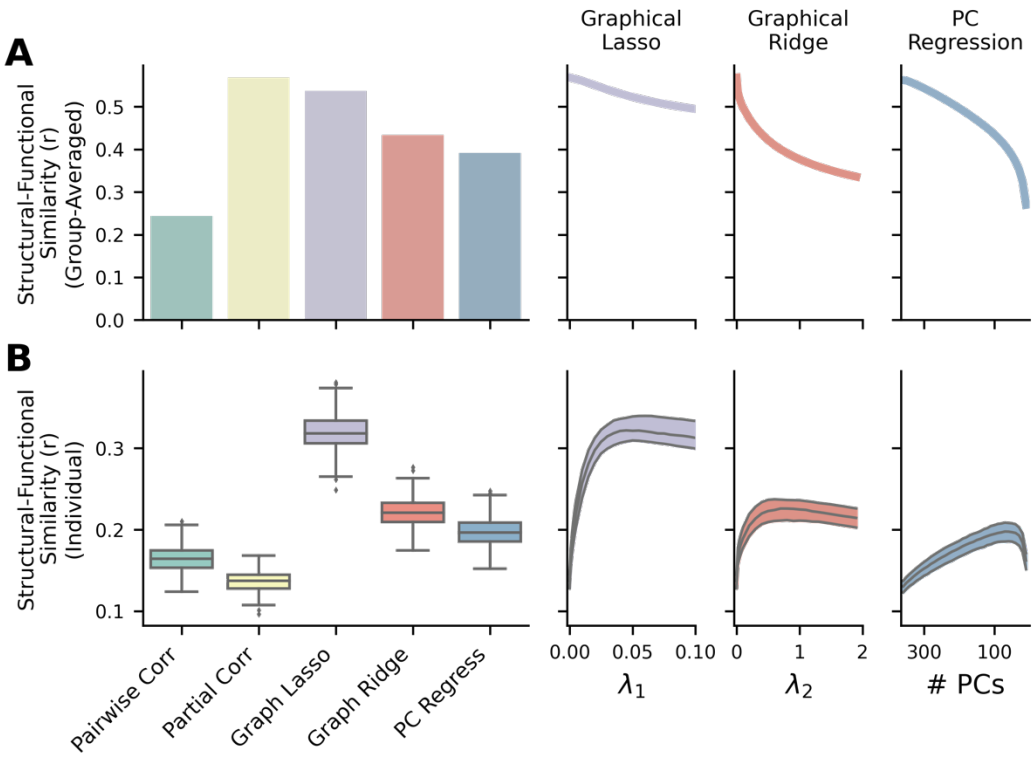

Figure 4 - Similarity between empirical FC and structural connectivity (diffusion MRI tractography). The bar and boxplots show results where the regularization hyperparameters have been optimized for each FC matrix, while the right plots show the single values or the medians and IQRs across different hyperparameter values for the regularized methods. For PC regression, number of PCs is plotted in descending order because fewer PCs correspond with more regularization. **A)** Structural-functional similarity between group-averaged FC and structural matrices (n = 1). Averaging nullifies much of the noise in individual connectivity weights to show validity without the effects of low reliability. **B)** Structural-functional similarity between individual FC matrices and the same subjects' structural matrices (n = 236). Individual measurement accuracy is vastly improved by recovering reliability through regularization.

As expected, the regularized FC methods also showed relatively high levels of structural-functional similarity when using group-averaged matrices (graphical lasso: r = 0.537; graphical ridge: r = 0.434; PC regression: r = 0.391), although unregularized partial correlation scored significantly higher than graphical ridge and PC regression (both p < .00001, Table 1, rows 6 and 7). Graphical lasso did not show a significant difference from partial correlation at the group level (p = .061, Table 1, row 5). These results indicate that some types of regularization may alter the group-level FC network structures in ways that lessen their validity. This is further illustrated by the decline in group-averaged structural-functional similarity with hyperparameters that induce greater regularization (Figure 4A, right panels). Meanwhile, the regularized FC methods all performed significantly better than basic partial correlation on individual structural-functional similarity (graphical lasso: mean r = 0.319, SD = 0.022; graphical ridge: mean r = 0.221, SD = 0.018; PC regression: mean r = 0.197, SD = 0.0177; all p < .00001, Table 1, rows 5-7). Graphical lasso again performed significantly better than graphical ridge and PC regression (both p < .0001, Table 1, rows 8 and 9). This supports our hypothesis that regularization can improve individual measurement accuracy by

enhancing reliability. These findings show graphical lasso to be especially promising, as it vastly improves the accuracy of individual FC estimates while preserving high validity of underlying (group-averaged) FC networks seen with partial correlation (Table 1, row 5). Again, these patterns were also demonstrated in held-out subjects (Supplementary Figure S2 and Supplementary Table S1), although partial correlation and graphical lasso no longer showed a significant difference in group-averaged structural-functional similarity.

Reliability and ground truth similarity of FC with simulated data

Since we do not have access to ground truth empirical FC, we also tested the validity and individual measurement accuracy of FC methods on simulated data whose true underlying connectivity we know. An example simulated network and the FC matrices estimated from a single session are shown in Figure 5. When estimating FC for the regularized methods, the mean optimal hyperparameter values were $\lambda_1 = 0.113$ for graphical lasso (SD = 0.007), $\lambda_2 = 1.00$ for graphical ridge (SD = 0.24), and number of PCs = 12.8 for PC regression (SD = 4.7). We again quantified the reliability of FC methods using between-session similarity, comparing the first two sessions' matrices for each of 100 different networks. We evaluated group-level validity using the similarity of group-averaged FC (50 sessions each) with the ground truth networks, and we assessed individual measurement accuracy as the similarity of individual FC estimates (1 session each) with the corresponding true networks.

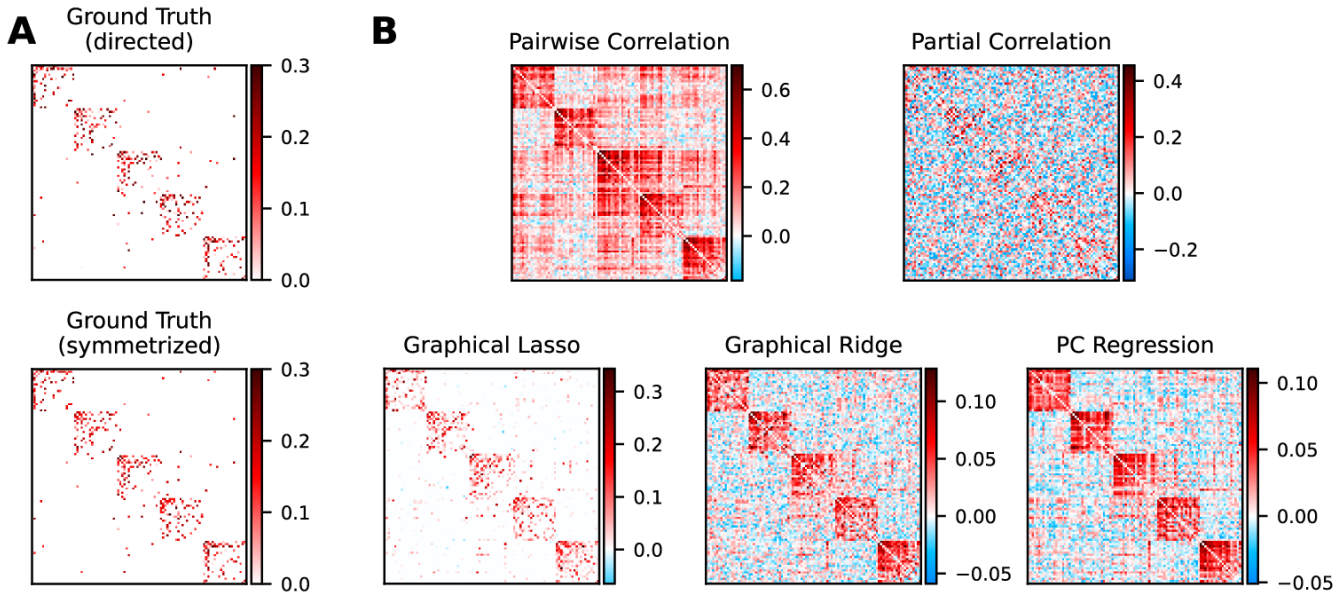

Figure 5 – Simulated networks and FC estimated by different methods. **A)** An example ground truth network. The directed, asymmetric version (top) was used to generate nodes' timeseries while the symmetrized version (bottom) was compared with FC estimates. **B)** Individual FC matrices estimated from a simulated timeseries using each FC method.

The results using simulated data largely mirror those from empirical fMRI data. In evaluating reliability, partial correlation was again shown to be the least reliable FC method (mean r = 0.160, SD = 0.028; Figure 6A), scoring considerably worse than

pairwise correlation (mean r = 0.813, SD = 0.069; p < .00001, Table 2, row 1). Again, all three of the regularized methods improved on the scores of unregularized partial correlation (graphical lasso: mean r = 0.793, SD = 0.073; graphical ridge: mean r = 0.589, SD = 0.032; PC regression: mean r = 0.580, SD = 0.069; all p < .00001, Table 2, rows 5-7). Graphical lasso performed significantly better than graphical ridge and PC regression (both p < .00001, Table 2, rows 8 and 9), and the scores of graphical ridge and PC regression were not significantly different (p = .428, Table 2, row 10). While graphical ridge and PC regression did not achieve as high reliability as pairwise correlation (p < .00001, Table 2, rows 3 and 4), graphical lasso was not significantly different from pairwise correlation when adjusting for multiple comparisons (p = .033, Table 2, row 2). In examining between-session similarity across a range of hyperparameters for the regularized methods, we can again observe that a higher degree of regularization tends to result in greater reliability (Figure 6A, right panels).

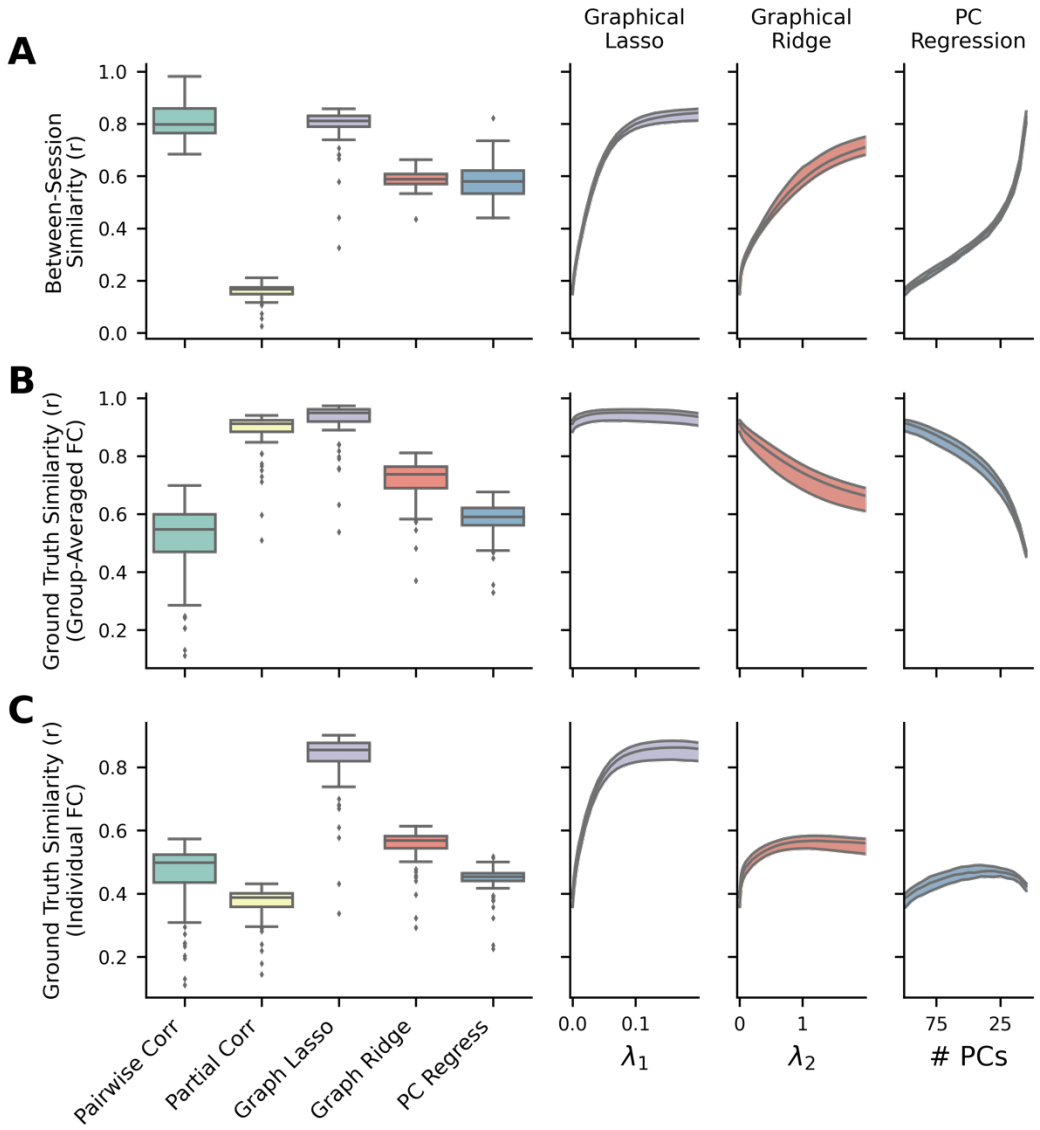

Figure 6 – Reliability and ground truth similarity of FC methods with simulated data. The boxplots show results where the regularization hyperparameters have been optimized for each FC matrix, while the right plots show the medians and IQRs across different hyperparameter values for the regularized methods. For PC regression, number of PCs is plotted in descending order because fewer PCs correspond with more regularization. **A)** Between-session similarity, calculated between one pair of session matrices for each simulated network (n = 100). **B)** Ground truth similarity between group-averaged FC matrices (50 sessions each) and the ground truth for each simulated network (n = 100). Averaging nullifies much of the noise in individual connectivity weights to show validity without the effects of low reliability. **C)** Ground truth similarity between an individual session's estimated FC matrix and the ground truth for each simulated network (n = 100). The accuracy of single measurements is vastly improved by recovering reliability through regularization.

Table 2 – Statistical tests comparing FC methods on reliability and ground truth similarity using simulated data. Between-session similarity, group-averaged ground truth similarity, and individual ground truth similarity were compared using two-tailed, dependent-sample t-tests, their Pearson's r values first normalized using Fisher's z transformation (arctanh). For all reported correlation differences, correlations were transformed to Fisher's z, subtracted, and then transformed back to Pearson's r. Alpha levels were adjusted to .005 (.05/10) following Bonferroni correction to account for the multiple comparisons within each column.

Methods		Between-session similarity (Pearson's r)	Ground truth similarity (Pearson's r)	
			Group-averaged	Individual
1	Pairwise correlation vs. partial correlation	mean Δr = 0.773, T(99) = 33.4, p < .00001	mean Δr = -0.710, T(99) = -109, p < .00001	mean Δr = 0.115, T(99) = 16.0, p < .00001
2	Pairwise correlation vs. graphical lasso	mean Δr = 0.091, T(99) = 2.16, p = 0.033	mean Δr = -0.820, T(99) = -71.7, p < .00001	mean Δr = -0.613, T(99) = -73.0, p < .00001
3	Pairwise correlation vs. graphical ridge	mean Δr = 0.472, T(99) = 17.9, p < .00001	mean Δr = -0.319, T(99) = -52.8, p < .00001	mean Δr = -0.118, T(99) = -20.0, p < .00001
4	Pairwise correlation vs. PC regression	mean Δr = 0.478, T(99) = 17.8, p < .00001	mean Δr = -0.087, T(99) = -8.97, p < .00001	mean Δr = 0.027, T(99) = 3.31, p = 0.001
5	Partial correlation vs. graphical lasso	mean Δr = -0.734, T(99) = -76.2, p < .00001	mean Δr = -0.263, T(99) = -28.6, p < .00001	mean Δr = -0.680, T(99) = -54.5, p < .00001
6	Partial correlation vs. graphical ridge	mean Δr = -0.475, T(99) = -92.8, p < .00001	mean Δr = 0.505, T(99) = 55.2, p < .00001	mean Δr = -0.229, T(99) = -116, p < .00001
7	Partial correlation vs. PC regression	mean Δr = -0.469, T(99) = -42.3, p < .00001	mean Δr = 0.664, T(99) = 54.4, p < .00001	mean Δr = -0.088, T(99) = -34.8, p < .00001
8	Graphical lasso vs. graphical ridge	mean Δr = 0.398, T(99) = 27.8, p < .00001	mean Δr = 0.678, T(99) = 44.0, p < .00001	mean Δr = 0.534, T(99) = 43.3, p < .00001
9	Graphical lasso vs. PC regression	mean Δr = 0.405, T(99) = 22.4, p < .00001	mean Δr = 0.789, T(99) = 45.5, p < .00001	mean Δr = 0.630, T(99) = 45.9, p < .00001
10	Graphical ridge vs. PC regression	mean Δr = 0.008, T(99) = 0.797, p = 0.428	mean Δr = 0.239, T(99) = 45.6, p < .00001	mean Δr = 0.144, T(99) = 49.8, p < .00001

Meanwhile, partial correlation demonstrated its worth in our test of group-level validity (mean r = 0.890, SD = 0.066; Figure 6B), scoring significantly higher than pairwise correlation (mean r = 0.517, SD = 0.122; p < .00001, Table 2, row 1). Unregularized partial correlation also scored better than graphical ridge (mean r = 0.718, SD = 0.068) and PC regression (mean r = 0.583, SD = 0.058; both p < .00001, Table 2, rows 6 and 7) for group-level ground truth similarity, although graphical lasso (mean r = 0.927, SD =0.066) displayed slightly higher scores than partial correlation (p < .00001, Table 2, row 5). Among the regularized methods, graphical lasso was

followed by graphical ridge and then PC regression (all $p < .00001$, Table 2, rows 8-10). All regularized methods in turn achieved greater validity than pairwise correlation (all $p < .00001$, Table 2, rows 2-4). In a way, regularization has an opposite impact on validity from reliability, with group-level validity generally decreasing with greater levels of regularization (Fig. 6B right panels). This effect was minimal to nonexistent for graphical lasso, however, and somewhat limited for graphical ridge and PC regression at the optimal hyperparameter values.

The accuracy of individual FC estimates depends on the underlying estimated network structures being correct (validity) and the individual estimates being minimally perturbed by noise (reliability). Partial correlation produced diminished individual ground truth similarity scores (mean $r = 0.373$, $SD = 0.048$; Figure 6C), attributable to its low reliability. These were even lower than those of pairwise correlation (mean $r = 0.464$, $SD = 0.099$; $p < .00001$, Table 2, row 1). However, in improving reliability, all regularized methods were able to increase individual measurement accuracy above that of unregularized partial correlation (graphical lasso: mean $r = 0.829$, $SD = 0.087$; graphical ridge: mean $r = 0.554$, $SD = 0.051$; PC regression: mean $r = 0.446$, $SD = 0.042$; $p < .00001$, Table 2, rows 5-7). Graphical lasso showed superior accuracy among regularized methods, which is expected from its higher reliability and validity (all $p < .00001$, Table 2, rows 8 and 9). Graphical lasso and graphical ridge also performed expectedly better than pairwise correlation (both $p < .00001$, Table 2, rows 2 and 3), but PC regression showed significantly worse accuracy than pairwise correlation ($p = .001$; Table 2, row 4). Evaluating their performances at different hyperparameter values, we can see that the optimal levels of regularization are those that balance validity and reliability (Figure 6C right panels). These findings corroborate the results of structural-functional similarity in empirical fMRI data that regularized FC methods can improve individual measurement accuracy. They also substantiate graphical lasso as the most effective of the regularized methods we tested.

Accuracy of FC with varying amounts of data and noise levels

After confirming that regularization can indeed improve the reliability and individual measurement accuracy of FC, we proceeded to test its benefits to datasets of varied quality. Two factors that can exacerbate model overfitting are low number of datapoints and high degree of noise relative to the signal of interest (Blum et al., 2020; Hastie et al., 2009; Ying, 2019), which remain limitations of fMRI data. We therefore generated 100 additional networks and simulated datasets with differing numbers of timepoints and noise levels, covering a range of low to improbably large amounts of data (1-5, 10, and 100 times as many observations as nodes) and three plausible ratios of Gaussian noise relative to the signal (0.25, 0.50, and 1.00). We then computed the FC from each dataset and calculated the ground truth similarities of individual FC estimates.

Our first observation on estimating the FC matrices was that the optimal hyperparameters for the regularized methods vary across conditions, the methods generally preferring a greater degree of regularization when the data contains more noise or fewer timepoints (Figure 7A). For graphical lasso, the average optimal hyperparameter for each condition was strongly correlated with the number of timepoints (Spearman correlation; $r_s = -0.991$, $p < .00001$) but not with noise level ($r_s =$

0.067, p = .772). For graphical ridge, optimal hyperparameter value did not have a significant monotonic relationship with number of timepoints (r_s = -0.377, p = .092) but did with noise level (r_s = 0.616, p = .003). For PC regression, optimal hyperparameter value was significantly correlated with both number of timepoints (r_s = 0.771, p = .00004) and noise level (r_s = -0.539, p = .012). This indicates a stronger relevance of regularization when datasets have higher noise or fewer timepoints. Note too that this dependence of hyperparameters on timepoints can have practical implications for model selection, necessitating that cross-validation schemes use a similar number of timepoints in training models as will be present in the final model (see Methods).

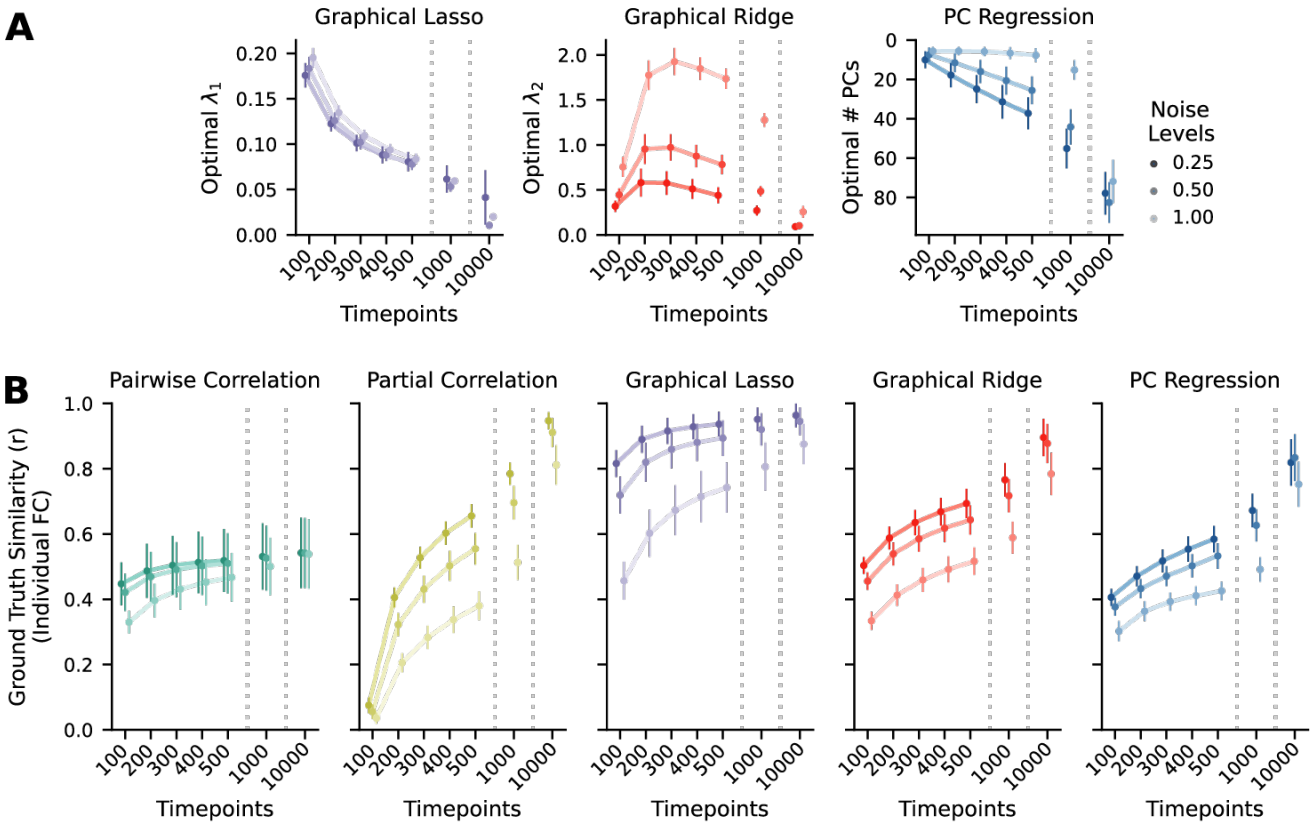

Figure 7 – Performance of FC methods when varying the amounts of data and noise levels in simulated timeseries. **A**) Optimal regularization hyperparameters of each regularized method with different numbers of timepoints (x-axis) and noise ratios (line color). Generally, a larger degree of regularization is preferred when there is more noise or fewer timepoints. For PC regression, number of PCs is plotted in descending order because fewer PCs correspond with more regularization. **B)** Ground truth similarity of individual FC estimates (n = 100) for different numbers of timepoints (x-axis) and noise ratios (line color). Plots indicate mean and standard deviation.

The extended results on ground truth similarity largely agree with the previous findings but provide greater insights into the methods and their utility to different datasets (Figure 7B; see Table 3 for statistical comparisons). We can see that pairwise correlation maintains relatively stable ground truth similarity across all conditions, showing less susceptibility to low data and high noise but plateauing at a meager level of performance. In contrast, partial correlation performs dismally in low data and high

noise conditions but does successfully recreate the networks when given sufficient data. The amounts of data needed to accomplish this, however, are not often feasible. By adding L1 regularization, graphical lasso can produce strong performances with far fewer timepoints. Its improvements over partial correlation were especially large with fewer timepoints, but it still offers marginal gains at 10000 timepoints. Graphical ridge and PC regression are also less susceptible to low data and high noise, although partial correlation surpasses them when more data is available. They also never reach the efficacy of graphical lasso, which scores significantly better than all FC methods across all conditions (all $p < .00001$, Table 3, rows 2, 5, 8, and 9). These results again lead us to recommend graphical lasso for FC estimation, for any scan length and noise level.

Table 3 – Statistical tests comparing FC methods on ground truth similarity of individual FC estimates (Pearson's r) using simulated data with varying amounts of data and noise levels. Individual ground truth similarities were compared using two-tailed, dependent-sample t-tests (99 degrees of freedom), their Pearson's r values first normalized using Fisher's z transformation (arctanh). For all reported correlation differences, correlations were transformed to Fisher's z, subtracted, and then transformed back to Pearson's r. Alpha levels were adjusted to .00024 (.05/210) following Bonferroni correction to account for the multiple comparisons across FC methods and dataset conditions.

Methods		Noise Level	Number of Timepoints						
			100	200	300	400	500	1000	10000
1	Pairwise correlation vs. partial correlation	0.25	mean Δr = 0.387 T=55.5, p<.00001	mean Δr = 0.106 T=13.3, p<.00001	mean Δr = -0.027 T=-3.15, p=.002	mean Δr = -0.124 T=-14.5, p<.00001	mean Δr = -0.201 T=-24.1, p<.00001	mean Δr = -0.431 T=-55.5, p<.00001	mean Δr = -0.838 T=-159, p<.00001
		0.50	mean Δr = 0.376 T=62.5, p<.00001	mean Δr = 0.176 T=27.9, p<.00001	mean Δr = 0.080 T=11.6, p<.00001	mean Δr = 0.005 T=0.657, p=.513	mean Δr = -0.058 T=-8.46, p<.00001	mean Δr = -0.265 T=-42.5, p<.00001	mean Δr = -0.743 T=-125, p<.00001
		1.00	mean Δr = 0.298 T=83.4, p<.00001	mean Δr = 0.209 T=55.0, p<.00001	mean Δr = 0.170 T=39.1, p<.00001	mean Δr = 0.138 T=29.2, p<.00001	mean Δr = 0.108 T=21.8, p<.00001	mean Δr = -0.013 T=-2.35, p=.021	mean Δr = -0.491 T=-106, p<.00001
2	Pairwise correlation vs. graphical lasso	0.25	mean Δr = -0.585 T=-110, p<.00001	mean Δr = -0.720 T=-116, p<.00001	mean Δr = -0.775 T=-115, p<.00001	mean Δr = -0.806 T=-111, p<.00001	mean Δr = -0.826 T=-107, p<.00001	mean Δr = -0.866 T=-96.8, p<.00001	mean Δr = -0.905 T=-77.8, p<.00001
		0.50	mean Δr = -0.432 T=-83.3, p<.00001	mean Δr = -0.580 T=-89.1, p<.00001	mean Δr = -0.651 T=-87.6, p<.00001	mean Δr = -0.695 T=-85.4, p<.00001	mean Δr = -0.723 T=-85.3, p<.00001	mean Δr = -0.785 T=-81.2, p<.00001	mean Δr = -0.850 T=-77.3, p<.00001
		1.00	mean Δr = -0.151 T=-37.6, p<.00001	mean Δr = -0.275 T=-53.8, p<.00001	mean Δr = -0.348 T=-58.2, p<.00001	mean Δr = -0.397 T=-60.2, p<.00001	mean Δr = -0.432 T=-63.7, p<.00001	mean Δr = -0.528 T=-69.4, p<.00001	mean Δr = -0.657 T=-82.7, p<.00001
3	Pairwise correlation vs. graphical ridge	0.25	mean Δr = -0.071 T=-13.7, p<.00001	mean Δr = -0.137 T=-22.3, p<.00001	mean Δr = -0.189 T=-29.4, p<.00001	mean Δr = -0.232 T=-36.3, p<.00001	mean Δr = -0.269 T=-43.8, p<.00001	mean Δr = -0.396 T=-74.8, p<.00001	mean Δr = -0.709 T=-83.2, p<.00001
		0.50	mean Δr = -0.041 T=-10.4, p<.00001	mean Δr = -0.091 T=-17.6, p<.00001	mean Δr = -0.130 T=-22.4, p<.00001	mean Δr = -0.164 T=-27.3, p<.00001	mean Δr = -0.195 T=-32.7, p<.00001	mean Δr = -0.305 T=-55.1, p<.00001	mean Δr = -0.657 T=-93.2, p<.00001
		1.00	mean Δr = -0.004 T=-4.36, p=.00003	mean Δr = -0.019 T=-8.07, p<.00001	mean Δr = -0.034 T=-10.4, p<.00001	mean Δr = -0.049 T=-12.6, p<.00001	mean Δr = -0.063 T=-14.8, p<.00001	mean Δr = -0.122 T=-23.6, p<.00001	mean Δr = -0.431 T=-87.0, p<.00001
4	Pairwise correlation vs. PC regression	0.25	mean Δr = 0.053 T=9.18, p<.00001	mean Δr = 0.025 T=3.22, p=.002	mean Δr = -0.013 T=-1.52, p=.131	mean Δr = -0.051 T=-5.72, p<.00001	mean Δr = -0.088 T=-10.1, p<.00001	mean Δr = -0.215 T=-28.6, p<.00001	mean Δr = -0.514 T=-56.5, p<.00001
		0.50	mean Δr = 0.054 T=11.4, p<.00001	mean Δr = 0.049 T=7.25, p<.00001	mean Δr = 0.030 T=3.84, p=.0002	mean Δr = 0.005 T=0.619, p=.537	mean Δr = -0.026 T=-3.16, p=.002	mean Δr = -0.145 T=-19.2, p<.00001	mean Δr = -0.553 T=-52.7, p<.00001
		1.00	mean Δr = 0.031 T=10.8, p<.00001	mean Δr = 0.040 T=9.25, p<.00001	mean Δr = 0.048 T=9.11, p<.00001	mean Δr = 0.053 T=8.78, p<.00001	mean Δr = 0.055 T=8.16, p<.00001	mean Δr = 0.017 T=2.18, p=.032	mean Δr = -0.364 T=-65.1, p<.00001
5	Partial correlation vs. graphical lasso	0.25	mean Δr = -0.793 T=-111, p<.00001	mean Δr = -0.767 T=-83.9, p<.00001	mean Δr = -0.765 T=-72.4, p<.00001	mean Δr = -0.758 T=-66.8, p<.00001	mean Δr = -0.750 T=-60.9, p<.00001	mean Δr = -0.694 T=-48.2, p<.00001	mean Δr = -0.279 T=-18.0, p<.00001
		0.50	mean Δr = -0.695 T=-87.0, p<.00001	mean Δr = -0.686 T=-73.0, p<.00001	mean Δr = -0.695 T=-65.0, p<.00001	mean Δr = -0.697 T=-60.4, p<.00001	mean Δr = -0.694 T=-57.6, p<.00001	mean Δr = -0.657 T=-48.6, p<.00001	mean Δr = -0.292 T=-28.9, p<.00001
		1.00	mean Δr = -0.430 T=-68.6, p<.00001	mean Δr = -0.458 T=-62.5, p<.00001	mean Δr = -0.489 T=-59.0, p<.00001	mean Δr = -0.507 T=-56.4, p<.00001	mean Δr = -0.517 T=-55.0, p<.00001	mean Δr = -0.519 T=-49.8, p<.00001	mean Δr = -0.245 T=-33.3, p<.00001
6	Partial correlation vs. graphical ridge	0.25	mean Δr = -0.446 T=-152, p<.00001	mean Δr = -0.240 T=-98.2, p<.00001	mean Δr = -0.163 T=-64.2, p<.00001	mean Δr = -0.111 T=-38.1, p<.00001	mean Δr = -0.072 T=-23.3, p<.00001	mean Δr = 0.042 T=9.29, p<.00001	mean Δr = 0.319 T=32.9, p<.00001
		0.50	mean Δr = -0.411 T=-138, p<.00001	mean Δr = -0.263 T=-152, p<.00001	mean Δr = -0.207 T=-136, p<.00001	mean Δr = -0.168 T=-114, p<.00001	mean Δr = -0.139 T=-97.7, p<.00001	mean Δr = -0.043 T=-25.9, p<.00001	mean Δr = 0.167 T=59.9, p<.00001
		1.00	mean Δr = -0.302 T=-104, p<.00001	mean Δr = -0.227 T=-113, p<.00001	mean Δr = -0.203 T=-128, p<.00001	mean Δr = -0.186 T=-139, p<.00001	mean Δr = -0.170 T=-150, p<.00001	mean Δr = -0.109 T=-137, p<.00001	mean Δr = 0.076 T=52.6, p<.00001
7	Partial correlation vs. PC regression	0.25	mean Δr = -0.341 T=-117, p<.00001	mean Δr = -0.081 T=-28.1, p<.00001	mean Δr = 0.014 T=4.75, p<.00001	mean Δr = 0.074 T=22.0, p<.00001	mean Δr = 0.115 T=37.1, p<.00001	mean Δr = 0.238 T=59.6, p<.00001	mean Δr = 0.569 T=58.6, p<.00001
		0.50	mean Δr = -0.329 T=-104, p<.00001	mean Δr = -0.128 T=-46.9, p<.00001	mean Δr = -0.050 T=-17.8, p<.00001	mean Δr = 0.001 T=0.190, p=.850	mean Δr = 0.032 T=10.7, p<.00001	mean Δr = 0.125 T=37.9, p<.00001	mean Δr = 0.323 T=35.5, p<.00001
		1.00	mean Δr = -0.270 T=-78.5, p<.00001	mean Δr = -0.171 T=-49.2, p<.00001	mean Δr = -0.123 T=-38.8, p<.00001	mean Δr = -0.085 T=-26.4, p<.00001	mean Δr = -0.054 T=-16.5, p<.00001	mean Δr = 0.030 T=8.11, p<.00001	mean Δr = 0.155 T=43.2, p<.00001
8	Graphical lasso vs. graphical ridge	0.25	mean Δr = 0.536 T=77.8, p<.00001	mean Δr = 0.646 T=72.0, p<.00001	mean Δr = 0.687 T=68.4, p<.00001	mean Δr = 0.706 T=66.9, p<.00001	mean Δr = 0.717 T=64.2, p<.00001	mean Δr = 0.715 T=61.0, p<.00001	mean Δr = 0.549 T=54.6, p<.00001
		0.50	mean Δr = 0.398 T=55.2, p<.00001	mean Δr = 0.516 T=53.9, p<.00001	mean Δr = 0.570 T=52.0, p<.00001	mean Δr = 0.599 T=50.6, p<.00001	mean Δr = 0.615 T=50.1, p<.00001	mean Δr = 0.632 T=47.6, p<.00001	mean Δr = 0.438 T=48.9, p<.00001
		1.00	mean Δr = 0.147 T=32.4, p<.00001	mean Δr = 0.257 T=38.0, p<.00001	mean Δr = 0.318 T=38.9, p<.00001	mean Δr = 0.355 T=39.2, p<.00001	mean Δr = 0.380 T=40.0, p<.00001	mean Δr = 0.434 T=40.7, p<.00001	mean Δr = 0.316 T=43.0, p<.00001
9	Graphical lasso vs. PC regression	0.25	mean Δr = 0.619 T=82.3, p<.00001	mean Δr = 0.732 T=74.6, p<.00001	mean Δr = 0.770 T=70.4, p<.00001	mean Δr = 0.787 T=67.6, p<.00001	mean Δr = 0.796 T=65.2, p<.00001	mean Δr = 0.800 T=62.1, p<.00001	mean Δr = 0.732 T=52.2, p<.00001
		0.50	mean Δr = 0.475 T=59.1, p<.00001	mean Δr = 0.612 T=57.0, p<.00001	mean Δr = 0.668 T=54.6, p<.00001	mean Δr = 0.698 T=52.9, p<.00001	mean Δr = 0.710 T=53.2, p<.00001	mean Δr = 0.723 T=50.7, p<.00001	mean Δr = 0.561 T=48.9, p<.00001

		1.00	mean Δr = 0.181 T=30.3, p<.00001	mean Δr = 0.311 T=36.3, p<.00001	mean Δr = 0.390 T=38.1, p<.00001	mean Δr = 0.441 T=39.0, p<.00001	mean Δr = 0.476 T=40.1, p<.00001	mean Δr = 0.540 T=41.9, p<.00001	mean Δr = 0.386 T=47.0, p<.00001
10	Graphical ridge vs. PC regression	0.25	mean Δr = 0.123 T=67.6, p<.00001	mean Δr = 0.162 T=62.1, p<.00001	mean Δr = 0.176 T=59.9, p<.00001	mean Δr = 0.183 T=55.1, p<.00001	mean Δr = 0.186 T=56.1, p<.00001	mean Δr = 0.198 T=54.6, p<.00001	mean Δr = 0.306 T=38.0, p<.00001
		0.50	mean Δr = 0.095 T=53.8, p<.00001	mean Δr = 0.139 T=56.6, p<.00001	mean Δr = 0.159 T=55.3, p<.00001	mean Δr = 0.169 T=51.4, p<.00001	mean Δr = 0.170 T=57.5, p<.00001	mean Δr = 0.167 T=55.6, p<.00001	mean Δr = 0.164 T=20.7, p<.00001
		1.00	mean Δr = 0.035 T=14.9, p<.00001	mean Δr = 0.059 T=22.2, p<.00001	mean Δr = 0.082 T=30.1, p<.00001	mean Δr = 0.102 T=33.8, p<.00001	mean Δr = 0.118 T=36.4, p<.00001	mean Δr = 0.138 T=39.8, p<.00001	mean Δr = 0.080 T=24.0, p<.00001

Sensitivity of FC to motion artifacts

In addition to the random measurement noise whose impact we tested above, fMRI data is also contaminated with subject motion artifacts (Power et al., 2012). Previous studies have shown partial correlation to effectively mitigate the confounding effects of motion (Mahadevan et al., 2021; Power et al., 2015), but given the pervasiveness of the issue, we sought to also assess the regularized methods tested here. We did so by measuring the QC-FC correlation of each edge, or the correlation across subjects between their mean head motion during scanning and their estimated connectivity weights for each edge of interest (Ciric et al., 2017; Mahadevan et al., 2021).

Our findings replicate those of previous studies, showing partial correlation to be largely robust to motion artifacts with 0% of edges significantly correlated with motion (median absolute QC-FC = 0.045). Pairwise correlation was very susceptible to motion artifacts with 51.1% significant edges (median absolute QC-FC = 0.148), also in agreement with prior findings. Graphical lasso (0.03% significant edges, median absolute QC-FC = 0.048), graphical ridge (0.03% significant edges, median absolute QC-FC = 0.051), and PC regression (0.02% significant edges, median absolute QC-FC = 0.051) were impacted slightly more than unregularized partial correlation, but nonetheless they were also largely insensitive to motion. In case sparsity alone was providing some FC methods with an advantage, we repeated the analyses while only assessing nonzero edges (see Methods). The results showed minimal changes, with pairwise correlation having 50.2% significant edges (n = 64619 edges included in the analysis, median absolute QC-FC = 0.149), partial correlation 0% significant edges (n = 59985, median absolute QC-FC = 0.049), graphical lasso 1.6% significant edges (n = 705, median absolute QC-FC = 0.055), graphical ridge 1.2% significant edges (n = 2837, median absolute QC-FC = 0.062), and PC regression 0.8% significant edges (n = 3076, median absolute QC-FC = 0.061). Their insensitivity to motion artifacts further supports the suitability of regularized partial correlation or similar methods for FC estimation.

Predicting task activations from empirical FC

The remaining analyses further test the accuracy of regularized multivariate FC in representing brain networks and test their advantage for FC-based neuroscience applications. For simplicity, we limited our main results to pairwise correlation, partial correlation, and graphical lasso FC, but the results of graphical ridge and PC regression are presented in Supplementary Figures S3 and S4 and Supplementary Tables S2 and S3, with graphical ridge and PC regression achieving similar performances to graphical lasso.

Here we implemented activity flow modeling to predict held-out neurocognitive function (task activations) from FC matrices (Cole et al., 2016). This model formalizes a functional role for FC, and in this application conveys how well different FC methods

captured communication pathways underlying activity flow in the brain. To perform activity flow mapping, the activity of a target node *j* is predicted as the sum of activity from all other source nodes *i*, weighted by the FC coefficients between sources and target (Figure 8A; see Methods). We applied this to the "hit" and "correct rejection" events of a Go/No-go task, and we present the actual and predicted activations for a single subject's "hit" trials in Figure 8D. The prediction accuracies were calculated for each subject as the R^2 score or Pearson correlation between the predicted and actual activities.

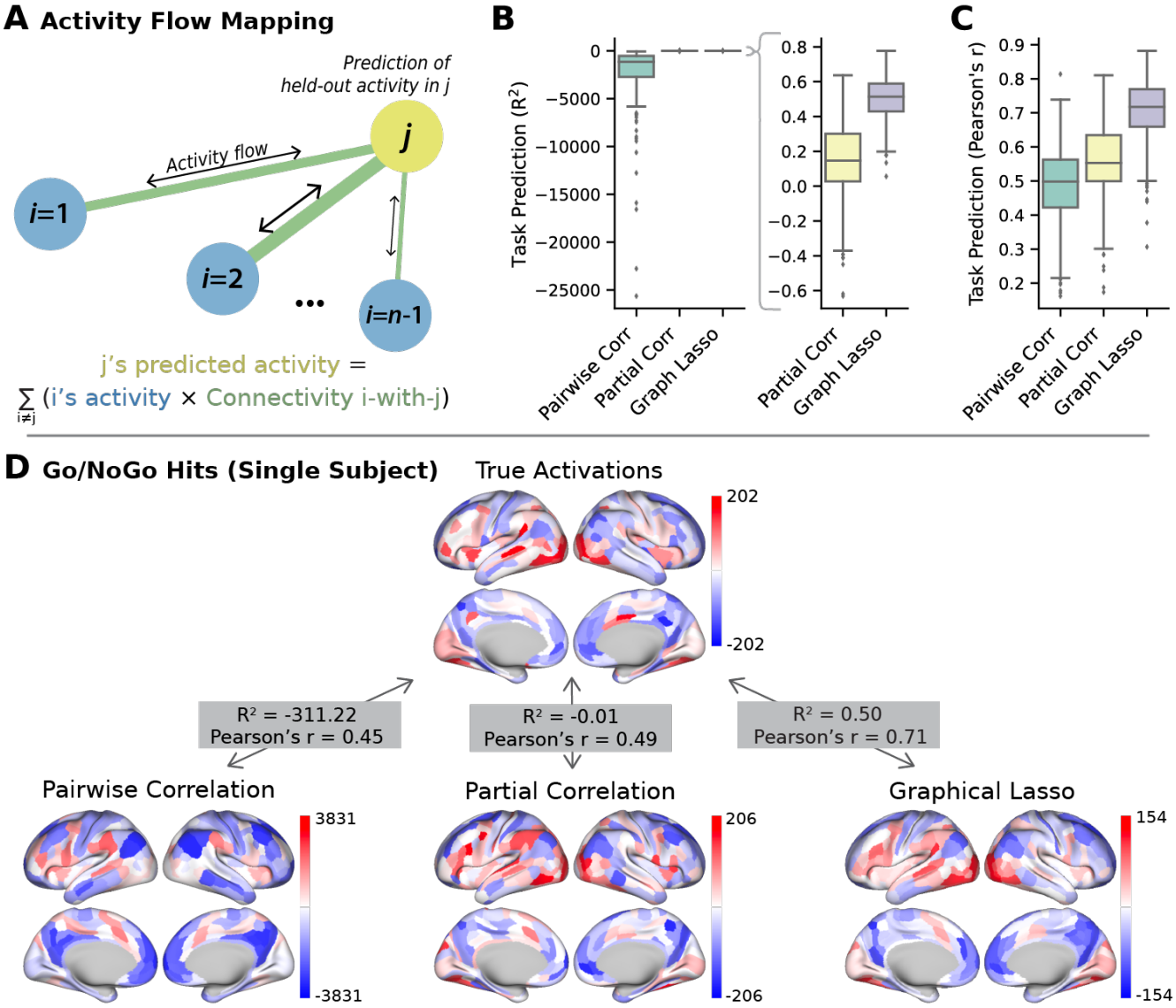

Figure 8 – Predicting task activations using FC estimated from rest fMRI data. **A)** The activity flow mapping paradigm (Cole et al., 2016). Each target region's (*j*) activity is predicted as the sum of all other regions' (*i*) activity flows to the target region, which are the products of each source region's connectivity with the target and the source region's actual activity. **B)** Prediction accuracy by FC method calculated as the coefficient of determination (R^2) between regions' actual and predicted activations, computed in each subject (n = 236) across regions and task conditions. Partial correlation and graphical lasso results are expanded in a separate panel. **C)** Prediction accuracy calculated as Pearson correlation (r), which unlike R^2 is scale-invariant. **D)** A single subject's actual and predicted task activations for the go/no-go hit events.

Pairwise correlation FC generated large negative R^2 scores for task prediction (mean R^2 = -2228, SD = 3180; Figure 8B), meaning that its predicted activations were very different from the actual activations. However, this does not necessarily mean that the pattern of activations across nodes and conditions was incorrect. Rather, since R^2 is a measure of unscaled similarity, these low scores were primarily driven by the large size difference (as opposed to pattern difference) between predicted and actual values (see Figure 8D color bars). Because pairwise correlation FC contains many higher weighted edges, it predicted activations with much larger scales than were actually measured. Scoring prediction accuracy with Pearson correlation instead of R^2 aided the performance of pairwise correlation FC (mean r = 0.490, SD = 0.119; Figure 8C), although still it remained below the other FC methods. Diverging from the previous tests of FC accuracy, partial correlation produced significantly higher scores than pairwise correlation for R^2 (mean R^2 = 0.140, SD = 0.220; T(235) = 10.74, p < .00001) as well as Pearson's r (mean r = 0.551, SD = 0.113; T(235) = 12.31, p < .00001; all tests two-tailed, dependent-samples; α = .005 (.05/10) for all tests following Bonferroni correction for multiple comparisons). This performance metric appears to be less impacted by low reliability than previous tests, likely because some effects of noisy edges can cancel during the summation of activity flows. These results again support the higher validity of partial correlation than pairwise correlation FC, as the network created by partial correlation better reflects the integration of brain activity over nodes. Activity prediction was further improved by adding regularization through graphical lasso (mean R^2 = 0.506, SD = 0.123; mean r = 0.708, SD = 0.091), which performed significantly better than both pairwise correlation (R^2: T(235) = 10.74, p < .00001; Pearson's r: T(235) = 51.61, p < .00001) and partial correlation (R^2: T(235) = 49.54, p < .00001; Pearson's r: T(235) = 81.15, p < .00001; α = .005). This pattern of results was reproduced in the replication dataset as well (Supplementary Figure S3 and Supplementary Table S2. Graphical lasso was previously shown to increase reliability while preserving the validity of the underlying estimated network structure, allowing for greater accuracy of individual FC estimates. This improved accuracy allowed for better simulations of activity flowing across brain networks.

Predicting subject age and intelligence from empirical FC

An increasingly common goal in neuroscience research is to link individual differences in brain structure or function with cognitive or other biological characteristics. This can help reveal the functional relevance of brain network organization. However, individual difference studies have been shown to have small effect sizes (Marek et al., 2022), and they therefore may benefit from regularized FC methods that can reduce the

impact of measurement noise. If edge weights experience high noise variability compared to true individual variability, then any real individual difference patterns will be obscured, and results will be too easily swayed by random chance. We therefore hypothesized that using more reliable FC estimates would lead to better predictions of individual differences. We tested this by using FC edge weights to predict subject age and intelligence (psychometric *g*; Johnson et al., 2008), comparing prediction accuracies across FC methods. These tests demonstrate the utility of regularization to this research approach while providing additional validation of these methods, confirming their ability to capture the patterns of brain function that underly cognitive differences.

We did indeed find that the more reliable FC methods, pairwise correlation and graphical lasso, predicted individual differences better than the less reliable method, partial correlation. All methods allowed for above-chance prediction of subject age (pairwise correlation: $r = 0.710$, $p < .00001$; partial correlation: $r = 0.544$, $p < .00001$; graphical lasso: $r = 0.694$, $p < .00001$; $\alpha = .005$ (.05/10) for all tests following Bonferroni correction for multiple comparisons; Figure 9A). Pairwise correlation and graphical lasso both produced significantly higher predicted-to-actual age correlations than partial correlation (pairwise vs. partial correlation: $z = 4.162$, $p = .00003$; graphical lasso vs. partial correlation: $z = 5.211$, $p < .00001$; $\alpha = .005$ (.05/10) for all tests following Bonferroni correction for multiple comparisons), but their correlations were not significantly different from each other (pairwise correlation vs. graphical lasso: $z = 0.576$, $p = .564$; statistical tests described by Meng et al., 1992). These results were reproduced in the replication dataset (pairwise correlation: $r = 0.726$, $p < .00001$; partial correlation: $r = 0.619$, $p < .00001$; graphical lasso: $r = 0.740$, $p < .00001$; pairwise vs. partial correlation: $z = 3.010$, $p = .0026$; graphical lasso vs. partial correlation: $z = 5.194$, $p < .00001$; pairwise correlation vs. graphical lasso: $z = -0.582$, $p = .561$; $\alpha = .005$).

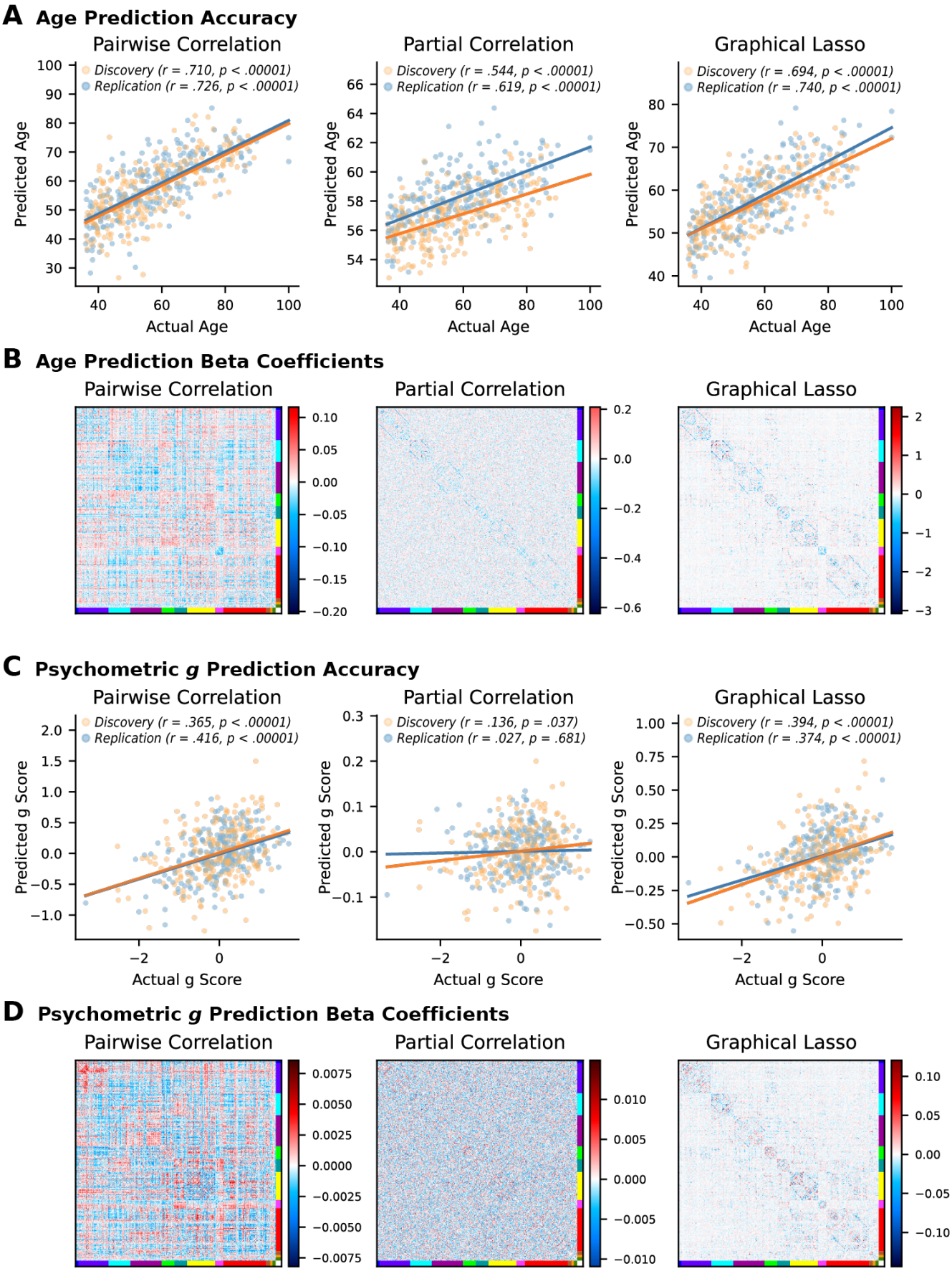
A Age Prediction Accuracy
Pairwise Correlation
Discovery (r = .710, p < .00001)
Replication (r = .726, p < .00001)
Partial Correlation
Discovery (r = .544, p < .00001)
Replication (r = .619, p < .00001)
Graphical Lasso
Discovery (r = .694, p < .00001)
Replication (r = .740, p < .00001)
Predicted Age
Actual Age
B Age Prediction Beta Coefficients
Pairwise Correlation
Partial Correlation
Graphical Lasso
C Psychometric g Prediction Accuracy
Pairwise Correlation
Discovery (r = .365, p < .00001)
Replication (r = .416, p < .00001)
Partial Correlation
Discovery (r = .136, p = .037)
Replication (r = .027, p = .681)
Graphical Lasso
Discovery (r = .394, p < .00001)
Replication (r = .374, p < .00001)
Predicted g Score
Actual g Score
D Psychometric g Prediction Beta Coefficients
Pairwise Correlation
Partial Correlation
Graphical Lasso

Figure 9 – Predicting individual differences in age and intelligence using estimated FC from rest fMRI data. **A)** Actual and predicted ages of each subject by FC method, from both the discovery and replication datasets. **B)** Average beta coefficients assigned to each connection by the regression models for estimating subject age. Beta coefficients shown here are the averages over cross-validation folds and discovery and replication datasets. Blue indicates that connection strength decreased with age and red indicates that connection strength increased. **C)** Actual and predicted intelligence (psychometric *g*) of each subject by FC method. **D)** Average beta coefficients assigned to each connection by the regression models for estimating psychometric *g*.

The same pattern was observed when predicting intelligence from FC, although prediction accuracies were generally lower than for age, and partial correlation no longer showed a significantly above-chance prediction accuracy (pairwise correlation: $r = 0.365$, $p < .00001$; partial correlation: $r = 0.136$, $p = .037$; graphical lasso: $r = 0.394$, $p < .00001$; $\alpha = .017$ (.05/3) for all tests following Bonferroni correction for multiple comparisons). Again, pairwise correlation and graphical lasso produced significantly better prediction accuracies than partial correlation (pairwise vs. partial correlation: $z = 3.030$, $p = .0024$; graphical lasso vs. partial correlation: $z = 3.906$, $p = .00009$; $\alpha = .017$ (.05/3) for all tests following Bonferroni correction for multiple comparisons), but again their performances were not significantly different from each other (pairwise correlation vs. graphical lasso: $z = -0.505$, $p = .613$). The same occurred in the replication dataset (pairwise correlation: $r = 0.416$, $p < .00001$; partial correlation: $r = 0.027$, $p = .681$; graphical lasso: $r = 0.374$, $p < .00001$; pairwise vs. partial correlation: $z = 4.779$, $p < .00001$; graphical lasso vs. partial correlation: $z = 5.155$, $p < .00001$; pairwise correlation vs. graphical lasso: $z = 0.718$, $p = .473$; $\alpha = .017$). Together, these results indicate that regularization improves the ability of partial correlation FC to reflect individual differences in cognition, and they provide additional evidence for graphical lasso FC as accurately representing brain connectivity. Further, these results indicate the importance of the increased reliability of regularized partial correlation for matching the behavioral prediction accuracy of pairwise correlation, but now with more interpretable (i.e., valid) connectivity due to the substantial reduction in the number of confounded functional connections.

Discussion

Accurate estimation of human brain connectivity is a major goal of neuroscience, given evidence that connectivity is a major determinant of neurocognitive function (Bassett & Sporns, 2017). The current field standard approach for FC estimation – pairwise Pearson correlation – is well known to be sensitive to false positive connectivity, resulting in confounded and indirect connections that often do not reflect the brain's true underlying connectivity (Honey et al., 2009; Reid et al., 2019; Sanchez-Romero & Cole, 2021). It appears that the field would benefit from changing to a method that removes many of these false connections, such as partial correlation. However, several reports have demonstrated that partial correlation FC has much lower reliability than pairwise FC (Fiecas et al., 2013; Mahadevan et al., 2021), substantially reducing its utility. Here we tested the hypothesis that low reliability in partial correlation (and related methods) is due to excessive model complexity resulting in overfitting to noise. Consistent with our hypothesis, we found that multiple regularization techniques

significantly improve the repeat reliability and accuracy of partial correlation FC. These results suggest regularized partial correlation is a strong candidate for replacing pairwise correlation as the next field standard for estimating FC using fMRI. These results also suggest that FC methods used with high temporal resolution data (such as electroencephalography) could benefit from regularization as well, especially given the additional model complexity that comes with modeling temporal lags (e.g., with multivariate autoregressive modeling; Fiecas et al., 2010; Haufe et al., 2010; Mill et al., 2022).

We found that regularization improved FC estimation along multiple dimensions. First, reliability was substantially improved (e.g., mean r = 0.68 for graphical lasso between-session similarity) relative to unregularized partial correlation (e.g., mean r = 0.28; Figure 3). This result essentially "rescues" partial correlation as a method, given how essential reliability is to FC estimation (Noble et al., 2021). This was especially beneficial in conditions of low data and high noise, which can further impair reliability (Birn et al., 2013). Second, the validity of FC estimates was much higher for regularized partial correlation relative to pairwise correlation. This was demonstrated in terms of increased similarity of regularized partial correlation FC to empirical structural connectivity (Figure 4) and simulated ground truth FC (Figure 6). Notably, this likely reflects partial correlation reducing the number of confounded (e.g., a false A-C connection due to A←B→C) and indirect (e.g., a false A-C connection due to A→B→C) connections (Figure 1A), which structural connectivity and ground truth FC are insensitive to. Further supporting validity, regularized partial correlation FC was found to be less sensitive to in-scanner motion than pairwise correlation FC, demonstrating reduced confounding from non-neural factors as well. Finally, given the importance of reliability and validity for a variety of FC applications, we tested for improvements to FC-based applications in cognitive neuroscience. As expected, we found that regularized partial correlation significantly improved FC-based predictions of task-evoked fMRI activations (Figure 8) and FC-based predictions of individual differences in behavior (Figure 9). Together, these results demonstrate the improved reliability, validity, and general utility of regularized partial correlation – especially graphical lasso – relative to pairwise correlation and unregularized partial correlation.

The central finding of this study is that regularization stabilizes the connectivity estimates of partial correlation (or multiple regression) FC, increasing the method's reliability. We demonstrated this in multiple ways. First, we provided a visual depiction of measurement instability in FC matrices, where an individual subject's partial correlation FC varied discernibly from the group-averaged matrix (Figure 2). In the individual matrix, random noise obscured an underlying network structure, while in the group matrix, that noise had been reduced by averaging. Regularization reduced this noise at the individual subject level. We then quantified repeat reliability in empirical data using between-session similarity and intraclass correlation, where all regularized methods substantially improved on the reliability of partial correlation FC.

Our findings also agree with those of several previous studies. Brier et al. (2015) showed that Ledoit-Wolf shrinkage (Ledoit & Wolf, 2003), a regularization technique not investigated here, improved the reliability of partial covariance FC. Mejia et al. (2018) also demonstrated that L_2-regularized partial correlation achieved greater reliability as the L_2 penalty (degree of regularization) was increased. Our results somewhat contrast

with those of Fiecas et al. (2013) and Mahadevan et al. (2021), who tested regularized partial correlation, in that their results indicate a more subtle impact of regularization on reliability. We suspect, however, that this apparent small difference in regularized and unregularized partial correlation is due to their measuring reliability using intraclass correlation for all possible edges. We found that null edges achieve poor ICC scores regardless of their reliability (see Methods for elaboration), meaning that the abundance of invariably low scores from these sparse FC methods would have eclipsed any meaningful changes in ICCs for non-null edges. Our analysis using intraclass correlation, which instead examined a conservative set of non-null edges, showed a clear improvement of ICCs with regularization. In utilizing several regularization methods, two reliability metrics, and empirical and simulated datasets, our study provides comprehensive support that regularization can substantially improve the repeat reliability of multivariate FC methods which are otherwise prone to overfitting to noise.

We also analyzed the validity and individual measurement accuracy of the FC methods, which we define as the systematic correctness of estimates in aggregate (independent of reliability) and the closeness of individual estimates to the truth (dependent on validity and reliability). Previous studies have tested regularized FC methods based on simulations (Nie et al., 2017; Ryali et al., 2012; Smith et al., 2011), generalization to held-out data (Varoquaux et al., 2010), network modularity (Brier et al., 2015; Mahadevan et al., 2021; Ryali et al., 2012; Varoquaux et al., 2010), sensitivity to task state (Brier et al., 2015; Duff et al., 2013; Sala-Llonch et al., 2019), prediction of individual differences (Pervaiz et al., 2020), and similarity to structural connectivity (Liégeois et al., 2020). We chose to validate FC with structural connectivity and simulated networks because these both (attempt to) represent direct connections only – a property which allows FC to better reflect the causal mechanisms and physical reality of brain networks (Reid et al., 2019). Structural connectivity estimates direct connections by mapping white matter tracts across the brain volume to their grey matter endpoints (Le Bihan & Johansen-Berg, 2012). These anatomical connections are the basis for brain-wide communication and are expected to exist between regions with direct functional connections. Structural connectivity does not measure precisely the same phenomena as FC, as structural connectivity shows predominantly static pathways while FC is also influenced by dynamic factors such as cognitive state (Cole et al., 2014). Structural connectivity estimated from diffusion MRI is also prone to systematic errors, such as from failing to resolve crossing fibers or underestimating long-distance tracts (Maier-Hein et al., 2017; Rheault et al., 2020; Sotiropoulos & Zalesky, 2019). Nevertheless, the two measures largely overlap (Damoiseaux & Greicius, 2009; Honey et al., 2009; Straathof et al., 2019; van den Heuvel et al., 2009). Because structural connectivity is not susceptible to confounded and indirect connectivity, it should better converge with the FC whose methods limit these errors (Damoiseaux & Greicius, 2009; Honey et al., 2009; Liégeois et al., 2020). For instance, Liégeois et al. (2020) have shown partial correlation FC to be more similar to structural connectivity than is pairwise correlation FC.

We also tested the validity and accuracy of the FC methods using simulations, which allow estimated FC to be compared with the true causal connections. However, simulations also require the selection of many model parameters, the choices of which can potentially bias performance among FC methods. We opted for a simple linear

model to simulate activity over bidirectional networks, which were generated with modular and approximately scale-free structure – properties the human brain has been shown to have (Bullmore & Sporns, 2009; van den Heuvel & Sporns, 2011). As with structural connectivity, prior simulation studies have shown partial correlation FC to emulate true connectivity more closely than pairwise correlation FC (Nie et al., 2017; Smith et al., 2011). We thus validated the FC methods by comparing them against approximate (structural) or certain (simulated) causal networks to further increase confidence in our results.

One aim of these analyses was to ensure that the regularization techniques did not reduce the underlying validity of FC measures, estimated as the correctness of the group-level means. We hypothesized that unregularized partial correlation FC would exhibit high validity, and indeed, after averaging across estimates, it performed substantially better than pairwise correlation FC at reflecting structural and simulated network connectivity. We found graphical lasso FC to perform similarly to partial correlation, scoring slightly worse when tested against structural connectivity but slightly better when tested with simulations. As regularization was only hypothesized to aid reliability, this was the expected result for analyses of group-level FC where the effect of reliability had been controlled via cross-subject averaging. Unexpectedly, however, we found graphical ridge and PC regression to have significantly worse validity, a pattern that emerged for both structural connectivity and simulation analyses. This result also aligns with the FC matrix visualizations (Figure 2), where graphical ridge and PC regression FC show visibly different group-level network structures relative to partial correlation and graphical lasso FC. Evidently, the more diffuse connectivity patterns produced by graphical ridge and PC regression are less valid representations of direct FC than the sparser alternatives. To our knowledge, this is the only published study to examine the effect of these regularization techniques on group-level validity, controlling for the benefit of improved reliability. Our study suggests the novel finding that graphical ridge and PC regression may reduce the group-level validity of FC estimates while graphical lasso can better maintain the validity achieved by unregularized partial correlation.

Our second aim was to determine whether the improved reliability of the regularized FC methods could grant them greater individual measurement accuracy than pairwise correlation and unregularized partial correlation FC. Individual measurement accuracy is a function of both reliability and the underlying validity of the measurement (see Figure 1). We found that the individual estimates of pairwise correlation FC are hindered by its low validity while the scores of partial correlation FC are hindered by its low reliability. The regularized methods generally perform better, led by graphical lasso FC, which most increased reliability and best preserved underlying measurement validity.

The results were largely consistent between the structural connectivity and simulation analyses, and they agree with results from prior studies. Liégeois et al. (2020) also compared FC estimates with structural connectivity and showed graphical ridge to be more similar than both partial correlation and pairwise correlation FC. Smith et al. (2011) and Nie et al. (2017) showed graphical lasso to recreate simulated networks more accurately than pairwise correlation FC, although there was not always a large difference between graphical lasso and unregularized partial correlation. We

suspect that this was due to the simulations using very few nodes (usually 5 or 10) relative to the amount of data (usually 200 TRs or more), causing less overfitting to noise than should be expected from an empirical dataset. Regularization had a much larger benefit during the simulation with 50 nodes, where partial correlation was impaired by a more realistic degree of overfitting. Varoquaux et al. (2010) tested the generalizability (via loglikelihood) of FC models to held-out rest data, showing graphical lasso and graphical ridge to both perform better than partial correlation when applied to individual subjects' data. They also found graphical lasso to exhibit higher generalizability than graphical ridge, further supporting our finding that graphical lasso is the more valid regularized FC method.

Limitations and directions for future research

The present study has several limitations, each of which suggest important directions for future research. For example, the present results are based only on empirical and simulated fMRI data, and only at the scale of brain regions. This suggests future work could test our hypotheses regarding the utility of regularized partial correlation or other multivariate methods for estimating FC with different types of brain recordings. For example, regularization could be useful for estimating multivariate autoregressive models with MEG/EEG data. Indeed, supporting the possibility that our hypotheses generalize to EEG data, Mill et al. (2022) found some evidence that PC regression improved the stability of multivariate autoregressive models. Additionally, Haufe et al. (2010) show that a group lasso penalty can also stabilize multivariate autoregressive models, and Fiecas et al. (2010) found that shrinkage can stabilize partial coherence estimates from EEG data.

Perhaps because we did not set out to promote any single regularization approach, we tested more forms of regularization than most prior studies investigating regularization for FC estimation. However, many other forms of regularization have been proposed that we did not test (Mejia et al., 2018; Nie et al., 2017; Ryali et al., 2012; Varoquaux et al., 2010). One promising form of regularization to test in future work will be elastic net (Zou & Hastie, 2005), which combines both L_1 and L_2 regularization with potential benefits of both (Ryali et al., 2012). For instance, elastic net is purported to induce sparsity while allowing more sharing of weights between similar variables (Zou & Hastie, 2005). Another adaptation of interest to us is group regularization, where data is pooled across subjects so that the larger quantity of data can improve model fit. Typically, these will estimate the underlying, group-level network structure and the individual variations of each subject from that baseline (Mejia et al., 2018; Varoquaux et al., 2010). Some methods also offer the advantage of not requiring hyperparameter selection, making them easier for researchers to apply (Brier et al., 2015; Ledoit & Wolf, 2003; Nie et al., 2017). For regularization methods that do require hyperparameters, there are various additional model selection methods not tested here that may improve FC accuracy or computational efficiency. We chose to use cross-validated R^2, but other options include cross-validated loglikelihood (Friedman et al., 2008), extended Bayesian information criterion (Foygel & Drton, 2010), and *Dens* criterion based on sparsity (Wang et al., 2016). Many other forms of FC remain to be tested with regularization as well, given that regularization is applicable to any form of data fitting (e.g., artificial neural network learning) and not only partial correlation and

multiple regression. A promising direction for future work is to utilize regularization to counter the excessive flexibility of nonlinear FC methods, reducing the tendency for such methods to overfit to noise.

We used simulations to further test the utility of regularization across a variety of situations based on ground truth connectivity. These simulations were necessarily limited to only a subset of biological details, given our focus on fMRI data and connectivity between brain regions (as opposed to voxels or neurons). This provides an opportunity for future studies to conduct more comprehensive simulations with more biological detail, multiple scales simulated (e.g., voxels and brain regions), and with different underlying network architectures. These studies could reveal more general conclusions about the utility of each form of regularization as a function of the specific data analysis scenario.

The complexity of regularized partial correlation FC contrasts with many neuroscientists' preference for reduced FC measure complexity – to minimize assumptions, minimize compute time, and fully comprehend and easily communicate methodological details. However, we showed that the lower complexity of the field-standard Pearson correlation FC method comes with lower validity of the estimated connections. Further, while more compute time is needed for regularized partial correlation (for single instances and for repetitions over different regularization hyperparameters), recent increases in the availability of multi-core processors makes the increased compute time much less severe than just a few years ago. This also opens up an opportunity for future research to reduce processing time by identifying standard regularization parameters for common situations, such as when a particular fMRI sequence, scan duration, and region set are used. It is also possible (given the low variability of hyperparameters across subjects in our study) that compute time could be reduced by using the optimal hyperparameters from a small subset of subjects as the basis for hyperparameters for all remaining subjects in a dataset.

Our primary motivation for using partial correlation FC (rather than the more common pairwise correlation FC) has been to reduce the number of false connections due to confounders (Reid et al., 2019). However, it is possible that hidden/unobserved confounders result in false connections even with regularized partial correlation. This possibility is substantially reduced (relative to, e.g., multi-unit recording) with wide-field-of-view imaging methods like fMRI or MEG/EEG. However, one straightforward way for future studies to reduce the chance of false positives from confounders is to include subcortical regions in addition to the cortical regions used here, given that there is substantial interaction between cortical and subcortical regions (Ji et al., 2019). It may also be advantageous to use regularized partial correlation at the highest spatial resolution available (e.g., voxels) rather than brain regions, given the possibility that confounder time series may not be fully accounted for when averaged into a larger neural population's time series.

Conclusion

The brain is a complex network, producing cognition through distributed processing by many regions. To understand the brain, it is therefore crucial to understand how its regions communicate, as is attempted with FC analyses. However, the quality of inferences depends on the performance of the specific FC method. Here,

we explore the issue of instability in multivariate FC methods, which are otherwise advantageous, and demonstrate their enhancement with regularization techniques. We thoroughly examine the performance of these methods, utilizing held-out resting-state fMRI data, structural connectivity, fMRI task activations, behavior, and simulations to assess both validity and reliability. In all, we show the regularized methods (especially graphical lasso) to be robust estimators of functional connections, which have strong potential to improve the quality of future FC studies.

Ethics

This study used data from the Human Connectome Project in Aging dataset, released through the NIMH Data Archive (NDA). All subjects gave signed informed consent in accordance with the protocol approved by the institutional review board associated with each data collection site (Washington University St. Louis, University of Minnesota, Massachusetts General Hospital, and University of California, Los Angeles). We followed the terms set in the NDA Data Use Certification, and our use of this data was approved by the Rutgers University institutional review board.

Author contributions

Kirsten L. Peterson: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - Original Draft, Writing - Review & Editing, Visualization
Ruben Sanchez-Romero: Methodology, Software, Writing - Review & Editing
Ravi D. Mill: Methodology, Writing - Review & Editing
Michael W. Cole: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding acquisition

Declaration of competing interest

The authors have no conflicts of interest to declare.

Data and code availability

Data and code to reproduce these analyses will be made available upon publication of this manuscript.

Acknowledgements

This work was supported by the US National Science Foundation under award 2219323. The empirical data used here (Human Connectome Project in Aging) was supported by the National Institute On Aging of the National Institutes of Health under Award Number U01AG052564 and by funds provided by the McDonnell Center for Systems Neuroscience at Washington University in St. Louis. We thank the Office of Advanced Research Computing at Rutgers, The State University of New Jersey, for providing access to the Amarel cluster and associated research computing resources that have contributed to the results reported here. This content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies.

References

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 47–97.

Bassett, D. S., & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience*, *20*(3), 353–364.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage*, *37*(1), 90–101.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Birn, R. M., Molloy, E. K., Patriat, R., Parker, T., Meier, T. B., Kirk, G. R., Nair, V. A., Meyerand, M. E., & Prabhakaran, V. (2013). The effect of scan length on the reliability of resting-state fMRI connectivity estimates. *NeuroImage*, *83*, 550–558.

Biswal, B., Yetkin, F. Z., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine: Official Journal of the Society of Magnetic Resonance in Medicine / Society of Magnetic Resonance in Medicine*, *34*(4), 537–541.

Blum, A., Hopcroft, J., & Kannan, R. (2020). Machine Learning. In *Foundations of Data Science* (pp. 109–158). Cambridge University Press.

Bookheimer, S. Y., Salat, D. H., Terpstra, M., Ances, B. M., Barch, D. M., Buckner, R. L., Burgess, G. C., Curtiss, S. W., Diaz-Santos, M., Elam, J. S., Fischl, B., Greve, D. N., Hagy, H. A., Harms, M. P., Hatch, O. M., Hedden, T., Hodge, C., Japardi, K. C., Kuhn, T. P., … Yacoub, E. (2019). The Lifespan Human Connectome Project in Aging: An overview. *NeuroImage*, *185*, 335–348.

Bowie, C. R., & Harvey, P. D. (2006). Administration and interpretation of the Trail Making Test. *Nature Protocols*, *1*(5), 2277–2281.

Brier, M. R., Mitra, A., McCarthy, J. E., Ances, B. M., & Snyder, A. Z. (2015). Partial covariance based functional connectivity computation using Ledoit-Wolf covariance regularization. *NeuroImage*, *121*, 29–38.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews. Neuroscience*, *10*(3), 186–198.

Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., Shinohara, R. T., Elliott, M. A., Eickhoff, S. B., Davatzikos, C., Gur, R. C., Gur, R. E., Bassett, D. S., & Satterthwaite, T. D. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, *154*, 174–187.

Cole, M. W., Bassett, D. S., Power, J. D., Braver, T. S., & Petersen, S. E. (2014). Intrinsic and task-evoked network architectures of the human brain. *Neuron*, *83*(1), 238–251.

Cole, M. W., Ito, T., Bassett, D. S., & Schultz, D. H. (2016). Activity flow over resting-state networks shapes cognitive task activations. *Nature Neuroscience*, *19*(12), 1718–1726.

Cole, M. W., Ito, T., Cocuzza, C., & Sanchez-Romero, R. (2021). The Functional Relevance of Task-State Functional Connectivity. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *41*(12), 2684–2702.

Damoiseaux, J. S., & Greicius, M. D. (2009). Greater than the sum of its parts: a review of studies combining structural connectivity and resting-state functional connectivity. *Brain Structure & Function*, *213*(6), 525–533.

Donahue, C. J., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Behrens, T. E., Dyrby, T. B., Coalson, T., Kennedy, H., Knoblauch, K., Van Essen, D. C., & Glasser, M. F. (2016). Using Diffusion Tractography to Predict Cortical Connection Strength and Distance: A Quantitative Comparison with Tracers in the Monkey. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *36*(25), 6758–6770.

Duff, E., Makin, T., Madugula, S., Smith, S. M., & Woolrich, M. W. (2013). Utility of Partial Correlation for Characterising Brain Dynamics: MVPA-based Assessment of Regularisation and Network Selection. *2013 International Workshop on Pattern Recognition in Neuroimaging*, 58–61.

Fiecas, M., Ombao, H., Linkletter, C., Thompson, W., & Sanes, J. (2010). Functional connectivity: shrinkage estimation and randomization test. *NeuroImage*, *49*(4), 3005–3014.

Fiecas, M., Ombao, H., van Lunen, D., Baumgartner, R., Coimbra, A., & Feng, D. (2013). Quantifying temporal correlations: a test-retest evaluation of functional connectivity in resting-state fMRI. *NeuroImage*, *65*, 231–241.

Foygel, R., & Drton, M. (2010). Extended Bayesian information criteria for Gaussian graphical models. *Advances in Neural Information Processing Systems*, *23*. https://proceedings.neurips.cc/paper/2010/hash/072b030ba126b2f4b2374f342be9ed44-Abstract.html

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* , *9*(3), 432–441.

Friston, K. J. (2011). Functional and Effective Connectivity: A Review. *Brain Connectivity*, *1*(1), 13–36.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., Jenkinson, M., & WU-Minn HCP Consortium. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage*, *80*, 105–124.

Harada, C. N., Natelson Love, M. C., & Triebel, K. L. (2013). Normal cognitive aging. *Clinics in Geriatric Medicine*, *29*(4), 737–752.

Harms, M. P., Somerville, L. H., Ances, B. M., Andersson, J., Barch, D. M., Bastiani, M., Bookheimer, S. Y., Brown, T. B., Buckner, R. L., Burgess, G. C., Coalson, T. S., Chappell, M. A., Dapretto, M., Douaud, G., Fischl, B., Glasser, M. F., Greve, D. N., Hodge, C., Jamison, K. W., … Yacoub, E. (2018). Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. *NeuroImage*, *183*, 972–984.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning* (2nd ed.). Springer New York.

Haufe, S., Tomioka, R., Nolte, G., Müller, K.-R., & Kawanabe, M. (2010). Modeling sparse connectivity between underlying brain sources for EEG/MEG. *IEEE Transactions on Bio-Medical Engineering*, *57*(8), 1954–1963.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Applications to nonorthogonal problems. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences*, *12*(1), 69.

Honey, C. J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J. P., Meuli, R., & Hagmann, P. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(6), 2035–2040.

Ito, T., Brincat, S. L., Siegel, M., Mill, R. D., He, B. J., Miller, E. K., Rotstein, H. G., & Cole, M. W. (2020). Task-evoked activity quenches neural correlations and variability across cortical areas. *PLoS Computational Biology*, *16*(8), e1007983.

Ji, J. L., Spronk, M., Kulkarni, K., Repovš, G., Anticevic, A., & Cole, M. W. (2019). Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*, *185*, 35–57.

Johnson, W., Nijenhuis, J. te, & Bouchard, T. J. (2008). Still just 1 g: Consistent results from five test batteries. *Intelligence*, *36*(1), 81–95.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, *31*(3), 300.

Le Bihan, D., & Johansen-Berg, H. (2012). Diffusion MRI at 25: exploring brain tissue structure and function. *NeuroImage*, *61*(2), 324–341.

Ledoit, O., & Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *Journal of Empirical Finance*, *10*(5), 603–621.

Lever, J., Krzywinski, M., & Altman, N. (2016). Model selection and overfitting. *Nature Methods*, *13*(9), 703–704.

Liégeois, R., Santos, A., Matta, V., Van De Ville, D., & Sayed, A. H. (2020). Revisiting correlation-based functional connectivity and its relationship with structural connectivity. *Network Neuroscience (Cambridge, Mass.)*, *4*(4), 1235–1251.

Mahadevan, A. S., Tooley, U. A., Bertolero, M. A., Mackey, A. P., & Bassett, D. S. (2021). Evaluating the sensitivity of functional connectivity measures to motion artifact in resting-state fMRI data. *NeuroImage*, *241*(118408), 118408.

Maier-Hein, K. H., Neher, P. F., Houde, J.-C., Côté, M.-A., Garyfallidis, E., Zhong, J., Chamberland, M., Yeh, F.-C., Lin, Y.-C., Ji, Q., Reddick, W. E., Glass, J. O., Chen, D. Q., Feng, Y., Gao, C., Wu, Y., Ma, J., He, R., Li, Q., … Descoteaux, M. (2017). The challenge of mapping the human connectome based on diffusion tractography. *Nature Communications*, *8*(1), 1349.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*, 654–660.

Marrelec, G., Krainik, A., Duffau, H., Pélégrini-Issac, M., Lehéricy, S., Doyon, J., & Benali, H. (2006). Partial correlation for functional brain interactivity investigation in functional MRI. *NeuroImage*, *32*(1), 228–237.

McCormick, E. M., Arnemann, K. L., Ito, T., Hanson, S. J., & Cole, M. W. (2022). Latent functional connectivity underlying multiple brain states. *Network Neuroscience (Cambridge, Mass.)*, *6*(2), 570–590.

Mejia, A. F., Nebel, M. B., Barber, A. D., Choe, A. S., Pekar, J. J., Caffo, B. S., & Lindquist, M. A. (2018). Improved estimation of subject-level functional connectivity using full and partial correlation with empirical Bayes shrinkage. *NeuroImage*, *172*, 478–491.

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*(1), 172–175.

Mill, R. D., Hamilton, J. L., Winfield, E. C., Lalta, N., Chen, R. H., & Cole, M. W. (2022). Network modeling of dynamic brain interactions predicts emergence of neural information that supports human cognitive behavior. *PLoS Biology*, *20*(8), e3001686.

Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., & Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: are anti-correlated networks introduced? *NeuroImage*, *44*(3), 893–905.

Nie, L., Yang, X., Matthews, P. M., Xu, Z.-W., & Guo, Y.-K. (2017). Inferring functional connectivity in fMRI using minimum partial correlation. *International Journal of Automation and Computing*, *14*(4), 371–385.

Noble, S., Scheinost, D., & Constable, R. T. (2021). A guide to the measurement and interpretation of fMRI test-retest reliability. *Current Opinion in Behavioral Sciences*, *40*, 27–32.

Pardoe, I. (2020). *Applied Regression Modeling*. John Wiley & Sons.

Peeters, C. F. W., Bilgrau, A. E., & van Wieringen, W. N. (2022). rags2ridges: A One-Stop-ℓ2-Shop for Graphical Modeling of High-Dimensional Precision Matrices. *Journal of Statistical Software*, *102*, 1–32.

Pervaiz, U., Vidaurre, D., Woolrich, M. W., & Smith, S. M. (2020). Optimising network modelling methods for fMRI. *NeuroImage*, *211*, 116604.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154.

Power, J. D., Laumann, T. O., Plitt, M., Martin, A., & Petersen, S. E. (2017). On Global fMRI Signals and Simulations [Review of *On Global fMRI Signals and Simulations*]. *Trends in Cognitive Sciences*, *21*(12), 911–913.

Power, J. D., Plitt, M., Gotts, S. J., Kundu, P., Voon, V., Bandettini, P. A., & Martin, A. (2018). Ridding fMRI data of motion-related influences: Removal of signals with distinct spatial and physical bases in multiecho data. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(9), E2105–E2114.

Power, J. D., Schlaggar, B. L., & Petersen, S. E. (2015). Recent progress and outstanding issues in motion correction in resting state fMRI. *NeuroImage*, *105*, 536–551.

Reid, A. T., Headley, D. B., Mill, R. D., Sanchez-Romero, R., Uddin, L. Q., Marinazzo, D., Lurie, D. J., Valdés-Sosa, P. A., Hanson, S. J., Biswal, B. B., Calhoun, V.,

Poldrack, R. A., & Cole, M. W. (2019). Advancing functional connectivity research from association to causation. *Nature Neuroscience*. https://doi.org/10.1038/s41593-019-0510-4

Revelle, W. R. (2017). *psych: Procedures for Personality and Psychological Research*. https://www.scholars.northwestern.edu/en/publications/psych-procedures-for-personality-and-psychological-research

Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique [The Rey Auditory Verbal Learning Test, RVLMT]. *Archives de Psychologie*.

Rheault, F., Poulin, P., Valcourt Caron, A., St-Onge, E., & Descoteaux, M. (2020). Common misconceptions, hidden biases and modern challenges of dMRI tractography. *Journal of Neural Engineering*, *17*(1), 011001.

Ryali, S., Chen, T., Supekar, K., & Menon, V. (2012). Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage*, *59*(4), 3852–3861.

Sala-Llonch, R., Smith, S. M., Woolrich, M., & Duff, E. P. (2019). Spatial parcellations, spectral filtering, and connectivity measures in fMRI: Optimizing for discrimination. *Human Brain Mapping*, *40*(2), 407–419.

Salthouse, T. A. (2010). Selective review of cognitive aging. *Journal of the International Neuropsychological Society: JINS*, *16*(5), 754–760.

Sanchez-Romero, R., & Cole, M. W. (2021). Combining Multiple Functional Connectivity Methods to Improve Causal Inferences. *Journal of Cognitive Neuroscience*, *33*(2), 180–194.

Sanchez-Romero, R., Ito, T., Mill, R. D., Hanson, S. J., & Cole, M. W. (2023). Causally informed activity flow models provide mechanistic insight into network-generated cognitive activations. *NeuroImage*, *278*, 120300.

Schaipp, F., Vlasovets, O., & Müller, C. (2021). GGLasso - a Python package for General Graphical Lasso computation. *Journal of Open Source Software*, *6*(68), 3865.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428.

Smith, S. M., Miller, K. L., Salimi-Khorshidi, G., Webster, M., Beckmann, C. F., Nichols, T. E., Ramsey, J. D., & Woolrich, M. W. (2011). Network modelling methods for FMRI. *NeuroImage*, *54*(2), 875–891.

Smith, S. M., Vidaurre, D., Beckmann, C. F., Glasser, M. F., Jenkinson, M., Miller, K. L., Nichols, T. E., Robinson, E. C., Salimi-Khorshidi, G., Woolrich, M. W., Barch, D. M., Uğurbil, K., & Van Essen, D. C. (2013). Functional connectomics from resting-state fMRI. *Trends in Cognitive Sciences*, *17*(12), 666–682.

Sotiropoulos, S. N., Jbabdi, S., Xu, J., Andersson, J. L., Moeller, S., Auerbach, E. J., Glasser, M. F., Hernandez, M., Sapiro, G., Jenkinson, M., Feinberg, D. A., Yacoub, E., Lenglet, C., Van Essen, D. C., Ugurbil, K., Behrens, T. E. J., & WU-Minn HCP Consortium. (2013). Advances in diffusion MRI acquisition and processing in the Human Connectome Project. *NeuroImage*, *80*, 125–143.

Sotiropoulos, S. N., & Zalesky, A. (2019). Building connectomes using diffusion MRI: why, how and but. *NMR in Biomedicine*, *32*(4), e3752.

Srinivasan, R., Winter, W. R., Ding, J., & Nunez, P. L. (2007). EEG and MEG coherence: measures of functional connectivity at distinct spatial scales of neocortical dynamics. *Journal of Neuroscience Methods*, *166*(1), 41–52.

Straathof, M., Sinke, M. R., Dijkhuizen, R. M., & Otte, W. M. (2019). A systematic review on the quantitative relationship between structural and functional network connectivity strength in mammalian brains. *Journal of Cerebral Blood Flow and Metabolism: Official Journal of the International Society of Cerebral Blood Flow and Metabolism*, *39*(2), 189–209.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, *58*(1), 267–288.

van den Heuvel, M. P., Mandl, R. C. W., Kahn, R. S., & Hulshoff Pol, H. E. (2009). Functionally linked resting-state networks reflect the underlying structural connectivity architecture of the human brain. *Human Brain Mapping*, *30*(10), 3127–3141.

van den Heuvel, M. P., & Sporns, O. (2011). Rich-club organization of the human connectome. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(44), 15775–15786.

Varoquaux, G., Gramfort, A., Poline, J.-B., & Thirion, B. (2010). Brain covariance selection: better individual functional connectivity models using population prior. *Advances in Neural Information Processing Systems*, *23*. https://proceedings.neurips.cc/paper/2010/hash/db576a7d2453575f29eab4bac787b919-Abstract.html

Wang, Y., Kang, J., Kemmer, P. B., & Guo, Y. (2016). An Efficient and Reliable Statistical Method for Estimating Functional Connectivity in Large Scale Brain Networks Using Partial Correlation. *Frontiers in Neuroscience*, *10*, 123.

Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., Carlozzi, N. E., Slotkin, J., Blitz, D., Wallner-Allen, K., Fox, N. A., Beaumont, J. L., Mungas, D., Nowinski, C. J., Richler, J., Deocampo, J. A., Anderson, J. E., Manly, J. J., Borosh, B., … Gershon, R. C. (2013). Cognition assessment using the NIH Toolbox. *Neurology*, *80*(11 Suppl 3), S54-64.

Yeh, F.-C., Verstynen, T. D., Wang, Y., Fernández-Miranda, J. C., & Tseng, W.-Y. I. (2013). Deterministic diffusion fiber tracking improved by quantitative anisotropy. *PloS One*, *8*(11), e80713.

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics. Conference Series*, *1168*(2), 022022.

Zalesky, A., Fornito, A., & Bullmore, E. (2012). On the use of correlation as a measure of network connectivity. *NeuroImage*, *60*(4), 2096–2106.

Zou, H., & Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *67*(2), 301–320.

Supplementary Materials

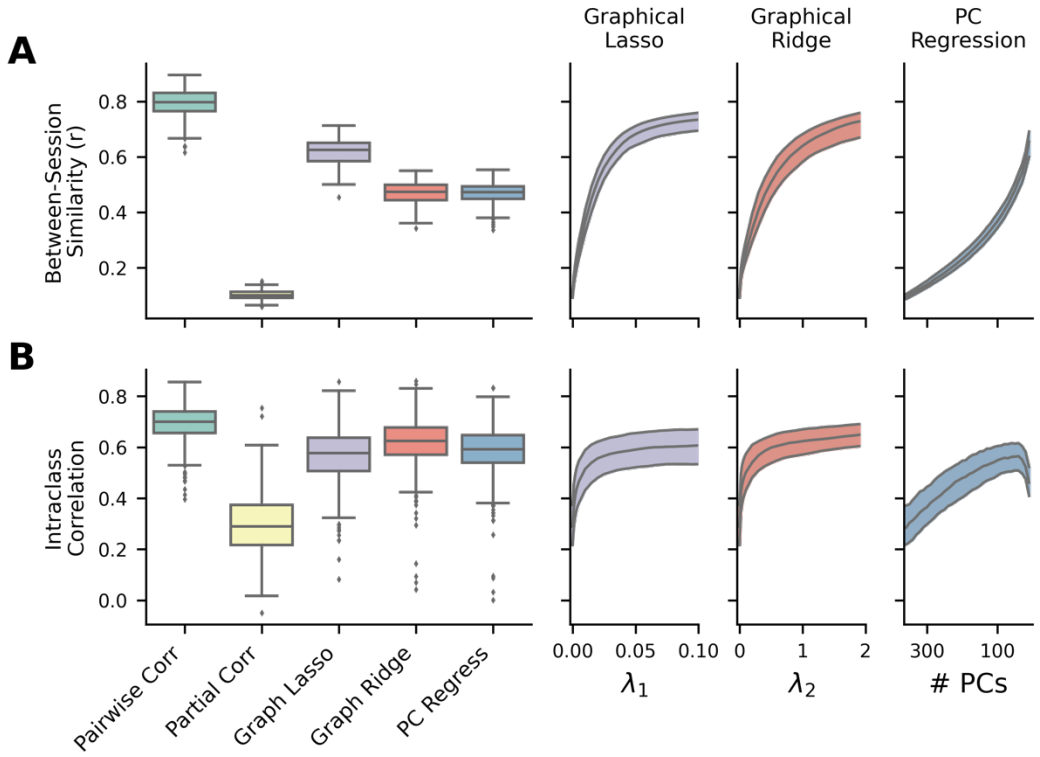

Figure S1 – Reliability of FC methods with empirical fMRI data – replication dataset. The boxplots show results where the regularization hyperparameters have been optimized for each FC matrix, while the right plots show the medians and IQRs across different hyperparameter values for the regularized methods. For PC regression, number of PCs is plotted in descending order because fewer PCs correspond with more regularization. **A)** Between-session similarity, the Pearson correlation between each subject's session 1 and session 2 FC matrices (n = 236). **B)** Intraclass correlation (Shrout & Fleiss, 1979), calculated for each edge in a conservative subset of nonzero edges (n = 535).

Table S1 – Statistical tests comparing FC methods on reliability and structural-functional similarity using empirical data – replication dataset. Between-session similarity, intraclass correlation, and individual structural-functional similarity were compared using two-tailed, dependent-sample t-tests. The Pearson's r values for between-session similarity and individual structural-functional similarity were normalized using Fisher's z transformation (arctanh) before statistical comparisons. Group-averaged structural-functional similarity (single Pearson's r per method) was compared using the method described by Meng et al. (1992), which tests for a significant difference between "correlated correlations" (correlations with a shared variable – structural connectivity weights in this case). For all reported correlation differences, correlations were transformed to Fisher's z, subtracted, and then transformed back to Pearson's r. Alpha levels were adjusted to .005 (.05/10) following Bonferroni correction to account for the multiple comparisons within each column.

	Methods	Between-session similarity (Pearson's r)	Intraclass correlation (ICC)	Structural-functional similarity (Pearson's r)	
				Group-averaged	Individual
	Pairwise correlation	mean r = 0.793, SD = 0.053	mean ICC = 0.695, SD = 0.068	r = 0.244	mean r = 0.164, SD = 0.019
	Partial correlation	mean r = 0.102, SD = 0.016	mean ICC = 0.294, SD = 0.115	r = 0.566	mean r = 0.135, SD = 0.013
	Graphical lasso	mean r = 0.616, SD = 0.046	mean ICC = 0.567, SD = 0.108	r = 0.535	mean r = 0.316, SD = 0.023
	Graphical ridge	mean r = 0.470, SD = 0.038	mean ICC = 0.617, SD = 0.095	r = 0.433	mean r = 0.219, SD = 0.019
	PC regression	mean r = 0.468, SD = 0.037	mean ICC = 0.586, SD = 0.096	r = 0.391	mean r = 0.195, SD = 0.019
1	Pairwise correlation vs. partial correlation	mean Δr = 0.758, T(235) = 108, p < .00001	mean ΔICC = 0.401, T(534) = 80.1, p < .00001	Δr = -0.374 z = -5.42 p < .00001	mean Δr = 0.030, T(235) = 29.2, p < .00001
2	Pairwise correlation vs. graphical lasso	mean Δr = 0.357, T(235) = 45.0, p < .00001	mean ΔICC = 0.128, T(534) = 28.4, p < .00001	Δr = -0.336 z = -4.97 p < .00001	mean Δr = -0.160, T(235) = -130, p < .00001
3	Pairwise correlation vs. graphical ridge	mean Δr = 0.525, T(235) = 70.7, p < .00001	mean ΔICC = 0.078, T(534) = 22.2, p < .00001	Δr = -0.212 z = -3.80 p = .0001	mean Δr = -0.057, T(235) = -58.7, p < .00001
4	Pairwise correlation vs. PC regression	mean Δr = 0.527, T(235) = 70.8, p < .00001	mean ΔICC = 0.108, T(534) = 31.1, p < .00001	Δr = -0.163 z = -2.96 p = .003	mean Δr = -0.032, T(235) = -32.9, p < .00001
5	Partial correlation vs. graphical lasso	mean Δr = -0.551, T(235) = -149, p < .00001	mean ΔICC = -0.272, T(534) = -74.9, p < .00001	Δr = 0.043 z = 1.80, p = 0.072	mean Δr = -0.189, T(235) = -170, p < .00001
6	Partial correlation vs. graphical ridge	mean Δr = -0.387, T(235) = -157, p < .00001	mean ΔICC = -0.323, T(534) = -70.2, p < .00001	Δr = 0.176 z = 4.43 p < .00001	mean Δr = -0.087, T(235) = -121, p < .00001
7	Partial correlation vs. PC regression	mean Δr = -0.386, T(235) = -137, p < .00001	mean ΔICC = -0.292, T(534) = -58.8, p < .00001	Δr = 0.225 z = 4.90 p < .00001	mean Δr = -0.062, T(235) = -84.9, p < .00001
8	Graphical lasso vs. graphical ridge	mean Δr = 0.208, T(235) = 85.7, p < .00001	mean ΔICC = -0.050, T(534) = -16.0, p < .00001	Δr = 0.133 z = 3.44, p = 0.0006	mean Δr = 0.104, T(235) = 123, p < .00001
9	Graphical lasso vs. PC regression	mean Δr = 0.209, T(235) = 66.0, p < .00001	mean ΔICC = -0.020, T(534) = -5.30, p < .00001	Δr = 0.183 z = 4.14 p = .00004	mean Δr = 0.129, T(235) = 140, p < .00001
10	Graphical ridge vs. PC regression	mean Δr = 0.002, T(235) = 0.995, p = 0.321	mean ΔICC = 0.030, T(534) = 22.7, p < .00001	Δr = 0.051 z = 5.55 p < .00001	mean Δr = 0.025, T(235) = 114, p < .00001

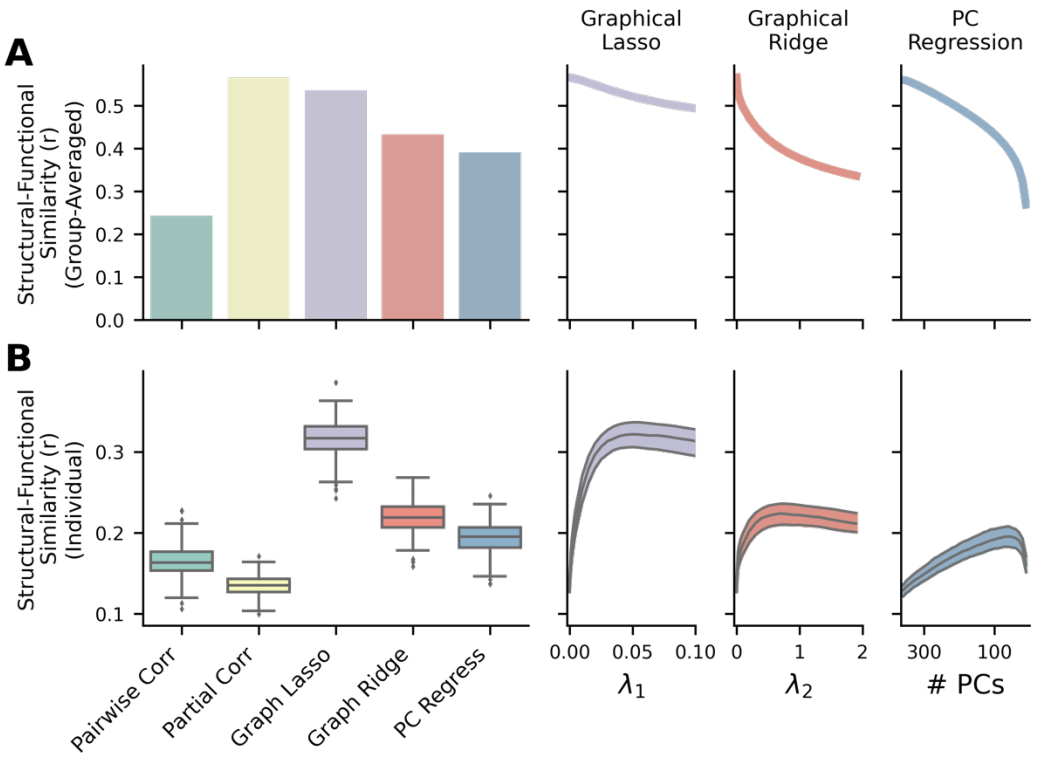

Figure S2 - Similarity between empirical FC and structural connectivity (diffusion MRI tractography) – replication dataset. The bar and boxplots show results where the regularization hyperparameters have been optimized for each FC matrix, while the right plots show the single values or the medians and IQRs across different hyperparameter values for the regularized methods. For PC regression, number of PCs is plotted in descending order because fewer PCs correspond with more regularization. **A)** Structural-functional similarity between group-averaged FC and structural matrices (n = 1). Averaging nullifies much of the noise in individual connectivity weights to show validity without the effects of low reliability. **B)** Structural-functional similarity between individual FC matrices and the same subjects'

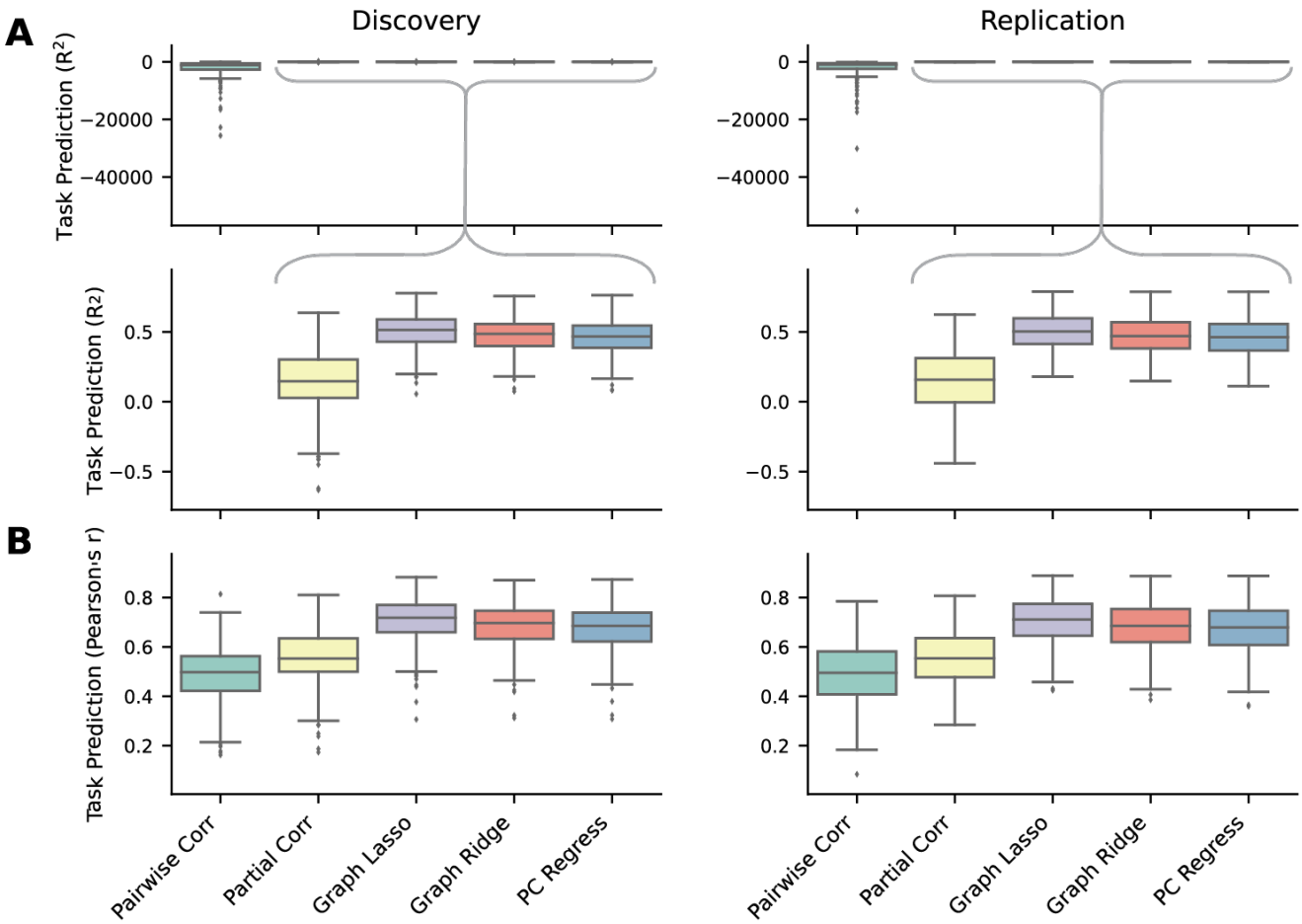

Figure S3 – Predicting task activations using FC estimated from rest fMRI data – discovery and replication datasets. **A)** Prediction accuracy by FC method calculated as the coefficient of determination (R^2) between regions' actual and predicted activations, computed in each subject (n = 236) across regions and task conditions. Results from the discovery dataset are shown on the left and replication dataset on the right. Partial correlation, graphical lasso, graphical ridge, and PC regression results are expanded in separate panels. **B)** Prediction accuracy calculated as Pearson correlation (r), which unlike R^2 is scale-invariant.

Table S2 – Statistical tests comparing FC methods on task activation prediction accuracy – discovery and replication datasets. Prediction accuracies with R^2 and Pearson's r were compared using two-tailed, dependent-sample t-tests. The Pearson's r values were normalized using Fisher's z transformation (arctanh) before statistical comparisons. For reported correlation differences, correlations were transformed to Fisher's z, subtracted, and then transformed back to Pearson's r. Alpha levels were adjusted to .005 (.05/10) following Bonferroni correction to account for the multiple comparisons within each column.

		Task activation prediction accuracy (R^2)		Task activation prediction accuracy (Pearson's r)	
	Methods	Discovery dataset	Replication dataset	Discovery dataset	Replication dataset
	Pairwise correlation	mean R^2 = -2228, SD = 3180	mean R^2 = -2429, SD = 4649	mean r = 0.490, SD = 0.119	mean r = 0.493, SD = 0.121
	Partial correlation	mean R^2 = 0.140, SD = 0.220	mean R^2 = 0.143, SD = 0.211	mean r = 0.551, SD = 0.113	mean r = 0.551, SD = 0.109
	Graphical lasso	mean R^2 = 0.506, SD = 0.123	mean R^2 = 0.502, SD = 0.122	mean r = 0.708, SD = 0.091	mean r = 0.705, SD = 0.088
	Graphical ridge	mean R^2 = 0.475, SD = 0.124	mean R^2 = 0.472, SD = 0.124	mean r = 0.684, SD = 0.094	mean r = 0.682, SD = 0.092
	PC regression	mean R^2 = 0.463, SD = 0.126	mean R^2 = 0.461, SD = 0.126	mean r = 0.675, SD = 0.097	mean r = 0.673, SD = 0.095
1	Pairwise correlation vs. partial correlation	mean ΔR^2 = -2229, T(235) = -10.7, p < .00001	mean ΔR^2 = -2429.338, T(235) = -8.01, p < .00001	mean Δr = -0.086, T(235) = -12.3, p < .00001	mean Δr = -0.081, T(235) = -11.5, p < .00001
2	Pairwise correlation vs. graphical lasso	mean ΔR^2 = -2229, T(235) = -10.7, p < .00001	mean ΔR^2 = -2429.698, T(235) = -8.01, p < .00001	mean Δr = -0.343, T(235) = -51.6, p < .00001	mean Δr = -0.334, T(235) = -51.5, p < .00001
3	Pairwise correlation vs. graphical ridge	mean ΔR^2 = -2229, T(235) = -10.7, p < .00001	mean ΔR^2 = -2429.667, T(235) = -8.01, p < .00001	mean Δr = -0.299, T(235) = -48.1, p < .00001	mean Δr = -0.292, T(235) = -47.1, p < .00001
4	Pairwise correlation vs. PC regression	mean ΔR^2 = -2229, T(235) = -10.7, p < .00001	mean ΔR^2 = -2429.656, T(235) = -8.01, p < .00001	mean Δr = -0.284, T(235) = -45.2, p < .00001	mean Δr = -0.278, T(235) = -44.5, p < .00001
5	Partial correlation vs. graphical lasso	mean ΔR^2 = -0.366, T(235) = -49.5, p < .00001	mean ΔR^2 = -0.360, T(235) = -47.9, p < .00001	mean Δr = -0.265, T(235) = -81.2, p < .00001	mean Δr = -0.260, T(235) = -68.5, p < .00001
6	Partial correlation vs. graphical ridge	mean ΔR^2 = -0.335, T(235) = -44.1, p < .00001	mean ΔR^2 = -0.329, T(235) = -43.4, p < .00001	mean Δr = -0.219, T(235) = -63.6, p < .00001	mean Δr = -0.217, T(235) = -54.8, p < .00001
7	Partial correlation vs. PC regression	mean ΔR^2 = -0.322, T(235) = -41.5, p < .00001	mean ΔR^2 = -0.318, T(235) = -40.4, p < .00001	mean Δr = -0.203, T(235) = -51.5, p < .00001	mean Δr = -0.202, T(235) = -44.4, p < .00001
8	Graphical lasso vs. graphical ridge	mean ΔR^2 = 0.032, T(235) = 23.7, p < .00001	mean ΔR^2 = 0.030, T(235) = 21.9, p < .00001	mean Δr = 0.049, T(235) = 25.9, p < .00001	mean Δr = 0.046, T(235) = 23.6, p < .00001
9	Graphical lasso vs. PC regression	mean ΔR^2 = 0.044, T(235) = 25.1, p < .00001	mean ΔR^2 = 0.042, T(235) = 24.5, p < .00001	mean Δr = 0.065, T(235) = 27.3, p < .00001	mean Δr = 0.062, T(235) = 26.1, p < .00001
10	Graphical ridge vs. PC regression	mean ΔR^2 = 0.012, T(235) = 15.3, p < .00001	mean ΔR^2 = 0.011, T(235) = 14.3, p < .00001	mean Δr = 0.017, T(235) = 15.6, p < .00001	mean Δr = 0.016, T(235) = 14.6, p < .00001

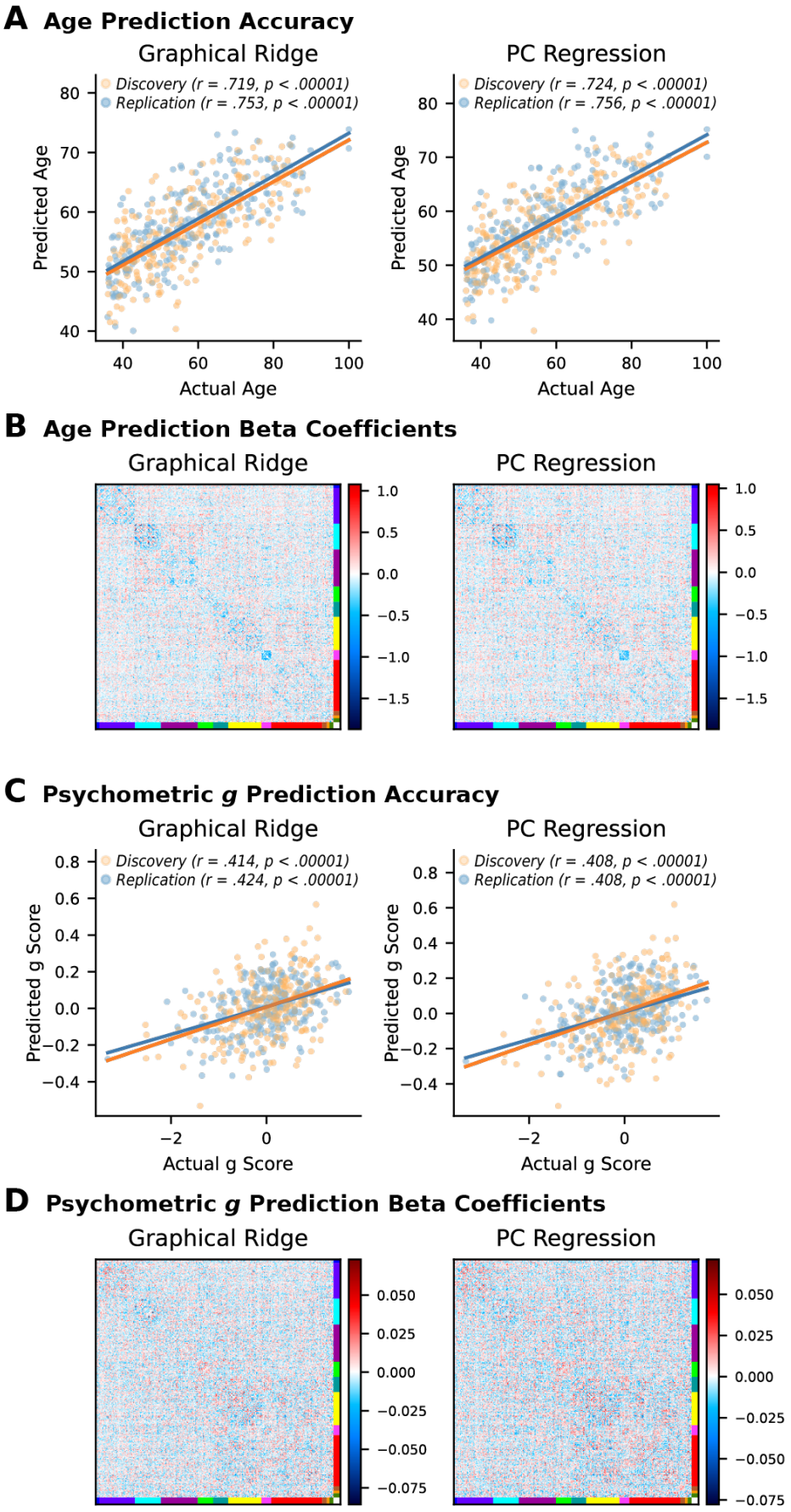
A Age Prediction Accuracy
Graphical Ridge
Discovery (r = .719, p < .00001)
Replication (r = .753, p < .00001)
Predicted Age
Actual Age
PC Regression
Discovery (r = .724, p < .00001)
Replication (r = .756, p < .00001)
Predicted Age
Actual Age
B Age Prediction Beta Coefficients
Graphical Ridge
PC Regression
C Psychometric g Prediction Accuracy
Graphical Ridge
Discovery (r = .414, p < .00001)
Replication (r = .424, p < .00001)
Predicted g Score
Actual g Score
PC Regression
Discovery (r = .408, p < .00001)
Replication (r = .408, p < .00001)
Predicted g Score
Actual g Score
D Psychometric g Prediction Beta Coefficients
Graphical Ridge
PC Regression

Figure S4 – Predicting individual differences in age and intelligence using estimated FC from rest fMRI data – graphical ridge and PC regression. **A)** Actual and predicted ages of each subject by FC method, from both the discovery and replication datasets. **B)** Average beta coefficients assigned to each connection by the regression models for estimating subject age. Beta coefficients shown here are the averages over cross-validation folds and discovery and replication datasets. Blue indicates that connection strength decreased with age and red indicates that connection strength increased. **C)** Actual and predicted intelligence (psychometric *g*) of each subject by FC method. **D)** Average beta coefficients assigned to each connection by the regression models for estimating psychometric *g*. See Figure 9 for results using pairwise correlation, partial correlation, and graphical lasso for FC estimation.

Table S3 – Statistical tests comparing FC methods on age and intelligence prediction accuracy – discovery and replication datasets. Predicted-to-actual correlations (single Pearson's r per method) were compared between methods using the two-tailed test described by Meng, Rosenthal, and Rubin (1992), which tests for a significant difference between "correlated correlations" (correlations with a shared variable – actual subject age and psychometric *g* in this case). For reported correlation differences, correlations were transformed to Fisher's z, subtracted, and then transformed back to Pearson's r. Alpha levels were adjusted to .005 (.05/10) following Bonferroni correction to account for the multiple comparisons within each column.

		Age prediction accuracy (Pearson's r)		Psychometric *g* prediction accuracy (Pearson's r)	
Methods		Discovery dataset	Replication dataset	Discovery dataset	Replication dataset
1	Pairwise correlation vs. partial correlation	$\Delta r = 0.269$, $z = 4.162$, $p = .00003$	$\Delta r = 0.194$, $z = 3.010$, $p = .0026$	$\Delta r = 0.242$, $z = 3.030$, $p = .002$	$\Delta r = 0.394$, $z = 4.779$, $p < .00001$
2	Pairwise correlation vs. graphical lasso	$\Delta r = 0.031$, $z = 0.576$, $p = .564$	$\Delta r = -0.031$, $z = -0.582$, $p = .561$	$\Delta r = -0.034$, $z = -0.505$, $p = .613$	$\Delta r = 0.050$, $z = 0.718$, $p = .473$
3	Pairwise correlation vs. graphical ridge	$\Delta r = -0.019$, $z = -0.365$, $p = .716$	$\Delta r = -0.061$, $z = -1.198$, $p = .231$	$\Delta r = -0.058$, $z = -0.958$, $p = .338$	$\Delta r = -0.010$, $z = -0.148$, $p = .882$
4	Pairwise correlation vs. PC regression	$\Delta r = -0.029$, $z = -0.572$, $p = .567$	$\Delta r = -0.068$, $z = -1.294$, $p = .196$	$\Delta r = -0.050$, $z = -0.837$, $p = .403$	$\Delta r = 0.010$, $z = 0.150$, $p = .881$
5	Partial correlation vs. graphical lasso	$\Delta r = -0.240$, $z = -5.211$, $p < .00001$	$\Delta r = -0.224$, $z = -5.194$, $p < .00001$	$\Delta r = -0.273$, $z = -3.906$, $p = .00009$	$\Delta r = -0.351$, $z = -5.155$, $p < .00001$
6	Partial correlation vs. graphical ridge	$\Delta r = -0.287$, $z = -6.377$, $p < .00001$	$\Delta r = -0.252$, $z = -5.619$, $p < .00001$	$\Delta r = -0.295$, $z = -4.487$, $p < .00001$	$\Delta r = -0.402$, $z = -6.201$, $p < .00001$
7	Partial correlation vs. PC regression	$\Delta r = -0.297$, $z = -6.372$, $p < .00001$	$\Delta r = -0.259$, $z = -5.613$, $p < .00001$	$\Delta r = -0.289$, $z = -3.983$, $p = .0001$	$\Delta r = -0.386$, $z = -5.725$, $p < .00001$
8	Graphical lasso vs. graphical ridge	$\Delta r = -0.050$, $z = -1.929$, $p = .054$	$\Delta r = -0.030$, $z = -1.249$, $p = .212$	$\Delta r = -0.024$, $z = -0.614$, $p = .540$	$\Delta r = -0.060$, $z = -1.422$, $p = .155$
9	Graphical lasso vs. PC regression	$\Delta r = -0.060$, $z = -2.259$, $p = .024$	$\Delta r = -0.037$, $z = -1.411$, $p = .158$	$\Delta r = -0.017$, $z = -0.388$, $p = .698$	$\Delta r = -0.040$, $z = -0.857$, $p = .391$
10	Graphical ridge vs. PC regression	$\Delta r = -0.011$, $z = -0.627$, $p = .531$	$\Delta r = -0.007$, $z = -0.442$, $p = .659$	$\Delta r = 0.007$, $z = 0.327$, $p = .744$	$\Delta r = 0.019$, $z = 0.845$, $p = .398$