2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)
April 4-7, 2018, Washington, D.C., USA

TASK INTEGRATION FOR CONNECTOME-BASED PREDICTION VIA CANONICAL CORRELATION ANALYSIS

Siyuan Gao[⋆], *Abigail S. Greene*[†], *R. Todd Constable* [‡], *Dustin Scheinost*[‡]

[⋆]Department of Biomedical Engineering, Yale University, New Haven, CT
[†]Interdepartmental Neuroscience Program, Yale School of Medicine, New Haven, CT
[‡]Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT

ABSTRACT

Generating models from functional connectivity data that predict behavioral measures holds great clinical potential. While the majority of the literature has focused on using only connectivity data from a single source, there is ample evidence that different cognitive conditions amplify individual differences in functional connectivity in a distinct, complementary manner. In this work, we introduce a computational model, labeled multidimensional Connectome-based Predictive Modeling (mCPM), that combines connectivity matrices collected from different task conditions in order to improve behavioral prediction by using complementary information found in different cognitive tasks. As proof of concept, we apply our algorithm to data from the Human Connectome Project. Using data from seven different tasks, mCPM generated models that better predicted fluid intelligence than models generated from any single task. Our results suggest that prediction of behavior can be improved by including multiple task conditions in computational models, that different tasks provide complementary information for prediction, and that mCPM provides a principal method for modeling such data.

1. INTRODUCTION

Advanced functional magnetic resonance imaging (fMRI) techniques, particularly functional connectivity (FC) analyses, are revealing robust individual differences in patterns of neural activity that predict continuous behavioral measures and clinical symptoms [1–4]. While FC is usually calculated from data acquired during resting-state, task conditions can perturb specific cognitive circuits in ways that can better reveal individual differences and improve behavioral prediction [5, 6]. However, depending on the behavior under study, different tasks may bring out different meaningful information. That is to say that a particular task may be better for generating predictive models for a specific behavior but worse for other behaviors. Further, different cognitive tasks may provide different complementary information, such that information from multiple tasks may be needed to achieve the best prediction. Thus, methods that incorporate FC information from a spectrum of cognitive tasks into a single predictive model may represent the best performing and most generalizable methods for prediction of behavior from FC data.

In order to principally combine different task FC matrices into a single predictive model, we propose a novel algorithm based on canonical correlation analysis (CCA) [7] and our Connectome-based Predictive Modeling (CPM) method [8], labeled multidimensional CPM (mCPM). To evaluate mCPM, we use data from the Human Connectome Project (HCP) [9] consisting of 515 subjects with task fMRI from seven different tasks. We show that models created by mCPM based on all seven tasks result in superior prediction of fluid intelligence compared to models built using a single task. These results exhibit the existence of complementary information in different tasks, highlight an opportunity to use multiple task FC matrices to more comprehensively characterize individual differences, and suggest the ability of mCPM to combine this information to maximize predictive power.

This paper is organized as follows. Section 2 summarizes related work. Section 3 introduces our mCPM algorithm. Section 4 describes experimental methods using mCPM to predict fluid intelligence from task connectivity matrices using subjects from the HCP. Section 5 presents our results. Finally section 6 offers some concluding remarks.

2. RELATED WORK

Development and application of predictive models of behavior from FC data is a mature area of research with many proposed approaches. The majority of this work has focused on binary or categorical classification of patient groups. A non-exhaustive list includes applications to Alzheimer's disease [10], attention deficit hyperactivity disorder [11], schizophrenia [12], depression [13], and autism [14]. Prediction of binary categorical outcomes—rather than dimensional outcomes—can misclassify subclinical or prodromal individuals and provide poor measures of disease or treatment progression (i.e., patients are classified the same regardless of changes in symptoms). In contrast, prediction of dimensional outcomes from FC data is an emerging field and is

978-1-5386-3636-7/18/$31.00

considerably more challenging than binary classification of disease state [8]. Association between FC and behavior in healthy participants have substantially lower effect sizes than differences due to disease. Additionally, prediction of continuous variables requires correct modeling over the whole range of the behavioral measure, whereas classification of binary groups largely requires correct grouping of participants near the margin. When participants are far from the margin, the correct classification is often trivial. Previous approaches to prediction of dimensional outcomes from FC data include support vector regression [2], elastic nets [3], pooled edge strength and linear models [4, 8], and partial least squares regression [15]. However, these approaches only use FC data from a single condition (typically resting-state) rather than FC data from multiple conditions.

3. MULTIDIMENSIONAL CONNECTOME-BASED PREDICTIVE MODELING (MCPM)

3.1. Overview

mCPM is an extension of our CPM algorithm to handle FC data from multiple sources. mCPM uses CCA to find orthogonal information from multiple connectivity matrices in order to maximize the correlation between connectivity and the behavioral measure. Here, we focus on connectivity data derived from multiple fMRI tasks. However, the method is general to the type of connectivity data and could easily incorporate structural connectivity data from diffusion tensor imaging or FC data from other modalities such as EEG.

3.2. Connectome-based Predictive Modeling (CPM)

CPM [8] is a validated method for extracting and pooling the most relevant features from connectivity data in order to construct linear models to predict behavioral measures. Briefly, edges of connectivity matrices that are significantly correlated with the behavior of interest are selected. The selected features are then pooled (*e.g.* averaged) and linear regression is used to predict the behavior in novel participants.

3.3. Canonical Correlation Analysis (CCA)

For two sets of observation matrices $\mathbf{X}$ and $\mathbf{Y}$, assuming that the variables are correlated, CCA seeks linear combinations of the columns of these two matrices that maximize correlation between them. In other words, we want to find vectors $\mathbf{a}$ and $\mathbf{b}$ such that the random variables $\mathbf{Xa}$ and $\mathbf{Yb}$ maximize the correlation. Assuming that $\mathbf{X}$ and $\mathbf{Y}$ are normalized such that each column of either matrix has mean zero and unit variance, the correlation to be maximized can be expressed by the following equation:

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

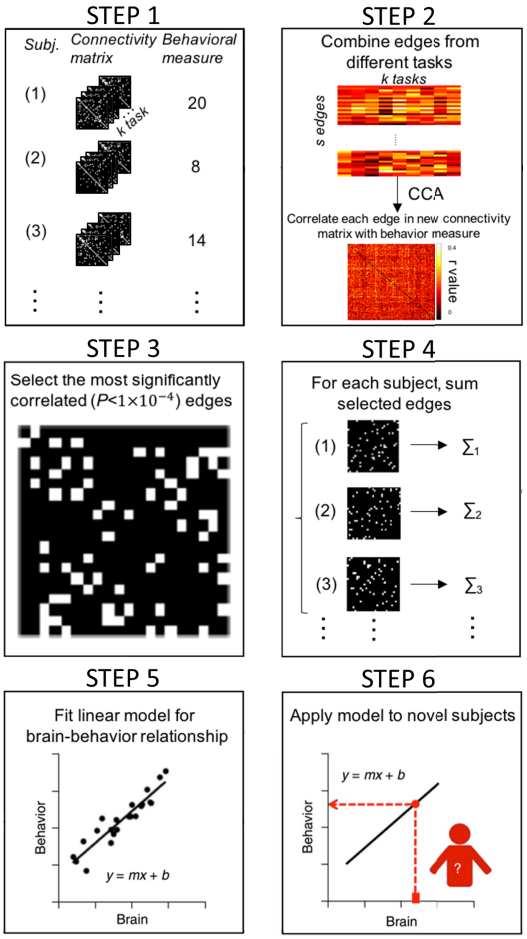

Fig. 1: Schematic of mCPM. Inputs to mCPM are connectivity matrices from multiple sources and behavioral measures. Step 1: Participants are divided into training and testing sets. Step 2: Across all participants in the training set, the same edge from different tasks are combined via CCA and correlated with behavior. Step 3: Significant edges are selected for further analysis. Step 4: For each participant, the significant edges are then pooled into a summary value of FC. Step 5: A linear regression model is built between the summary value of FC and the behavioral measure. Step 6: Summary values of FC are calculated for each participant in the testing set. This value is then inputted into the linear regression model. The resulting value is the predicted behavioral measure for the current test participant. Figure modified with permission from [8]

3.4. mCPM pipeline

The mCPM pipeline consists of six steps (Fig. 1). In the first step, participants are divided into training and testing sets for cross-validation. We use leave-one-out cross-validation as default. The second step is the combination of task FC edges. For the (i, j) edge, we have a matrix $\mathbf{E}_{i,j} \in \mathbf{R}^{(n-1)*t}$. Rows of the matrix denote each participant's (i, j) edge's different strength under t different tasks. Using CCA, we can find the canonical coefficients $\mathbf{w}_{i,j} \in \mathbf{R}^t$ for each edge. As each edge matrix $\mathbf{E}_{i,j}$ corresponds to the observation matrix $\mathbf{X}$ in Sec-

tion 3.3's definition equation for CCA, these coefficients $\mathbf{w}_{i,j}$ corresponds to the vector **a**, and the observation matrix **Y** will store the behavioral measures. We then combine FC matrices from all tasks into a total connectivity matrix using different canonical correlations. For $n-1$ subjects combined (i,j) edge, $\mathbf{E}_{total_{i,j}} = \sum_t \mathbf{E}^t_{i,j} w^t_{i,j}$, where the t-th column of $\mathbf{E}_{i,j}$, $\mathbf{E}^t_{i,j}$ contains (i,j) edge from $n-1$ subjects under the t-th task and each edge is demeaned across different participants and within the single task. For the third step, we assign the significantly correlated edges 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, instead of minimize, the correlation. Various significance thresholds can be used. In the fourth step, we calculate "network strength" $\mathbf{s}^{CN}$ by pooling (*i.e.* summing) the strength of all CN edges in each participant's total connectivity matrix, yielding a summary value s^{CN}_k for k-th participant.

$$\mathbf{s}^{CN} = \sum_{i,j} m^+_{i,j} \sum_t \mathbf{E}^t_{i,j} w^t_{i,j}$$

Where $\mathbf{s}^{CN}$ is the vector for summary values, $\mathbf{m}^+$ and $\mathbf{m}^-$ are binary matrices indexing the edges (i,j) that survived thresholding for CN. In the fifth step, we use linear regression to model the association between "network strength" and the behavioral measure in $n-1$ participants.

$$y_{behav_k} = \beta_0 * s^{CN}_k + \beta_1$$

In the sixth step, the "network strength" is calculated for the excluded participant(s), and is submitted to the corresponding regression model to generate a behavioral measure estimate for that participant. This process is repeated iteratively, with different participants in the training and testing dataset.

4. EXPERIMENT

We applied mCPM to FC data from the Human Connectome Project (HCP). After excluding participants for high motion or incomplete data, 515 subjects were retained for analysis. For proof of concept, we used fluid intelligence as the behavior of interest for prediction. However, mCPM can handle multiple behavioral measures. With a single behavioral measure mCPM simplifies to multi-linear regression. The seven task scans (gambling, language, motor, relational, social, working memory, and emotion) were processed with standard methods and parcellated into 268 nodes using a whole-brain, functional atlas defined previously in a separate sample [16]. Task FC was calculated based on the "raw" task timecourses, with no regression of task-evoked activity. Next, the mean timecourses of each node pair were correlated and correlation coefficients were Fisher transformed, generating seven 268x268 connectivity matrices per subject. We performed both CPM and mCPM on these matrices to generate cross-validated predictive models of fluid intelligence from whole-brain patterns of FC. Model performance was quantified as the Pearson correlation coefficient between predicted and true fluid intelligence (r) and mean squared error (MSE). Because the feature-selection thresholds used in CPM and mCPM are inevitably arbitrary, we tested five different thresholds. Except as otherwise noted, all reported results were generated using a feature-selection threshold of $p < 10^{-4}$.

5. RESULTS

5.1. mCPM improves prediction accuracy

All models from either CPM or mCPM produced significantly better predictions than chance ($p < 0.05$, permutation test with $5,000$ iterations). Using all the task FC data, mCPM generated models that achieved superior predictions compared to models generated by CPM based on a single task FC (see table 1). Using Steiger's test to compare two dependent correlations, predictions using the mCPM model were significantly ($p < 0.05$) greater than the best performing CPM model from a single task (*i.e.* the Gambling task).

Task	Feature selection threshold		
	1E−2	1E−3	1E−4
Gambling	0.359\17.63	0.337\17.92	0.342\17.63
Language	0.290\18.44	0.294\18.35	0.324\17.87
Motor	0.287\18.83	0.302\18.49	0.271\18.94
Relational	0.241\19.84	0.186\21.00	0.225\20.11
Social	0.277\19.04	0.256\19.34	0.209\20.11
WM	0.340\18.08	0.326\18.29	0.297\18.66
Emotion	0.296\18.83	0.324\18.23	0.235\19.52
mCPM	**0.401\17.01**	**0.384\17.41**	**0.396\17.16**

Table 1: Comparison of prediction performance of CPM using FC matrices from a single task and mCPM using FC matrices from all 7 tasks for different feature selection thresholds. Reported values are the correlation and MSE between predicted and observed fluid intelligence. mCPM (bolded) produced best performance.

5.2. Different tasks contribute differentially to prediction

To better understand the influence of each task on the mCPM results, we calculated the average score of each task involved in the prediction. First, we calculated the weighted sums of the selected edges, where weights $\mathbf{w}_{i,j}$ are determined by CCA. Then the average score was defined as the mean of those weighted sums across different participants.

$$as_t = \frac{\sum_{i,j} \mathbf{E}^t_{i,j} * \mathbf{w}^t_{i,j}}{n}$$

As shown in Fig. 2, all tasks provide positive contributions to our prediction. For all tasks, larger selected edge strengths correspond to larger predicted values of fluid intelligence. The largest portion of the summary statistic $\mathbf{s}^{CN}$ was composed of edge strength from the Language task. Interestingly, the Language task was near the middle of individual task prediction performance. Together, these suggest that the Language task, while not the individually most predictive task, contains a larger amount of unique variance when combined with the other tasks.

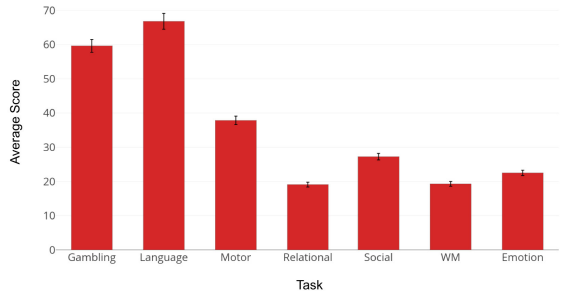

Fig. 2: Average score of different tasks. The error bar is showing the standard error. The language task shows the largest score, highlighting its importance in prediction.

5.3. Different thresholds generate similar results

Table 1 shows that mCPM is not significantly influenced by different, arbitrary thresholds for edge selection.

5.4. A subset of tasks improves mCPM performance

Certain tasks may contain redundant information for prediction. We adopted forward feature selection method to select the optimal combination of tasks. The optimal set of tasks were the Emotion, Social, Relational, Gambling, and Language tasks. The resulting correlation between observed and predicted fluid intelligence was $r = 0.4078$

5.5. Visualize selected connectivity features

We selected the edges that appeared in every cross-validation iteration for visualization using BioImage Suite (Fig. 3). In line with previous reports [1], predictive edges are mainly located in frontal-parietal networks.

6. SUMMARY

We proposed mCPM to combine connectivity matrices from multiple sources and find that it produces better prediction than using a single source. Using HCP fMRI data to predict fluid intelligence, we showed that predictive models built from FC matrices from multiple tasks result in superior prediction than models built using any single task. We further verified that our algorithm is robust under different feature selection thresholds. These results exhibit the existence of orthogonal information in different tasks, highlight an opportunity to use multiple task FC matrices to more comprehensively characterize individual differences, and suggest the ability of mCPM to combine this information to maximize predictive power. Future work will include incorporating structural connectivity data from diffusion tensor imaging into mCPM, testing our algorithm on other large open-source datasets, and bagging procedures to optimally select the best combinations of data sources.

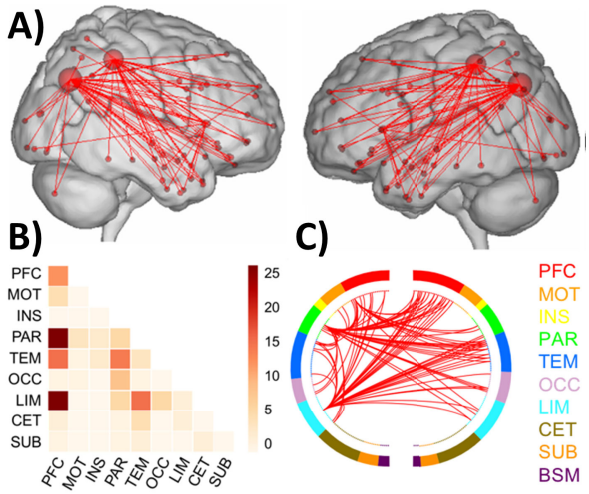

Fig. 3: Visualization of predictive edges. Edges are selected with a threshold of $P < 1 \times 10^{-4}$. **A)** Brain plots with each node represented as a sphere, where the size of the sphere indicates the number of selected edges connected to that node. **B)** Matrix plots: rows and columns represent canonical networks. The cells represent the total number of selected edges connecting the nodes in the two networks. **C)** Circle Plots: Nodes are arranged in two half circles approximately reflecting brain anatomy from anterior to posterior, and the nodes are color coded according to the cortical lobes. Lobes are prefrontal (PFC), motor (MOT), insula (INS), parietal (PAR), temporal (TEM), occipital (OCC), limbic (LIM), cerebellum (CER), subcortical (SUB), and brain stem (BSM).

7. ACKNOWLEDGEMENTS

Data were provided in part by the Human Connectome Project, WU-MinnConsortium (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.

References

[1] Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable, "Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity," *Nature Neuroscience*, vol. 18, no. 11, pp. 1664–1671, 2015.

[2] Nico U. F. Dosenbach, Binyam Nardos, Alexander L. Cohen, Damien A. Fair, Jonathan D. Power, Jessica A. Church, Steven M. Nelson, Gagan S. Wig, Alecia C. Vogel, Christina N. Lessov-Schlaggar, Kelly Anne Barnes, Joseph W. Dubis, Eric Feczko, Rebecca S. Coalson, John R. Pruett, Deanna M. Barch, Steven E. Petersen, and Bradley L. Schlaggar, "Prediction of individual brain maturity using fmri," *Science*, vol. 329, no. 5997, pp. 1358–1361, 2010.

[3] Stephen M Smith, Thomas E Nichols, Diego Vidaurre, Anderson M Winkler, Timothy EJ Behrens, Matthew F Glasser, Kamil Ugurbil, Deanna M Barch, David C Van Essen, and Karla L Miller, "A positive-negative mode of population covariation links brain connectivity, demographics and behavior," *Nature neuroscience*, vol. 18, no. 11, pp. 1565–1567, 2015.

[4] Monica D Rosenberg, Emily S Finn, Dustin Scheinost, Xenophon Papademetris, Xilin Shen, R Todd Constable, and Marvin M Chun, "A neuromarker of sustained attention from whole-brain functional connectivity," *Nature Neuroscience*, vol. 19, no. 1, pp. 165–171, 2016.

[5] Emily S Finn, Dustin Scheinost, Daniel M Finn, Xilin Shen, Xenophon Papademetris, and R Todd Constable, "Can brain state be manipulated to emphasize individual differences in functional connectivity?," *NeuroImage*, 2017.

[6] Tamara Vanderwal, Jeffrey Eilbott, Emily S. Finn, R. Cameron Craddock, Adam Turnbull, and F. Xavier Castellanos, "Individual differences in functional connectivity during naturalistic viewing conditions," *NeuroImage*, vol. 157, no. Supplement C, pp. 521 – 530, 2017.

[7] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.

[8] Xilin Shen, Emily S Finn, Dustin Scheinost, Monica D Rosenberg, Marvin M Chun, Xenophon Papademetris, and R Todd Constable, "Using connectome-based predictive modeling to predict individual behavior from brain connectivity," *nature protocols*, vol. 12, no. 3, pp. 506–518, 2017.

[9] David C Van Essen, Stephen M Smith, Deanna M Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al., "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[10] Gang Chen, B. Douglas Ward, Chunming Xie, Wenjun Li, Zhilin Wu, Jennifer L. Jones, Malgorzata Franczak, Piero Antuono, and Shi-Jiang Li, "Classification of alzheimer disease, mild cognitive impairment, and normal cognitive status with large-scale network analysis based on resting-state functional mr imaging," *Radiology*, vol. 259, no. 1, pp. 213–221, 2011, PMID: 21248238.

[11] Matthew Brown, Gagan Sidhu, Russell Greiner, Nasimeh Asgarian, Meysam Bastani, Peter Silverstone, Andrew Greenshaw, and Serdar Dursun, "Adhd-200 global competition: diagnosing adhd using personal characteristic data can outperform resting state fmri measurements," *Frontiers in Systems Neuroscience*, vol. 6, pp. 69, 2012.

[12] Mohammad Arbabshirani, Kent Kiehl, Godfrey Pearlson, and Vince Calhoun, "Classification of schizophrenia patients based on resting-state functional network connectivity," *Frontiers in Neuroscience*, vol. 7, pp. 133, 2013.

[13] Ling-Li Zeng, Hui Shen, Li Liu, Lubin Wang, Baojuan Li, Peng Fang, Zongtan Zhou, Yaming Li, and Dewen Hu, "Identifying major depression using whole-brain functional connectivity: a multivariate pattern analysis," *Brain*, vol. 135, no. 5, pp. 1498–1507, 2012.

[14] Mark Plitt, Kelly Anne Barnes, and Alex Martin, "Functional connectivity classification of autism identifies highly predictive brain features but falls short of biomarker standards," *NeuroImage: Clinical*, vol. 7, no. Supplement C, pp. 359 – 366, 2015.

[15] Kosuke Yoshida, Yu Shimizu, Junichiro Yoshimoto, Masahiro Takamura, Go Okada, Yasumasa Okamoto, Shigeto Yamawaki, and Kenji Doya, "Prediction of clinical depression scores and detection of changes in whole-brain using resting-state functional mri data with partial least squares regression," *PLoS ONE*, vol. 12, no. 7, pp. e0179638, 2017.

[16] Xilin Shen, Fuyuze Tokoglu, Xenios Papademetris, and R Todd Constable, "Groupwise whole-brain parcellation from resting-state fmri data for network node identification," *Neuroimage*, vol. 82, pp. 403–415, 2013.