NeuroImage 201 (2019) 116038

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Combining multiple connectomes improves predictive modeling of phenotypic measures

Siyuan Gao [a,*], Abigail S. Greene [b], R. Todd Constable [b,c,d], Dustin Scheinost [b,c,e,f]

[a] Department of Biomedical Engineering, Yale University, United States
[b] Interdepartmental Neuroscience Program, Yale School of Medicine, United States
[c] Department of Radiology and Biomedical Imaging, Yale School of Medicine, United States
[d] Department of Neurosurgery, Yale School of Medicine, United States
[e] Department of Statistics and Data Science, Yale University, United States
[f] Child Study Center, Yale School of Medicine, United States

ARTICLE INFO

Keywords:
Machine learning
Neural networks
Elastic net
Lasso
fMRI
Functional connectivity

ABSTRACT

Resting-state and task-based functional connectivity matrices, or connectomes, are powerful predictors of individual differences in phenotypic measures. However, most of the current state-of-the-art algorithms only build predictive models based on a single connectome for each individual. This approach neglects the complementary information contained in connectomes from different sources and reduces prediction performance. In order to combine different task connectomes into a single predictive model in a principled way, we propose a novel prediction framework, termed multidimensional connectome-based predictive modeling. Two specific algorithms are developed and implemented under this framework. Using two large open-source datasets with multiple tasks—the Human Connectome Project and the Philadelphia Neurodevelopmental Cohort, we validate and compare our framework against performing connectome-based predictive modeling (CPM) on each task connectome independently, CPM on a general functional connectivity matrix created by averaging together all task connectomes for an individual, and CPM with a naïve extension to multiple connectomes where each edge for each task is selected independently. Our framework exhibits superior performance in prediction compared with the other competing methods. We found that different tasks contribute differentially to the final predictive model, suggesting that the battery of tasks used in prediction is an important consideration. This work makes two major contributions: First, two methods for combining multiple connectomes from different task conditions in one predictive model are demonstrated; Second, we show that these models outperform a previously validated single connectome-based predictive model approach.

1. Introduction

Advanced functional magnetic resonance imaging (fMRI) techniques, particularly functional connectivity analyses, are revealing robust individual differences in patterns of neural activity that predict continuous phenotypic measures (Dubois and Adolphs, 2016; Rosenberg et al., 2018). Predictive modeling of the associations between phenotypic measures and the functional organization of an individual's brain improves generalization of results to novel individuals and increases their eventual clinical utility. Recent work has used functional connectivity matrices, or connectomes, to predict a wide range of phenotypic measures, including fluid intelligence (Finn et al., 2015), brain maturity (Dosenbach et al., 2010), and sustained attention (Rosenberg et al., 2015). However, most of the current state-of-the-art algorithms only build predictive models based on a single connectome for each individual (Dadi et al., 2018). This approach neglects the complementary information contained in connectomes from different sources and reduces prediction performance.

While functional connectivity is usually calculated from data acquired during rest, task conditions better reveal individual differences (Finn et al., 2017; Vanderwal et al., 2017) and improve phenotypic prediction (Greene et al., 2018; Rosenberg et al., 2015). Further, the observed improvement in predictive power appears to be task specific, suggesting that task conditions are likely better at generating models of phenotypes related to the circuits they perturb (Greene et al., 2018; Rosenberg et al., 2015). Overall, it is unlikely that a single task can be developed that is

* Corresponding author. Siyuan Gao Magnetic Resonance Research Center 300 Cedar St, PO Box 208043, New Haven, CT, 06520-8043, United States.
E-mail address: siyuan.gao@yale.edu (S. Gao).

https://doi.org/10.1016/j.neuroimage.2019.116038
Received 6 May 2019; Received in revised form 18 July 2019; Accepted 19 July 2019
Available online 20 July 2019
1053-8119/© 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

optimal for all phenotypes. Instead, methods that incorporate functional connectivity information from a spectrum of tasks into a single predictive model may yield the best performance and most generalizable method for predicting phenotypic measures from connectomes.

In order to combine different task connectomes into a single predictive model in a principled way, we propose a novel prediction framework, termed multidimensional connectome-based predictive modeling. Two algorithms, each with their own strengths and limitations, are provided to illustrate the advantage of utilizing multiple connectomes. Both are based on the previously validated connectome-based predictive modeling (CPM) method (Shen et al., 2017). The first utilizes canonical correlation analysis (CCA), while the second utilizes ridge regression. CCA combines multiple task connectomes by finding the projection direction which maximizes correlation between the combined connectomes and behavioral measure(s) to be predicted. In contrast, ridge regression directly incorporates the large number of edges in multiple connectomes through regularization.

Using two large open-source datasets with multiple tasks—the Human Connectome Project (HCP) (Van Essen et al., 2013) and the Philadelphia Neurodevelopmental Cohort (PNC) (Satterthwaite et al., 2016), we validate and compare our two algorithms against performing CPM on each task connectome independently, CPM on a general functional connectivity (GFC) (Elliott et al., 2019) matrix created by concatenating time series from all task conditions to create a single connectome for an individual, ridge regression on GFC and CPM with a naïve extension to multiple connectomes where each edge for each task is selected independently. Preliminary versions of this work were presented at the 2018 IEEE International Symposium on Biomedical Imaging (Gao et al., 2018a) and the 21st International Conference on Medical Image Computing and Computer Assisted Intervention (Gao et al., 2018b).

In all, our contribution in this paper is two-fold. First, we propose the combination of multiple connectomes from different task conditions in one predictive model. Second, we developed two algorithms for dealing with multiple connectomes and show that they outperformed the validated single connectome predictive model.

2. Methods

2.1. Connectome-based predictive modeling (CPM)

CPM (Shen et al., 2017) is a validated method for extracting and pooling the most relevant features from connectivity data in order to construct linear models to predict phenotypic measures (Fig. 1a). Briefly, edges of connectivity matrices that are significantly correlated with the phenotypic measure of interest are selected. The selected features are then pooled (e.g. averaged) and linear regression is used to predict the phenotypic measure in novel individuals. It is designed for single connectome-based prediction, but can be easily extended to multiple-connectome scenarios as each edge is selected independently. However, this simple approach is not equipped to efficiently incorporate the increasing number of features introduced by multiple connectomes.

2.2. Multidimensional connectome-based predictive modeling

Although CPM can be extended to leverage multiple connectomes, a specially designed framework will better utilize the complementary information in different brain connectivity patterns driven by corresponding task conditions. Here, we present two realizations of this framework to illustrate the feasibility and advantage of combining multiple connectomes for prediction.

2.2.1. CCA connectome-based predictive modeling (cCPM)

Canonical Correlation Analysis (CCA): For two sets of observation matrices **X** and **Y**, assuming that the variables are correlated, CCA seeks

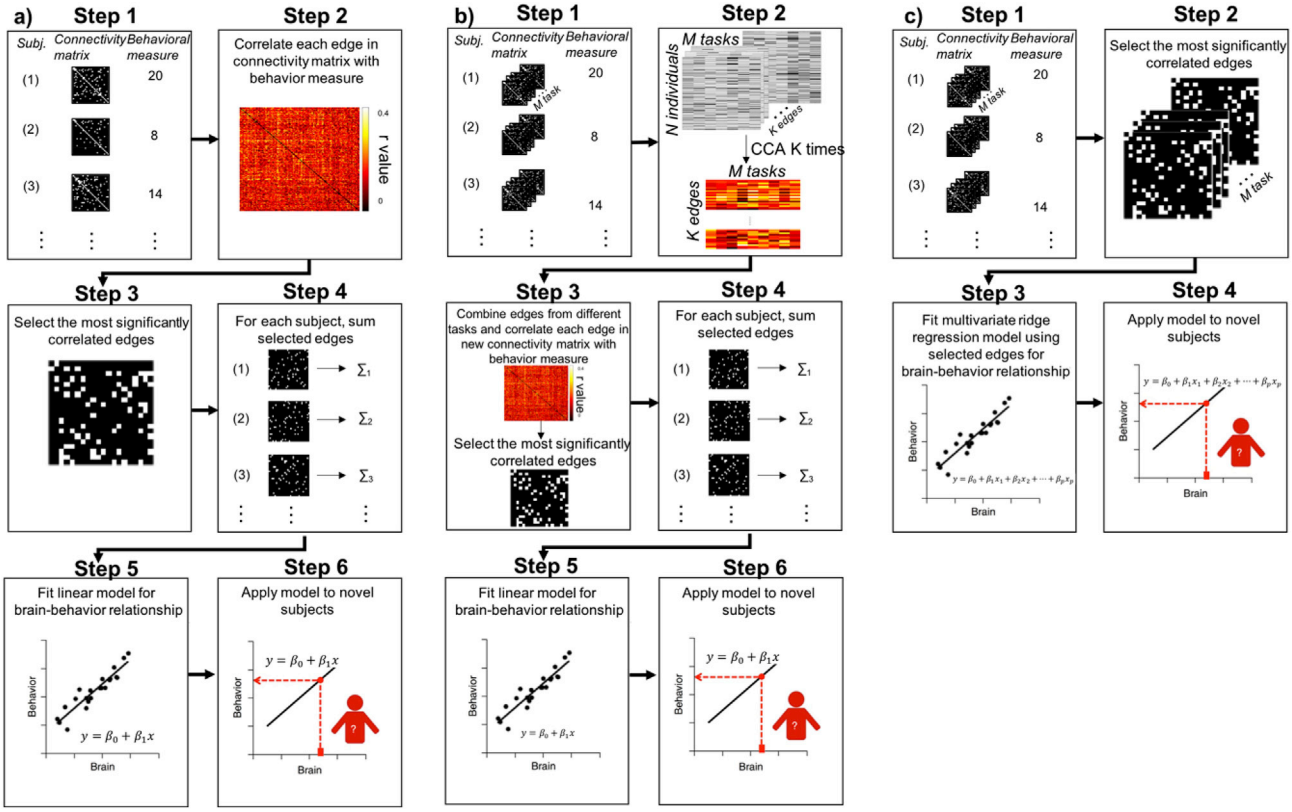

Fig. 1. Algorithm flow chart for three major models mentioned. a) The original CPM flow chart **b)** cCPM extends CPM to handle multiple connectomes per individual by replacing the correlation step in CPM with a canonical correlation analysis (CCA) step. **c)** rCPM extends CPM to handle multiple connectomes per individual by replacing the pooling (i.e. averaging) and linear regression step with a ridge regression step.

linear combinations of the columns of these two matrices that maximize correlation between them. In other words, we want to find vectors **a** and **b** such that the random variables **Xa** and **Yb** maximize the correlation. Assuming that **X** and **Y** are centered such that each column of either matrix has mean zero, the correlation to be maximized can be expressed by the following equation:

$$\rho = \frac{(\mathbf{Xa})^{\mathrm{T}}(\mathbf{Yb})}{\sqrt{\left[(\mathbf{Xa})^{\mathrm{T}}(\mathbf{Xa})\right]\left[(\mathbf{Yb})^{\mathrm{T}}(\mathbf{Yb})\right]}}$$

CCA Connectome-based Predictive Modeling: The cCPM pipeline consists of six steps (Fig. 1b). In the first step, individuals are divided into training and testing sets using 10-fold cross-validation. We denote the number of individuals for training as N^{train} and number of individuals for testing as N^{test}. In the second step, edges are combined. For the k-th edge, we have a matrix $\mathbf{E}_k \in \mathbf{R}^{N^{train} * M}$ containing edge strength of the k-th edge for all the individuals in the training set. Each row of the matrix $\mathbf{E}_k$ denotes each training individual's k-th edge's different strengths under M different tasks. Using CCA, we can find the canonical coefficients $\mathbf{w}_k \in \mathbf{R}^M$ for each edge. As each edge matrix $\mathbf{E}_k$ corresponds to the observation matrix **X** in the above definition equation for CCA, these coefficients $\mathbf{w}_k$ correspond to the vector **a**, and the observation matrix **Y** will store the behavioral measures. We then combine connectomes from all tasks into a total connectivity matrix using different canonical correlations, $\mathbf{E}_k^{\text{total}} = \sum_{m=1}^{M} \mathbf{E}_k(:, m)\mathbf{w}_k(m)$, where the m-th column of $\mathbf{E}_k$s denoted as $\mathbf{E}_k(:, m)$. Within each single task, each edge is demeaned across different individuals so that each column of $\mathbf{E}_k$ has mean 0. In the third step, we assign the combined edges that are significantly correlated with the behavioral measures to the "correlated network" (CN). The significance of the correlation is found from the CCA. Here, we assume that CCA always maximizes the positive correlation between combined edge strength and behavioral measure as the sign of the canonical coefficients can trivially be changed to maximize the positive correlation. Various significance thresholds for feature selection can be used. In the fourth step, we calculate "network strength" $\mathbf{s}^{\mathrm{CN}}$ by pooling (i.e. summing) the strength of all CN edges in each individual's total connectivity matrix, yielding a summary value $\mathbf{s}^{\mathrm{CN}}(\mathrm{i})$ for each individual:

$$\mathbf{s}^{\mathrm{CN}} = \sum_{\mathrm{k}}\left(\mathbf{B}(\mathrm{k})\sum_{\mathrm{m}}\mathbf{E}_{\mathrm{k}}(:, \mathrm{m})\mathbf{w}_{\mathrm{k}}(\mathrm{m})\right)$$

where $\mathbf{s}^{\mathrm{CN}}$ is the vector of summary values, $\mathbf{B}(k)$ is the indicator of whether the k-th edge passes the thresholding for CN. In the fifth step, we use linear regression ($\mathbf{y} = \beta_0 + \beta_1\mathbf{s}^{\mathrm{CN}}$) to model the association between "network strength" and the phenotypic measure in N^{train} individuals. In the sixth step, the "network strength" is calculated for the excluded N^{test} individuals, and is submitted to the corresponding regression model to generate phenotypic measure estimates for those testing individuals. This process is repeated iteratively, with different individuals in the training and testing sets.

2.2.2. Ridge regression connectome-based predictive modeling (rCPM)

Ridge regression: In ordinary least-squares (OLS) regression, a greater number of independent variables compared to the number of observations leads to an ill-posed problem and overfitting. To solve this ill-posed problem, regularization on regression coefficients can be applied to shrink the coefficients. Ridge regression shrinks the regression coefficients by imposing a L^2-norm penalty on their size. Compared with OLS regression, the coefficients from ridge regression minimize a penalized residual sum of squares,

$$\widehat{\beta}^{ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\sum_{i=1}^{N}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\} \quad (1)$$

where λ is the complexity parameter that controls the shrinkage strength: $\lambda = 0$ gives rise to the unregularized OLS, while increasing λ shrinks the coefficients towards zero. If we write the criterion in equation (1) in matrix form, $\mathrm{RSS}(\lambda) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}$, the ridge regression solutions can be solved by

$$\widehat{\boldsymbol{\beta}}^{\text{ridge}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

where **I** is the identity matrix. Compared with the solution for OLS, $\widehat{\boldsymbol{\beta}}^{\text{OLS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$, adding a positive constant to the diagonal of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ before inversion makes the problem nonsingular, even if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not of full rank.

Ridge regression CPM: Based on ridge regression, we modify the original CPM framework to better suit the high-dimensional nature of connectivity data (Fig. 1c). Specifically, due to the positive semi-definite nature of a functional connectivity matrix, the edges are not independent. Ridge regression is more robust than OLS in this case.

Instead of summing selected edges and fitting a one-dimensional OLS model, we directly fit a ridge regression model with training individuals using the selected edges from all the tasks and apply the model to testing individuals in the cross-validation framework. λ parameter in the ridge regression is chosen by another inner 10-fold cross-validation which uses only the training individuals. The largest λ value that has a mean squared error (MSE) within one standard error of the minimum MSE is chosen. In the Results (Section 3.3), we show that rCPM is not sensitive to λ.

2.3. Experiment setup

Datasets: We applied all algorithms (see Competing methods below) to the Human Connectome Project (HCP) 900 Subject Release and the Philadelphia Neurodevelopmental Cohort (PNC) first study release. These data releases were the only releases available at the time that this work began.

Phenotypic measure: In both datasets, a matrix reasoning test—a measure of fluid intelligence (gF)—was used as the phenotypic measure for prediction. In the HCP dataset, a 24-item version of the Penn Progressive Matrices test was used; this test is an abbreviated form of Raven's standard progressive matrices (Bilker et al., 2012). In the PNC dataset, 24- and 18-item versions of the Penn Matrix Reasoning Test were used (Bilker et al., 2012; Moore et al., 2015). Integer scores indicate number of correct responses (HCP: PMAT24_A_CR, range = 5–24, mean = 17.53, s.d. = 4.45, median = 19; PNC: PMAT_CR (phv00194834.v1.p1.c1), range = 0–23, mean = 12.27, s.d. = 4.04, median = 12).

HCP participants: From this dataset, we restricted our analyses to those individuals who participated in all nine fMRI conditions (seven task, two rest), whose mean frame-to-frame displacement was less than 0.1 mm and whose maximum frame-to-frame displacement was less than 0.15 mm (see HCP imaging parameters and preprocessing), and for whom gF measures were available ($n = 515$; 241 males; ages 22–37). This conservative threshold for exclusion due to motion was used to mitigate the substantial effects of motion on functional connectivity; following this exclusion, there was no significant correlation between motion and gF for most conditions (all $p > 0.05$, Bonferroni corrected) except the Social task, right-left (RL) phase encoding run ($r_s = -0.16$ ($p = 0.00017$)), the Relational task, left-right (LR) phase encoding run ($r_s = -0.15$ ($p = 0.0008$)), and the Emotion task, RL phase encoding run ($r_s = -0.14$ ($p = 0.0017$)).

PNC participants: From this dataset, we used behavioral, structural imaging, and functional imaging data. We restricted our analyses to those individuals who participated in all three fMRI runs (two task, one rest), on whom registration was successful (nine individuals were excluded for failed registrations), whose mean frame-to-frame displacement was less than 0.1 mm and whose maximum frame-to-frame displacement was less than 0.15 mm (as for the HCP dataset, and with the same motivation),

and for whom gF measures were available (n = 571; 251 male, ages 8–21). Following exclusion for motion, there was no significant correlation between motion and gF for any condition (all $p > 0.05$, Bonferroni corrected).

fMRI processing: fMRI data were processed with standard methods and parcellated into 268 nodes using a whole-brain, functional atlas defined in a separate sample (see (Greene et al., 2018) for more details). Task functional connectivity was calculated based on the "raw" task timecourses, with no regression of task-evoked activity: the mean timecourses of each node pair were correlated and correlation coefficients were Fisher transformed. Matrices were generated for both the LR and RL phase encoding runs in the HCP data, and these matrices were averaged for each condition, thus generating one 268×268 connectivity matrix per individual per task condition. Nodes that have missing coverage during any individual's scan were excluded from all individuals (9 nodes in HCP and 18 nodes in PNC were excluded). These matrices were used to generate cross-validated predictive models of gF.

Competing methods: We compared cCPM and rCPM to four simpler CPM-based approaches. For the first approach, we performed CPM on each task independently as previously demonstrated (Greene et al., 2018). For the second approach, we performed CPM on a general functional connectivity (GFC) matrix created by averaging brain connectivity information across all task conditions (Elliott et al., 2019). As time courses are z-score normalized before creating connectomes, averaging connectomes is similar to first, concatenating time courses and, then, correlating them. However, as the time length of each task is different, by concatenating time series first avoids biasing FC estimates toward the shortest tasks. Here, we generated GFC as suggested in the original paper by first concatenate and then correlate time courses. For the third approach, we compared ridge regression with CPM on GFC matrices. This is similar to performing rCPM on GFC matrices. However, we have chosen not to use the "rCPM" term here to avoid confusion between using GFC and our more direct way of combining multiple connectomes. Finally, for the fourth approach, we used a naïve extension to CPM, where all task connectomes were vectorized and concatenated to create a feature space that contained all task data for CPM. In contrast to the cCPM and rCPM approaches, this naïve implementation does not consider any shared or unique information offered by each task. Corrected resampled t-tests (Bouckaert and Frank, 2010) were used to compare competing methods.

Internal validation: 10-fold cross-validation was used to train all models. In 10-fold cross-validation, the sample was randomly divided into 10, approximately equal-sized groups; on each fold, the model was trained on 9 groups and tested on the excluded 10th group. This process was iteratively repeated 10 times, with each group excluded once. This procedure was repeated for 100 random divisions. CPM was performed with a range of p-value edge selection thresholds from 0.001 to 0.5. Model performance was evaluated by the cross-validated R^2,

$$R^2_{CV} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - y)^2}$$

R^2_{CV} can be negative (Scheinost et al., 2019) and negative values were set to 0. In this paper, $\sqrt{R^2_{CV}}$ is reported as it is comparable to, but less biased than, the normally used Pearson correlation between observed and predicted measures when using cross-validation. $\sqrt{R^2_{CV}}$ is averaged over the cross validation folds.

External validation: Additionally, we trained models using one of the datasets (either HCP or PNC) and applied the model to the other dataset. For external validation, we only used the Emotion and Working Memory tasks from HCP for consistency with the available task data from the PNC. To fairly compare between models, CPM was performed with the 50%, 5% and 1% of edges with lowest p-values. Model performance was evaluated by Pearson correlation coefficient ($r_{Pearson}$) between the predicted and observed gF measures.

Quantification of task contribution: To quantify the contribution of each task to a given predictive model, we calculated the m-th task's average weight (labeled W_m) to the model as

$$W_m = \sum_k \mathbf{B}(k)\boldsymbol{\beta}_k^m std(\mathbf{E}_k(:, m))$$

where $\mathbf{B}(k)$ indexes whether the k-th edge is selected, $std(\mathbf{E}_k(:, m))$ represents the standard deviation of the k-th edge in the m-th task and $\boldsymbol{\beta}_k^m$ represents the weight learned by cCPM or rCPM for the k-th edge in the m-th task. To make the results more interpretable, $W_m$are then normalized to have sum 1, $\sum_m W_m = 1$, so that it represents each task's contribution proportion in the whole model.

Similarly, as certain tasks may contain redundant information for prediction, we adopted forward feature selection to select the optimal combination of tasks. Forward feature selection finds the optimal combination of tasks by adding each of the tasks in a stepwise way; in each step, the task that improves prediction the most will be added to the selected task list. The optimal combination is found when any additional task won't lead to further improvement in prediction. We found the optimal task combination for HCP with both cCPM and rCPM. Both of the two algorithms are performed with a p-value threshold of 0.1.

Sensitivity to hyperparameters: Although all of the tested approaches are relatively simple in terms of choosing hyperparameters before training the model, all are dependent on the chosen p-value threshold for edge selection. To evaluate the sensitivity of each approach to this hyperparameter, we repeated analyses with p-value thresholds of 0.001, 0.005, 0.01, 0.05, 0.1, and 0.5. The rCPM approach has an additional hyperparameter: the penalty weighting parameter λ. To show that the rCPM models are also not sensitive to the choice of λ, we fixed λ at the average chosen value in cross-validation and varied it in 10% steps to test whether fixing and perturbing λ changes the prediction performance of rCPM.

Data and code availability statement: The HCP data used in this study are publicly available on the ConnectomeDB database (https://db.humanconnectome.org). The PNC data used in this study are publicly available on the database of Genotypes and Phenotypes (dbGaP accession code phs000607.v1.p1); a data access request must be approved to protect the confidentiality of individuals. MATLAB scripts to run the cCPM and rCPM analyses can be found at (https://github.com/YaleMRRC/CPM). BioImage Suite tools used for analysis and visualization can be accessed at (https://bioimagesuiteweb.github.io/webapp/connviewer.html). MATLAB scripts written to perform additional post-hoc analyses are available from the authors upon request.

3. Results

3.1. Combining multiple connectomes improves prediction accuracy compared with single connectome-based prediction

As shown in Fig. 2, all models that incorporate task data significantly predicted fluid intelligence, whereas the models based only on rest did not predict fluid intelligence better than simply predicting the population mean (i.e. R^2_{CV} did not differ from zero). In both datasets, rCPM (HCP: 0.436 ± 0.0072, PNC: 0.356 ± 0.0078) outperformed (HCP: $p = 7.37 * 10^{-4}$, PNC: $p = 0.0721$, corrected resampled t-test comparing rCPM to the next best performing method, GFC-ridge) all competing approaches. GFC-ridge (HCP: 0.387 ± 0.0068, PNC: 0.329 ± 0.151) and cCPM (HCP: 0.386 ± 0.0084, PNC: 0.301 ± 0.0122) performed similarly to each other (HCP: $p = 0.43$, PNC: $p = 0.18$) in both the two datasets and outperformed the other competing methods. The naïve CPM implementation (HCP: 0.354 ± 0.0094, PNC: 0.293 ± 0.0130) and CPM using a GFC matrix (HCP: 0.333 ± 0.0086, PNC: 0.263 ± 0.0151) had similar performance to models built from the best-performing single task,

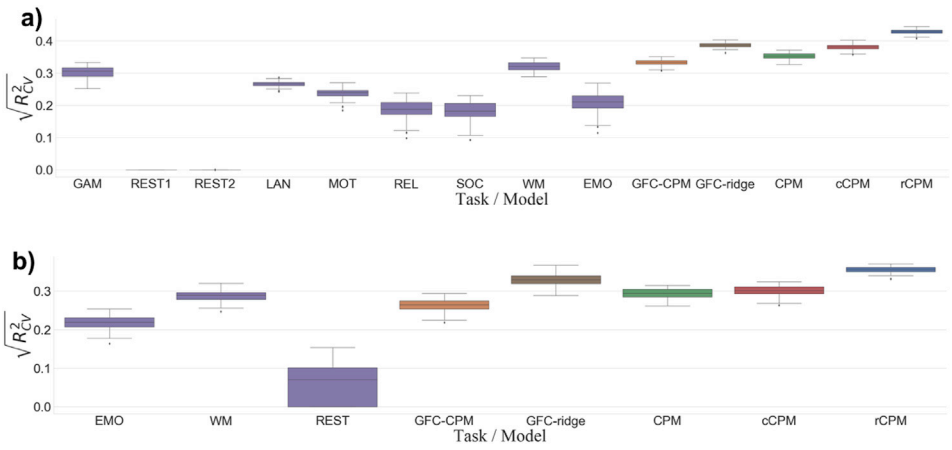

Fig. 2. Comparison of the predictive modeling approaches' ability to predict an individual's gF. a) HCP dataset. **b)** PNC dataset. Purple box plots show results from CPM on a single task. The orange, green, red, and blue box plots show results from combining multiple task connectomes using GFC-CPM, GFC-ridge, CPM, cCPM, and rCPM, respectively. Box plots show cross-validated $\sqrt{R^2_{CV}}$ with the error bars representing the 25th and 75th percentiles, respectively. Values below (or above) the 25th (or 75th percentiles) are shown as *. The best results across different edge selection thresholds are shown. Task acronyms: GAM: Gambling, LAN: Language, MOT: Motor, REL: Relational, SOC: Social, WM: Working Memory, EMO: Emotion.

working memory (HCP: 0.322 ± 0.0134, PNC: 0.288 ± 0.0144). We also tested the performance of ridge regression on single task connectomes. While the overall prediction performance increases when using ridge regression, rCPM still significantly outperforms ridge regression on single task connectomes (HCP: $p = 0.0012$, *PNC:* $p = 0.0146$, comparing rCPM to the next best performing single task with ridge regression).

3.2. Different tasks contribute differentially to the model

As shown in Fig. 3, different tasks contribute different numbers of edges to the final model. Tasks that are more predictive by themselves (e.g. Working Memory) tend to contribute more edges to the model, while less predictive tasks (e.g. Emotion) contribute fewer edges (Fig. 3a). Similarly, in terms of contribution W_m to the regression model, tasks that are predictive by themselves contribute more to the overall predictive model while less-predictive tasks contribute less (Fig. 3b). Additionally, tasks appear to select different edges in different networks for prediction (Fig. 3c). In the figure, percentage of edges in each network pair that are selected by the model are shown. While the presented results are mostly for rCPM (Fig. 3ac), the same trend is observed for cCPM (Fig. S2). As shown in Fig. S3, the predictive utility of a task is not dependent on the length of the task.

Finally, as shown in Fig. 4, using stepwise forward task selection, the second-best-performing individual task was not added to the combined model until step 5 or 6 (for rCPM and cCPM, respectively). Similarly, the best-performing model did not use all tasks. Together, these results suggest that certain tasks may contain redundant information for prediction and that an optimal combination of tasks with complementary information is needed to maximize prediction performance.

3.3. Evaluation of hyperparameters on model performance

As shown in Fig. 5ab, the performance of all approaches varies as a function of the p-value threshold in the edge-selection step. The p-value threshold controls the number of edges retained in the models. A lower p-value threshold represents more stringent edge selection and fewer edges will enter the final model. The GFC-CPM, GFC-ridge, the naïve CPM and cCPM all exhibit a decrease in performance as the p-value threshold is increased (i.e. more retained features), except for cCPM and GFC-ridge on the HCP dataset. In contrast, rCPM exhibits better performance as the p-value threshold is increased. As a result, the improvement in prediction performance offered by rCPM over the competing approaches is at its maximum at a higher p-value threshold (i.e. $p = 0.1$ or $p = 0.5$). However, the extra computation cost induced by more features should also be considered. As shown in Fig. 5c, results from rCPM are insensitive to the choice of λ over the tested range.

3.4. Models trained on one dataset can be generalized to another dataset

Fig. 6 shows that our models generalized to independent, external datasets. By showing the results under different edge sparsity levels, we validate that model generalizability does not decrease with more edges as more features often lead to overfitting. Actually, models based on more edges still outperform models based on fewer edges. GFC's generalizability is also tested here as it can potentially utilize all the available task scans in HCP dataset and apply it to PNC, which has fewer tasks. However, we didn't see improvement in generalizability by including more tasks.

3.5. Exploratory comparison of ridge regression to lasso (least absolute shrinkage and selection operator) and elastic net

Given the improved performance of rCPM, we explored the performance of CPM approaches based on two other regularized regression approaches, lasso and elastic net.

Instead of imposing a L^2-norm penalty on regression coefficient $\boldsymbol{\beta}$ as in ridge regression, lasso uses a L^1-norm penalty as $\lambda \sum_{j=1}^{p} \left|\beta_j\right|$. This encourages coefficients to be set to zero, while ridge regression only shrinks the size of the coefficients. Thus, lasso is often preferred for feature selection when sparsity is preferred. However, lasso tends to select only a small number of variables when the sample size is small. To overcome this limitation, elastic net regularization combines both the L^1-norm penalty and L^2-norm penalty. An α hyperparameter is used to balance between the weight of the two kinds of penalty. So the overall objective function for elastic net can be written as,

$$\widehat{\beta}^{EN} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \left(\frac{(1-\alpha)}{2} \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} \left|\beta_j\right| \right) \right\}$$

For this objective function, the bigger α is, the more lasso-type shrinkage will be put on the coefficients. When $\alpha = 1$, it is the same as lasso; when $\alpha = 0$, it is the same as ridge regression. Implementations of these approaches are identical to rCPM with the exception that the ridge regression step is replaced with either lasso or elastic net.

In our experiments, neither elastic net nor lasso performed as well as ridge regression (HCP: $p = 4.4 * 10^{-5}$, PNC: $p = 0.0035$). For both datasets, $\alpha = 0$ always produces the best prediction (Fig. 7), suggesting—again—the importance of including a large number of edges in a predictive model to best reflect distributed patterns of functional connectivity.

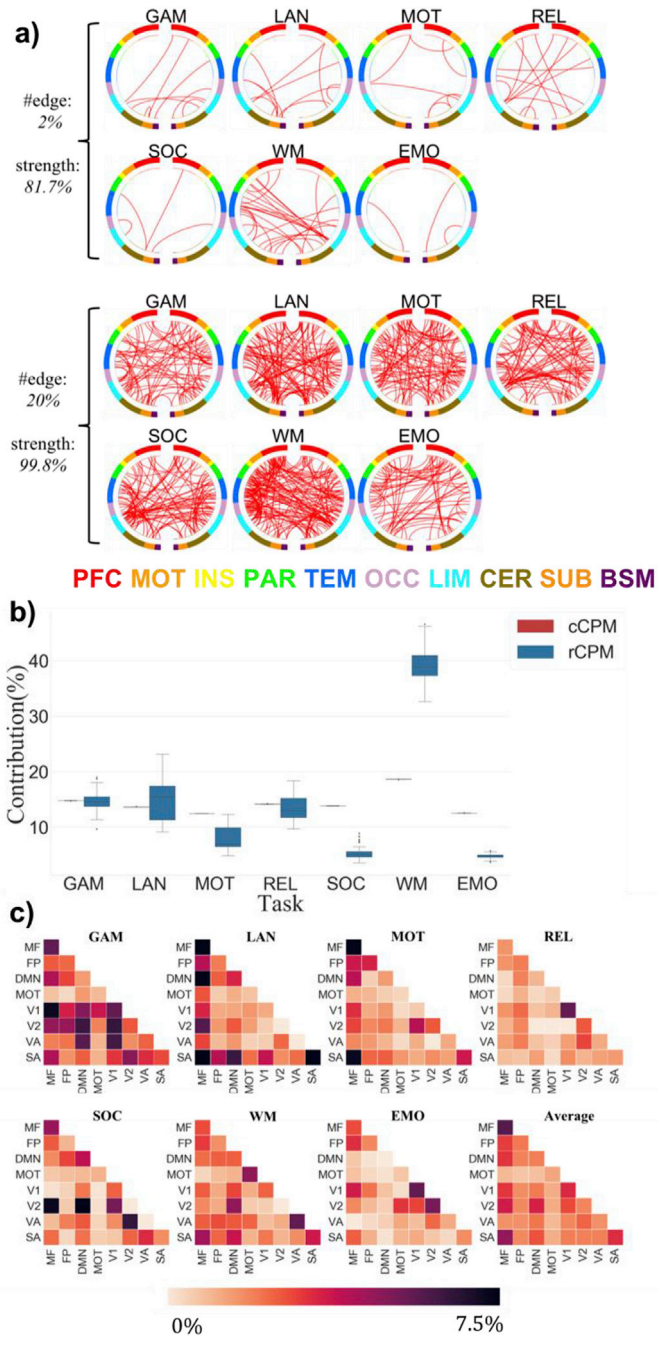

Fig. 3. Different tasks' contributions to the model. a) Visualization of the selected edges for different tasks in the model. Top row represents 2% and bottom row represents 20% of total number of selected edges. 81.7% and 99.8% of feature contribution (combined sum of each feature's regression coefficient times its standard deviation) in regression are possessed respectively by those networks. Anatomical acronyms: PFC = Prefrontal, MOT = MotorStrip, INS = Insula, PAR = Parietal, TEM = Temporal, OCC = Occipital, LIM = Limbic, CER = Cerebellum, SUB = Subcortical, BSM = Brainstem. **b)** Different tasks' average contribution fraction to the cCPM and rCPM model. **c)** Different tasks' contributions to the model, summarized at the network level. Network acronyms: MF = Medial Frontal, FP=Frontoparietal, DMN = Default Mode Network, MOT = Motor Cortex, V1=Visual I, V2=Visual II, VA=Visual Association, SA=Salience.

a)

Step	$\sqrt{R^2_{CV}}$
{WM}	0.315
{WM, MOT}	0.351
{WM, MOT, REL}	0.374
{WM, MOT, REL, EMO}	0.385
{WM, MOT, REL, EMO, GAM}	0.390
{WM, MOT, REL, EMO, GAM, SOC }	**0.391**
{WM, MOT, REL, EMO, GAM, SOC, LAN}	0.380

b)

Step	$\sqrt{R^2_{CV}}$
{WM}	0.358
{WM, MOT}	0.400
{WM, MOT, REL}	0.418
{WM, MOT, REL, LAN}	0.425
{WM, MOT, REL, LAN, SOC}	0.431
{WM, MOT, REL, LAN, SOC, GAM}	**0.433**
{WM, MOT, REL, LAN, SOC, GAM, EMO}	0.431

Fig. 4. Forward task selection for cCPM and rCPM. a) shows the results for cCPM while **b)** shows the results for rCPM. The optimal task combinations for the two algorithms are both using 6 tasks, where cCPM excludes the Language task while rCPM excludes the Emotion task. However, the algorithm performance of using all 7 available tasks is not significantly worse than using 6 tasks (cCPM: $p = 0.38$, rCPM: $p = 0.40$), and overall, including more tasks significantly improves prediction.

4. Discussion

In this paper, we proposed cCPM and rCPM, two general connectome-based prediction frameworks that utilize complementary information in different task connectomes to improve phenotypic prediction. We tested the two algorithms on two open-source datasets, HCP and PNC, to predict fluid intelligence using all the available task connectomes. rCPM shows superior performance in prediction for within-sample prediction compared to the competing methods, including cCPM, though all methods performed similarly for out-of-sample prediction. By looking at the contribution of specific tasks to the final prediction and the stepwise forward optimal task combination selection, we found that different tasks contribute differentially to the final predictive model. In contrast to other competing methods, rCPM performed better when the number of features included in the ridge regression step was large and included features only weakly correlated with intelligence. Last but not least, although the model takes in a large number of features, we validated its generalizability and robustness by using a range of hyperparameters and testing models on external datasets. Overall, cCPM and rCPM provide a powerful framework to combine all available functional imaging data into a single predictive model.

The major contribution of our paper is showing that combining connectomes in an appropriate manner improves prediction. Although task-based connectomes improve prediction performance compared to resting-state connectomes (Greene et al., 2018), approaches to combine multiple task-based connectomes into a single predictive model are limited. The GFC method is one approach to combine task-based connectomes (Elliott et al., 2019). However, by simply averaging over all conditions, GFC loses a large amount of task-specific information, which can be used to increase prediction performance, and performs worse than the other methods that combine multiple connectomes for predictive modeling.

Several regression approaches exist to shrink regression coefficients in the case of highly correlated features, which is common in connectome-based predictive modeling. Two common approaches are principal components regression (PCR), and partial least squares (PLS). Ridge regression shrinks both the high- and low-variance directions of the features, but applies greater shrinkage to the low-variance directions (Krishnan et al., 2011; Mwangi et al., 2014; Zhong et al., 2009) In contrast, PCR does not shrink the high-variance directions and simply discards the lower-variance directions (Frank and Friedman, 1993). PLS also shrinks the low-variance directions, but may inflate the higher variance directions (Frank and Friedman, 1993). Based on these theoretical considerations and additional experimental observations (see for example (Dubois et al., 2018; He et al., 2018)), ridge regression may be

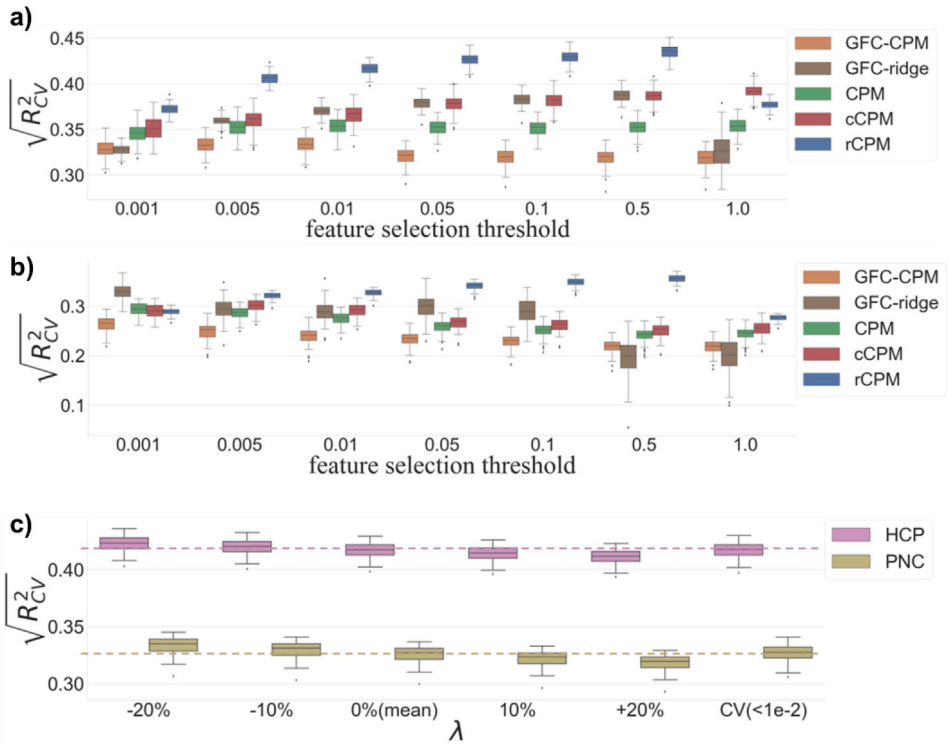

Fig. 5. Models' performance with various hyperparameters. a) Varying edge selection threshold for the HCP dataset. **b)** Varying edge selection threshold for the PNC dataset. **c)** Varying penalty weighting parameter for rCPM. In a) and b) edge selection threshold = 1.0 represents no edge selection. Horizontal line indicates prediction accuracy with λ chosen by inner cross validation.

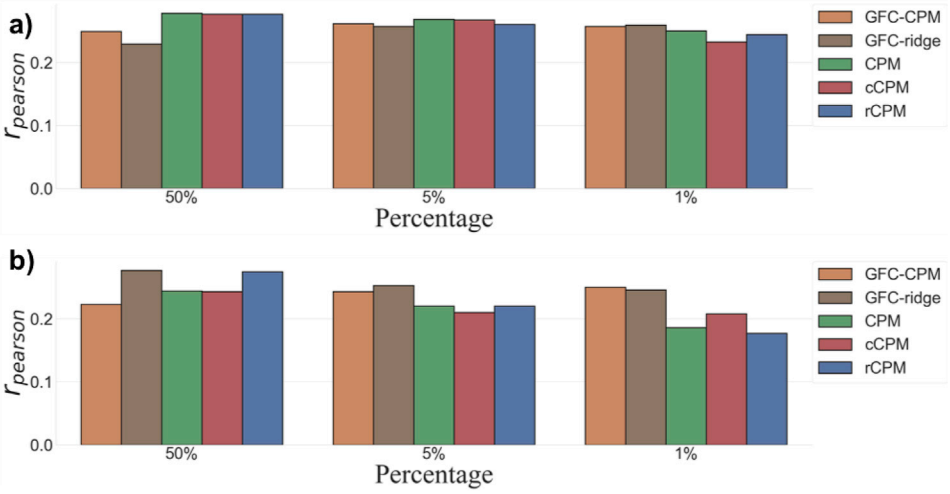

Fig. 6. Different models' generalizability to independent, external datasets. a) Trained on HCP and applied to PNC. **b)** Trained on PNC and applied to HCP. **b)** Trained on PNC and applied to HCP. The results are presented as Pearson correlation between predicted and actual measures. Models trained on either HCP or PNC datasets can significantly predict gF in the other dataset.

preferred for minimizing prediction error because of its smooth shrinkage (Frank and Friedman, 1993), though we did not test these other approaches in the context of connectome-based predictive modeling. Those related approaches will be tested in future work.

Similarly, ridge regression generated better predictions than approaches that perform shrinkage and feature selection, like lasso and elastic net. While popular in neuroimaging machine learning (Dadi et al., 2019), lasso has several limitations compared to ridge regression (Zou and Hastie, 2005). When the number of features is greater than the sample size, lasso limits the number of nonzero features to be the same as the sample size, even though additional features may be associated with the phenotype of interest. Similarly, lasso tends to retain only one feature from any set of highly correlated features, shrinking the other features to zero. In cCPM and rCPM (and other connectome-based predictive modeling approaches), the number of features (i.e. edges) is typically greater than the number of individuals and features are highly correlated. While elastic net regularization attempts to combine the strengths of ridge regression and lasso (Zou and Hastie, 2005), we did not observe any improvements in prediction performance over ridge regression with elastic net.

Despite both approaches incorporating complementary task information, rCPM significantly outperformed cCPM. As only one phenotypic

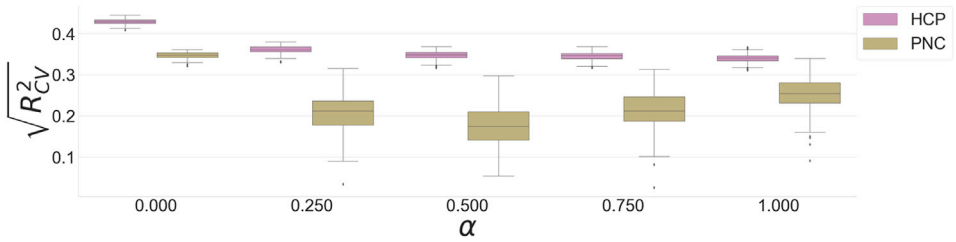

Fig. 7. Comparison of ridge regression with Elastic Net and Lasso. α, the weighting parameter between ridge and lasso-type regularization, is varied across different values. $\alpha = 0$ is the same as ridge regression while $\alpha = 1$ is the same as lasso. Ridge regression generates the most accurate prediction in both the HCP and PNC datasets.

measure was used in these analyses, cCPM simplifies to linear regression. Given this, we chose to follow the CPM design of pooling selected features, rather than constructing a single large matrix with all selected features for linear regression. This single matrix would be rank deficient, resulting in unstable solutions. Ridge regression is a natural answer to this problem as the regularization term allows for stable solutions that minimize the effect of noisy edges on model performance.

RCPM demonstrated improved prediction performance when more features were included in the model. Although rCPM incorporates a modest feature selection step, retaining nearly 60% of the edges (edge number retained with $p < 0.5$) during this step produced the best prediction performance. We attribute these results to ridge regression's ability to shrink noisy and correlated features, thereby, reducing their influence on the model. Moreover, these results suggest that even edges that are not strongly correlated with the behavioral measures still help with prediction, phenotypic information is encoded in spatially distributed connectivity patterns, and (almost) all available information should be used for prediction. These results also align with recent work in "double descent" test risk of predictive models, which suggests a lower test risk when large numbers of features—more than the number of samples—are included (Belkin et al., 2019).

A natural question that arises is: if each task differentially contributes to the final predictive model, what tasks should be included when designing a study? When using a single task, results suggest that tasks and phenotypes of interest should be matched such that the selected task perturbs brain circuits relevant to the phenotype (Greene et al., 2018; Rosenberg et al., 2015). Yet, when combining multiple tasks, including tasks with complementary information appears to be the most beneficial. For example, in the forward feature selection result (Fig. 4), the second task added is the motor task, which by itself is a poorer predictor of fluid intelligence than either the gambling or language task. However, the motor task likely provides more complementary information than the gambling or language task and, thus, provides the maximum gain in prediction power when added to the model. To generalize, we hypothesize that a battery of tasks that perturbs multiple and complementary brain circuits will yield better predictive models than a battery of tasks that perturbs a single brain circuit with different, subtle manipulations. While it may not be possible to collect a single battery of tasks that is optimal for all phenotypic information, we suspect that a standard battery could be developed that is good enough in most cases. Future work should develop and compare different batteries of tasks in terms of predictive modeling.

This work has some limitations. First, currently for cCPM and rCPM, all individuals are required to have complete data from all the tasks. As such, using cCPM and rCPM increases the likelihood of removing individuals from the analysis due to missing data. As the utility of predictive models is often dependent on sample size (Cui and Gong, 2018; Varoquaux et al., 2017), cCPM and rCPM may only be applicable to larger datasets that can support removing individuals due to missing data. Future work includes extending cCPM and rCPM to support data imputation methods to handle missing data. Second, cCPM and rCPM may make model interpretation and visualization more challenging. While cCPM and rCPM retain CPM's ability to simply map features back to the brain, the mappings from both of the two algorithms are task specific. Ultimately, the task at hand will determine if the added task information reduces interpretability of the model (see Rule #10 from (Scheinost et al., 2019) for a greater discussion). RCPM exhibited increased prediction performance when the number of features was larger, and while better prediction is good, the large number of edges can be difficult to visualize and interpret. It should be noted however that this may be a more accurate depiction of the complex systems involved in executing cognitive functions, and our tendency to reduce findings to one or two brain regions is likely grossly oversimplifying this complex system. In other words, the circuits involved in cognition may not lend themselves to easy visualization. Third, cCPM and rCPM can make external validation harder. The external dataset needs to have similar tasks for the model to be applicable. For example, when using external validation between the HCP and the PNC datasets, we limited the tasks from the HCP to only two for compatibility with the PNC dataset. Fourth, cCPM and rCPM did not show improvement for out-of-sample prediction. This result could suggest that cCPM and rCPM overfit for within-sample prediction or that the HCP and PNC datasets differ in important aspects, such as age, that limit generalizability of any model (see Rule #9 from (Scheinost et al., 2019)). Fifth, similarly, due to the large number of edges (features) involved in the prediction models compared with the sample size, overfitting is hard to eliminate completely (Whelan and Garavan, 2014) and this causes the difference between our results with the previous works (Finn et al., 2015). However, with the use of both cross-validation and independent dataset generalization, our evaluation results are less-prone to suffer from overestimation. Finally, while we focus on connectivity data derived from multiple fMRI tasks, cCPM and rCPM are agnostic to the type of input data and can easily incorporate structural connectivity data from DTI or from other functional modalities like EEG. In future work, we will explore the inclusion of other measures to further improve prediction.

In summary, we present cCPM and rCPM, two extensions to CPM to handle connectomes from multiple sources. Our results suggest that prediction of phenotypic measures can be improved by including multiple task conditions in computational models, that different tasks provide complementary information for prediction, and that cCPM and rCPM provide two principled methods for modeling such data.

Acknowledgements

Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The remainder of the data used in this study were provided by the Philadelphia Neurodevelopmental Cohort (Principal Investigators: Hakon Hakonarson and Raquel Gur; phs000607.v1.p1). Support for the collection of the data sets was provided by NIH grant RC2MH089983 awarded to Raquel Gur and RC2MH089924 awarded to Hakon

Hakonarson. All subjects were recruited through the Center for Applied Genomics at The Children's Hospital in Philadelphia. This work was also supported by NIH R01 MH111424, P40 MH115716, R24 MH114805, and T32GM007205 (A.S.G.).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.116038.

References

Belkin, M., Hsu, D., Xu, J., 2019. Two Models of Double Descent for Weak Features.

Bilker, W.B., Hansen, J.A., Brensinger, C.M., Richard, J., Gur, R.E., Gur, R.C., 2012. Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. Assessment. https://doi.org/10.1177/1073191112446655.

Bouckaert, R.R., Frank, E., 2010. Evaluating the Replicability of Significance Tests for Comparing Learning Algorithms. https://doi.org/10.1007/978-3-540-24775-3_3.

Cui, Z., Gong, G., 2018. The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. Neuroimage 192, 115–134. https://doi.org/10.1016/j.neuroimage.2018.06.001.

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., Varoquaux, G., 2019. Benchmarking functional connectome-based predictive models for resting-state fMRI. Neuroimage. https://doi.org/10.1016/j.neuroimage.2019.02.062.

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., Varoquaux, G., 2018. Benchmarking functional connectome-based predictive models for resting-state fMRI. HAL-inra. https://doi.org/10.1016/J.NEUROIMAGE.2019.02.062.

Dosenbach, N.U.F., Nardos, B., Cohen, A.L., Fair, D.A., Power, J.D., Church, J.A., Nelson, S.M., Wig, G.S., Vogel, A.C., Lessov-Schlaggar, C.N., Barnes, K.A., Dubis, J.W., Feczko, E., Coalson, R.S., Pruett, J.R., Barch, D.M., Petersen, S.E., Schlaggar, B.L., 2010. Prediction of individual brain maturity using fMRI. Science 80. https://doi.org/10.1126/science.1194144.

Dubois, J., Adolphs, R., 2016. Building a science of individual differences from fMRI. Trends Cogn. Sci. https://doi.org/10.1016/j.tics.2016.03.014.

Dubois, J., Galdi, P., Paul, L.K., Adolphs, R., 2018. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. Philos. Trans. R. Soc. Biol. Sci. https://doi.org/10.1098/rstb.2017.0284.

Elliott, M.L., Knodt, A.R., Cooke, M., Kim, M.J., Melzer, T.R., Keenan, R., Ireland, D., Ramrakha, S., Poulton, R., Caspi, A., Moffitt, T.E., Hariri, A.R., 2019. General functional connectivity: shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. Neuroimage 189, 516–532. https://doi.org/10.1016/j.neuroimage.2019.01.068.

Finn, E.S., Scheinost, D., Finn, D.M., Shen, X., Papademetris, X., Constable, R.T., 2017. Can brain state be manipulated to emphasize individual differences in functional connectivity? Neuroimage 160, 140–151. https://doi.org/10.1016/j.neuroimage.2017.03.064.

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., Constable, R.T., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. 18, 1664–1671. https://doi.org/10.1038/nn.4135.

Frank, L.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. Technometrics. https://doi.org/10.1080/00401706.1993.10485033.

Gao, S., Greene, A.S., Constable, R.T., Scheinost, D., 2018a. Task integration for connectome-based prediction via canonical correlation analysis. In: Proceedings - International Symposium on Biomedical Imaging. https://doi.org/10.1109/ISBI.2018.8363529.

Gao, S., Greene, A.S., Todd Constable, R., Scheinost, D., 2018b. Combining multiple connectomes via canonical correlation analysis improves predictive models. In: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). https://doi.org/10.1007/978-3-030-00931-1_40.

Greene, A.S., Gao, S., Scheinost, D., Constable, R.T., 2018. Task-induced brain state manipulation improves prediction of individual traits. Nat. Commun. https://doi.org/10.1038/s41467-018-04920-3.

He, T., Kong, R., Holmes, A.J., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., Yeo, B.T.T., 2018. Is deep learning better than kernel regression for functional connectivity prediction of fluid intelligence?. In: 2018 International Workshop on Pattern Recognition in Neuroimaging, PRNI 2018 https://doi.org/10.1109/PRNI.2018.8423958.

Krishnan, A., Williams, L.J., McIntosh, A.R., Abdi, H., 2011. Partial Least Squares (PLS) methods for neuroimaging: a tutorial and review. Neuroimage. https://doi.org/10.1016/j.neuroimage.2010.07.034.

Moore, T.M., Reise, S.P., Gur, R.E., Hakonarson, H., Gur, R.C., 2015. Psychometric properties of the penn computerized neurocognitive battery. Neuropsychology. https://doi.org/10.1037/neu0000093.

Mwangi, B., Tian, T.S., Soares, J.C., 2014. A review of feature reduction techniques in Neuroimaging. Neuroinformatics. https://doi.org/10.1007/s12021-013-9204-3.

Rosenberg, M.D., Casey, B.J., Holmes, A.J., 2018. Prediction complements explanation in understanding the developing brain. Nat. Commun. https://doi.org/10.1038/s41467-018-02887-9.

Rosenberg, M.D., Finn, E.S., Scheinost, D., Papademetris, X., Shen, X., Constable, R.T., Chun, M.M., 2015. A neuromarker of sustained attention from whole-brain functional connectivity. Nat. Neurosci. https://doi.org/10.1038/nn.4179.

Satterthwaite, T.D., Connolly, J.J., Ruparel, K., Calkins, M.E., Jackson, C., Elliott, M.A., Roalf, D.R., Hopson, R., Prabhakaran, K., Behr, M., Qiu, H., Mentch, F.D., Chiavacci, R., Sleiman, P.M.A., Gur, R.C., Hakonarson, H., Gur, R.E., 2016. NeuroImage the Philadelphia Neurodevelopmental Cohort : a publicly available resource for the study of normal and abnormal brain development in youth. Neuroimage 124, 1115–1119. https://doi.org/10.1016/j.neuroimage.2015.03.056.

Scheinost, D., Noble, S., Horien, C., Greene, A.S., Lake, E.M., Salehi, M., Gao, S., Shen, X., O'Connor, D., Barron, D.S., Yip, S.W., Rosenberg, M.D., Constable, R.T., 2019. Ten simple rules for predictive modeling of individual differences in neuroimaging. Neuroimage 193, 35–45. https://doi.org/10.1016/J.NEUROIMAGE.2019.02.057.

Shen, X., Finn, E.S., Scheinost, D., Rosenberg, M.D., Chun, M.M., Papademetris, X., Constable, R.T., 2017. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nat. Protoc. https://doi.org/10.1038/nprot.2016.178.

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K., Consortium, W.-M.H.C.P., 2013. The Wu-Minn human connectome project: an overview. Neuroimage 80, 62–79.

Vanderwal, T., Eilbott, J., Finn, E.S., Craddock, R.C., Turnbull, A., Castellanos, F.X., 2017. Individual differences in functional connectivity during naturalistic viewing conditions. Neuroimage. https://doi.org/10.1016/j.neuroimage.2017.06.027.

Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., Thirion, B., 2017. Assessing and Tuning Brain Decoders: Cross-Validation, Caveats, and Guidelines. Neuroimage. https://doi.org/10.1016/j.neuroimage.2016.10.038.

Whelan, R., Garavan, H., 2014. When optimism hurts: inflated predictions in psychiatric neuroimaging. Biol. Psychiatry. https://doi.org/10.1016/j.biopsych.2013.05.014.

Zhong, Y., Wang, H., Lu, G., Zhang, Z., Jiao, Q., Liu, Y., 2009. Detecting functional connectivity in fmri using pca and regression analysis. Brain Topogr. https://doi.org/10.1007/s10548-009-0095-4.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Ser. Soc. B Stat. Methodol. https://doi.org/10.1111/j.1467-9868.2005.00503.x.