Do Deep Neural Networks Outperform Kernel Regression for Functional Connectivity Prediction of Behavior?

Tong He[1,2], Ru Kong[1,2], Avram J. Holmes[3], Minh Nguyen[1,2], Mert R. Sabuncu[4], Simon B. Eickhoff[5,6], Danilo Bzdok[7,8,9], Jiashi Feng[2], B.T. Thomas Yeo[1,2,10,11,12]

[1] ASTAR-NUS Clinical Imaging Research Centre, Singapore Institute for Neurotechnology and Memory Networks Program, National University of Singapore, Singapore [2] Department of Electrical and Computer Engineering, National University of Singapore, Singapore [3] Departments of Psychology and Psychiatry, Yale University, New Haven, CT, USA [4] School of Electrical and Computer Engineering, Cornell University, Ithaca, NY, USA [5] Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany [6] Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Centre Jülich, Jülich, Germany [7] Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, Germany [8] JARA-BRAIN, Jülich-Aachen Research Alliance, Germany [9] Parietal team, INRIA, Neurospin, bat 145, CEA Saclay, 91191 Gif-sur-Yvette, France [10] Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA [11] Centre for Cognitive Neuroscience, Duke-NUS Medical School, Singapore [12] NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, Singapore

Address correspondence to:

B.T. Thomas Yeo

ECE, ASTAR-NUS CIRC, SINAPSE & MNP

National University of Singapore

Email: thomas.yeo@nus.edu.sg

Abstract

There is significant interest in the development and application of deep neural networks (DNNs) to neuroimaging data. A growing literature suggests that DNNs outperform their classical counterparts in a variety of neuroimaging applications, yet there are few direct comparisons of relative utility. Here, we compared the performance of three DNN architectures and a classical machine learning algorithm (kernel regression) in predicting individual phenotypes from whole-brain resting-state functional connectivity (RSFC) patterns. One of the DNNs was a generic fully-connected feedforward neural network, while the other two DNNs were recently published approaches specifically designed to exploit the structure of connectome data. By using a combined sample of almost 10,000 participants from the Human Connectome Project (HCP) and UK Biobank, we showed that the three DNNs do not outperform kernel regression across a wide range of behavioral and demographic measures. Furthermore, the generic feedforward neural network exhibited similar performance to the two state-of-the-art connectome-specific DNNs. We conclude with suggestions on future neuroimaging DNN research, including comparisons with stronger baseline algorithms, minimum sample sizes, transparency of hyperparameter tuning and code availability. Critically, we believe that deep learning remains a promising tool for analyzing neuroimaging data. However, researchers should carefully consider whether and how their applications might benefit from DNNs' advantages over classical alternatives, rather than treat deep learning as a panacea.

Keywords:

Behavioral prediction, deep learning, resting-state fMRI, fluid intelligence, personality, emotion

Introduction

Deep neural networks (DNNs) have enjoyed tremendous success in machine learning (Lecun et al., 2015). As such, there has been significant interest in the application of DNNs to neuroscience research. DNNs have been applied to neuroscience in at least two main ways. First, deep learning models have been used to simulate actual brain mechanisms, such as in vision (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014; Eickenberg et al., 2017) and auditory perception (Kell et al., 2018). Second, DNNs have been applied as tools to analyze neuroscience data, including lesion and tumor segmentation (Pinto et al., 2016; Havaei et al., 2017; Kamnitsas et al., 2017b; G. Zhao et al., 2018), anatomical segmentation (Wachinger et al., 2018; X. Zhao et al., 2018), image modality/quality transfer (Bahrami et al., 2016; Nie et al., 2017; Blumberg et al., 2018), image registration (Yang et al., 2017; Dalca et al., 2018), as well as behavioral and disease prediction (Plis et al., 2014; van der Burgh et al., 2017; Vieira et al., 2017; Nguyen et al., 2018).

Deep neural networks can perform well in certain scenarios where large quantities of data are unavailable, for example, winning multiple MICCAI predictive modeling challenges (Choi et al., 2016; Kamnitsas et al., 2017a; Hongwei Li et al., 2018). Yet, the conventional wisdom is that DNNs perform especially well when applied to well-powered samples, for instance, the 14 million images in ImageNet (Russakovsky et al., 2015) and Google 1 Billion Word Corpus (Chelba et al., 2014). However, in many neuroimaging applications, the available data often only involve hundreds or thousands of participants, while the associated feature dimensions can be significantly larger, such as entries of connectivity matrices with upwards of 100,000 edges. Consequently, we hypothesize that in certain neuroimaging applications, DNNs might not be the optimal choice for a machine learning problem (Bzdok and Yeo, 2017). Here, we investigated whether DNNs can outperform classical machine learning for behavioral prediction using resting-state functional connectivity (RSFC).

RSFC measures the synchrony of resting-state functional magnetic resonance image (rs-fMRI) signals between brain regions (Biswal et al., 1995; Fox and Raichle, 2007; Buckner et al., 2013), while participants are lying at rest without any explicit task. RSFC has been widely used for exploring human brain organization and mental disorders (Smith et al., 2009; Assaf et al., 2010; Power et al., 2011; Yeo et al., 2011; Bertolero et al., 2015). For a given brain parcellation scheme (e.g., Shen et al., 2013; Gordon et al., 2016; Glasser et al., 2017; Eickhoff et al., 2018), the parcels can be used as regions of interest (ROIs), such that a whole brain (or cortical) RSFC matrix can be computed for each participant. Each entry of the RSFC matrix corresponds to the strength of functional connectivity between two brain

regions. The entries of the RSFC matrices can then be used as features for predicting behavioral measures (e.g., fluid intelligence) in individual participants (Finn et al., 2015; Smith et al., 2015; Dubois and Adolphs, 2016; Rosenberg et al., 2016; Reinen et al., 2018).

In this work, we compared kernel regression with three DNN architectures in RSFC-based behavioral prediction. Kernel regression is a non-parametric classical machine learning algorithm (Murphy, 2012) that has previously been utilized in various neuroimaging prediction problems, including RSFC-based behavioral prediction (Raz et al., 2017; Zhu et al., 2017; Li et al., 2018; Kong et al., 2018). Our three DNN implementations included a generic, fully-connected feedforward neural network, and two state-of-the-art DNNs specifically developed for RSFC-based prediction (Kawahara et al., 2017; Parisot et al., 2017, 2018). An initial version of this study utilizing only the fluid intelligence measure in the HCP dataset has been previously presented at a workshop (He et al., 2018). By using RSFC data from nearly 10,000 participants and a broad range of behavioral (and demographic) measures from the HCP (Smith et al., 2013; Van Essen et al., 2013) and UK Biobank (Sudlow et al., 2015; Miller et al., 2016), this current extended study represents one of the largest empirical evaluations of DNN's utility in RSFC-based fingerprinting.

Methods

Datasets

Two datasets were considered: the Human Connectome Project (HCP) S1200 release (Van Essen et al., 2013) and the UK Biobank (Sudlow et al., 2015; Miller et al., 2016). Both datasets contained multiple types of neuroimaging data, including structural MRI, rs-fMRI, and multiple behavioral and demographic measures for each subject.

HCP S1200 release comprised 1206 healthy young adults (age 22-35). There were 1,094 subjects with both structural MRI and rs-fMRI. Both structural MRI and rs-fMRI were acquired on a customized Siemens 3T "Connectome Skyra" scanner at Washington University at St. Louis. The structural MRI was 0.7mm isotropic. The rs-fMRI was 2mm isotropic with TR of 0.72s and 1200 frames per run (14.4 minutes). Each subject had two sessions of rs-fMRI, and each session contained two rs-fMRI runs. A number of behavioral measures was also collected by HCP. More details can be found elsewhere (Van Essen et al., 2012; Barch et al., 2013; Smith et al., 2013) .

The UK Biobank is a prospective epidemiological study that have recruited 500,000 adults (age 40-69) between 2006-2010 (Sudlow et al., 2015). 100,000 of these 500,000 participants will be brought back for multimodal imaging by 2022 (Miller et al., 2016). Here we considered an initial release of 10065 subjects with both structural MRI and rs-fMRI data. Both structural MRI and rs-fMRI were acquired on harmonized Siemens 3T Skyra scanners at three UK Biobank imaging centres (Cheadle Manchester, Newcastle, and Reading). The structural MRI was 1.0mm isotropic. The rs-fMRI was 2.4mm isotropic with TR of 0.735s and 490 frames per run (6 minutes). Each subject had one rs-fMRI run. A number of behavioral measures was also collected by the UK Biobank. More details can be found elsewhere (Elliott and Peakman, 2008; Sudlow et al., 2015; Miller et al., 2016; Alfaro-Almagro et al., 2018) .

Preprocessing and RSFC

We utilized ICA-FIX MSM-All grayordinate rs-fMRI data provided by the HCP S1200 release (HCP S1200 manual; Van Essen et al., 2012, 2013; Glasser et al., 2013; Smith et al., 2013; Griffanti et al., 2014; Salimi-Khorshidi et al., 2014). To eliminate residual motion and respiratory-related artifacts (Burgess et al., 2016), we performed further censoring and nuisance regression (Li et al., 2018; Kong et al., 2018). Runs with more than 50% censored frames were discarded. We considered 400 cortical (Schaefer et al., 2018) and 19 sub-cortical (Fischl et al., 2002) ROIs. The preprocessed rs-fMRI time courses were

averaged across all grayordinate locations within each ROI. RSFC was then computed using Pearson's correlation of the averaged time courses for each run of each subject (with the censored frames excluded for the computation). The RSFC was averaged across all runs, resulting in one 419 x 419 RSFC matrix for each subject.

In the case of the UK Biobank, we utilized the 55 x 55 RSFC (Pearson's correlation) matrices provided by the Biobank (Miller et al., 2016; Alfaro-Almagro et al., 2018). The 55 ROIs were obtained from a 100-component whole-brain spatial-ICA (Beckmann and Smith, 2004), of which 45 components were considered to be artifactual (Miller et al., 2016). The use of a different parcellation scheme in the UK Biobank (compared with the HCP dataset) ensures that our results are robust to the particular choice of ROIs.

FC-based prediction setup

We considered 58 behavioral measures across cognition, emotion and personality from the HCP (Table S1; Kong et al., 2018). By restricting the dataset to participants with at least one run (that survived censoring) and all 58 behavioral measures, we were left with 953 subjects. 23, 67, 62 and 801 subjects had 1, 2, 3 and 4 runs respectively.

In the case of the UK Biobank, we considered four behavioral and demographic measures: age, sex, fluid intelligence and pairs matching[1] (number of incorrect matches). By restricting the dataset to participants with 55 x 55 RSFC matrices and all four measures, we were left with 8868 subjects.

For both datasets, kernel regression and three DNNs were applied to predict the behavioral and demographic measures of individual subjects based on individuals' RSFC matrices. More specifically, the RSFC data of each participant was summarized as an N x N matrix, where N is the number of brain ROIs. Each entry in the RSFC matrix represented the strength of functional connectivity between two ROIs. The entries of the RSFC matrix were then used as features to predict behavioral and demographic measures in individual participants.

Kernel ridge regression

Kernel regression (Murphy, 2012) is a non-parametric classical machine learning algorithm. Let y be the behavioral measure (e.g., fluid intelligence) and c be the RSFC matrix of a test subject. Let y_i be the behavioral measure (e.g., fluid intelligence) and c_i be the

[1] The pairs matching task requires participants to memorize the positions of matching pairs of cards.

RSFC matrix of the i-th training subject. Roughly speaking, kernel regression will predict the test subject's behavioral measure to be the weighted average of the behavioral measures of all training subjects: $y \approx \sum_{i \in training\ set} Similarity(c_i, c) y_i$, where $Similarity(c_i, c)$ is the similarity between the RSFC matrices of the test subject and i-th training subject. Here, we simply set $Similarity(c_i, c)$ to be the Pearson's correlation between the lower triangular entries of matrices c_i and c. In practice, an l_2 regularization term is needed to avoid overfitting (i.e., kernel ridge regression). The level of l_2 regularization is controlled by the hyperparameter λ. More details are found in Appendix A1.

Fully-connected neural network (FNN)

Fully-connected neural networks (FNNs) belong to a generic class of feedforward neural networks (Lecun et al., 2015) illustrated in Figure 1. A FNN takes in vector data as an input and outputs a vector. A FNN consists of several fully connected layers. Each fully connected layer consists of multiple nodes. Data enters the FNN via the input layer nodes. Each node (except input layer nodes) is connected to all nodes in the previous layer. The values at each node is the weighted sum of node values from the previous layer. The weights are the trainable parameters in FNN. The outputs of the hidden layer nodes typically go through a nonlinear activation function, e.g., Rectified Linear Units (ReLU; $f(x) = max(0, x)$), while the output layer tends to be linear. The value at each output layer node typically represents a predicted quantity. Thus, FNNs (and neural networks in general) allow the prediction of multiple quantities simultaneously. In this work, the inputs to the FNN are the vectorized RSFC (i.e., lower triangular entries of the RSFC matrices) and the outputs are the behavioral or demographic variables we seek to predict.

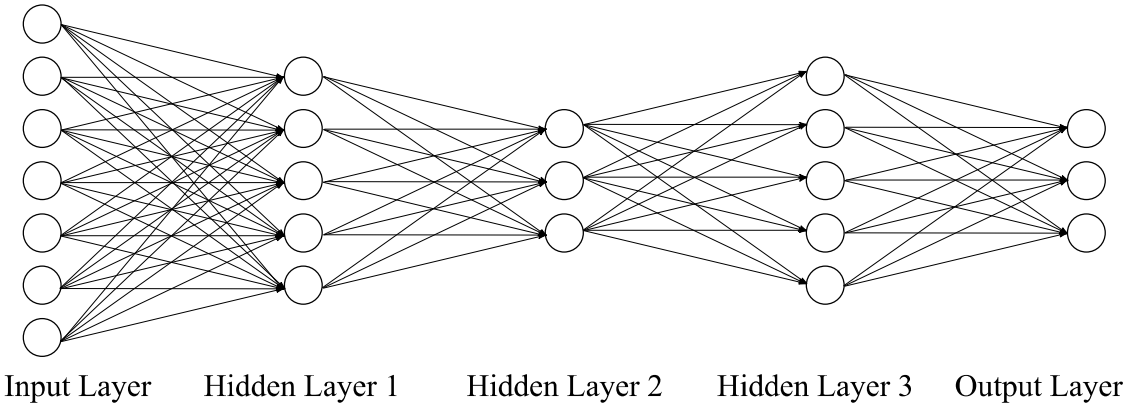

Figure 1. Schematic of a feedforward neural network (FNN). A FNN takes in vectorized RSFC matrix entries as inputs and outputs behavioral or demographic predictions. A FNN consists of an input layer, several hidden layers (three layers are shown here) and an output

layer. The number of nodes in the input layer is equal to the number of elements in the lower triangular portion of the RSFC matrix. The number of nodes in the output layer is typically equal to the number of behavioral measures we are predicting. The number of hidden layers and number of nodes in the hidden layers are among the many hyperparameters that have to be tuned.

BrainNetCNN

One potential weakness of the FNN is that it does not exploit the (mathematical and neurobiological) structure of the RSFC matrix, e.g., RSFC matrix is symmetric, positive definite and represents a network. On the other hand, BrainNetCNN (Kawahara et al., 2017) is a specially designed DNN for connectivity data, illustrated in Figure 2. BrainNetCNN allows the application of convolution to connectivity data, resulting in significantly less trainable parameters than the FNN. This leads to less parameters, which should theoretically improve the ease of training and reduce overfitting issues. In this work, the input to the BrainNetCNN is the $N \times N$ RSFC matrix and the outputs are the behavioral or demographic variables we seek to predict.

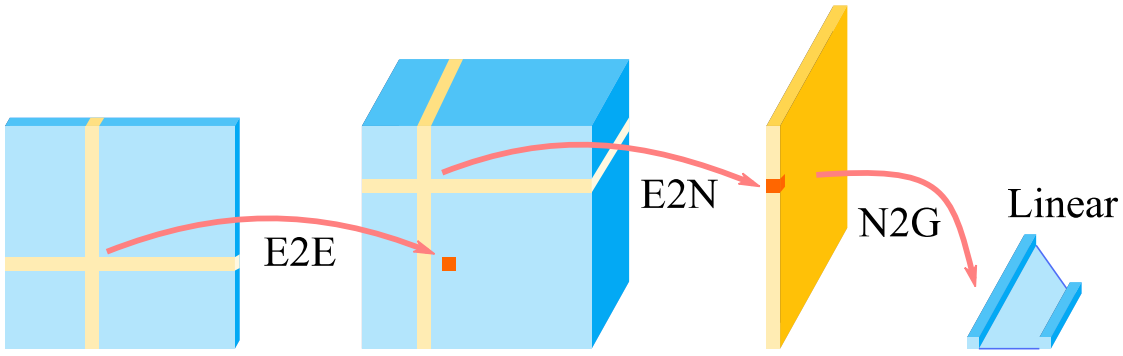

Figure 2. Schematic of the BrainNetCNN (Kawahara et al., 2017)**.** The BrainNetCNN takes in the RSFC matrix as an input and outputs behavioral or demographic predictions. BrainNetCNN consists of four types of layers, Edge-to-Edge (E2E) layer, Edge-to-Node (E2N) layer, Node-to-Graph (N2G) layer, and a final fully connected (Linear) layer. The number of the E2E layers can be any number greater than or equal to zero. On the other hand, there is one E2N layer and one N2G layer. The number of convolution filters and number of nodes in these layers are among the many hyperparameters that have to be tuned.

The BrainNetCNN takes in any connectivity matrix directly as an input and outputs behavioral or demographic predictions. Kawahara et al. (2017) used this model for predicting age and neurodevelopmental outcomes from structural connectivity data. BrainNetCNN consists of four types of layers: Edge-to-Edge (E2E) layer, Edge-to-Node (E2N) layer, Node-to-Graph (N2G) layer and a final fully connected (linear) layer. The first three types of layers

are specially designed layers introduced in the BrainNetCNN. The final fully connected layer is the same as that used in FNNs.

The Edge-to-Edge (E2E) layer is a convolution layer using cross-shaped filters (Figure 2). The cross-shaped filter is applied to each element of the input matrix. Thus, for each filter, the E2E layer takes in an $N \times N$ matrix and outputs an $N \times N$ matrix. The number of E2E layer is arbitrary and is a tunable hyperparameter. The outputs of the final E2E layer are inputs to the E2N layer. The E2N layer is similar to the E2E layer, except that the cross-shaped filter is applied to only the diagonal entries of the input matrix. Thus, for each filter, the E2N layer takes in an $N \times N$ matrix and outputs a $N \times 1$ vector. There is one E2N layer for BrainNetCNN. The outputs of the E2N layer are the inputs to the Node-to-Graph (N2G) layer. The N2G layer is simply a fully connected hidden layer similar to the a FNN's hidden layer. Finally, the outputs of the N2G layer are linearly summed by the final fully connected layer to provide a final set of prediction values.

Graph convolutional neural network (GCNN)

Standard convolution applies to data that lies on a Euclidean grid (e.g., images). Graph convolution exploits the graph Laplacian in order to generalize the concept of standard convolution to data lying on nodes connected together into a graph. This allows the extension of the standard CNN to graph convolutional neural networks (GCNNs; Defferrard et al., 2016; Bronstein et al., 2017; Kipf and Welling, 2017). There are many different ways that GCNN can be applied to neuroimaging data (Kipf and Welling, 2017; Ktena et al., 2018; Zhang et al., 2018). Here we considered the innovative GCNN developed by Kipf and Welling (2017) and extended to neuroimaging data by Parisot and colleagues (Parisot et al., 2017, 2018) . Figure 3 illustrates this approach.

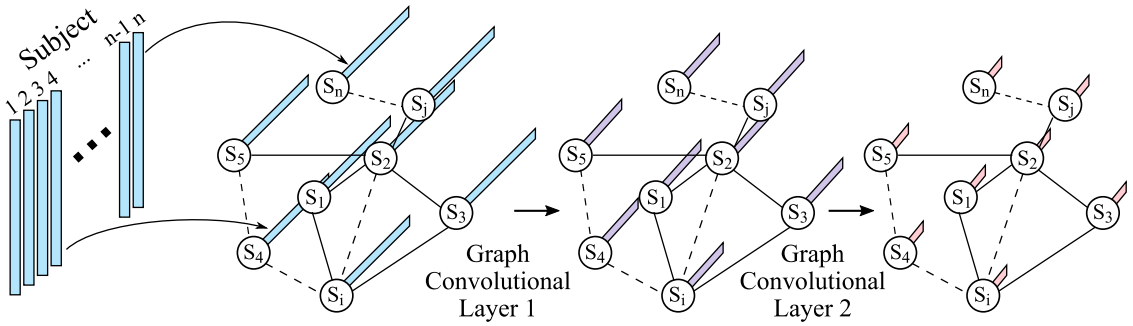

Figure 3. Schematic of a graph convolution neural network (GCNN; Parisot et al., 2017, 2018)**.** This particular GCNN takes in vectorized RSFC matrices of all subjects as input and outputs behavioral (or demographic) prediction of all subjects. Each node represents a

subject. The edges represent similarity between two subjects (e.g., based on the similarity between their RSFC matrices). The GCNN consists of several graph convolutional layers, which extends standard convolution to graph convolution.

The input to an FNN (Figure 1) or a BrainNetCNN (Figure 2) is the RSFC data of a single subject. By contrast, the GCNN takes in data (e.g., vectorized RSFC) of *all* subjects as input and outputs behavioral (or demographic) predictions of *all* subjects (Parisot et al., 2017, 2018). In other words, data from the training, validation, and testing sets are all input into the GCNN at the same time. To avoid leakage of information across training, validation and test sets, masking of data is applied during the calculation of the loss function and gradient descent.

More importantly, the graph in GCNN does not represent connectivity matrices (like in BrainNetCNN). Instead, each node represents a subject and edges are determined by the similarity between subjects. This similarity is problem dependent. For example, in the case of autism spectrum disorder (ASD) classification, similarity between two subjects is defined based on sex, sites and RSFC, i.e., two subjects are more similar if they have the same sex, from the same site and have similar RSFC patterns (Parisot et al., 2017, 2018). The use of sex and sites in the graph definition were particular important for this specific application, since ASD is characterized by strong sex-specific effects and the database included data from multiple unharmonized sites (Di Martino et al., 2014).

Similar to the original studies (Parisot et al., 2017, 2018), we utilized vectorized RSFC (lower triangular entries of the RSFC matrix) of all subjects as inputs to the GCNN. Edges between subjects were defined based on Pearson's correlation between lower triangular portions of RSFC matrices.

HCP training, validation and testing

For the HCP dataset, 20-fold cross-validation was performed. The 953 subjects were divided into 20 folds, such that family members were not split across folds. Inner-loop cross-validation was performed for hyperparameter tuning. More specifically, for a given test fold, cross-validation was performed on the remaining 19 folds with different hyperparameters. The best hyperparameters were then used to train on the 19 folds. The trained model was then applied to the test fold. This was repeated for all 20 test folds.

In the case of kernel regression, there was only one single hyperparameter λ (that controls the l_2 regularization; see Appendix A.1). A separate hyperparameter was tuned for

each test fold and each behavioral measure separately based on a grid search over the hyperparameter.

In the case of the DNNs, there was a large number of hyperparameters, e.g., number of layers, number of nodes, number of training epochs, dropout rate, optimizer (e.g., stochastic gradient or ADAM), weight initialization, activation functions, regularization, etc. GCNN also has additional hyperparameters tuned, e.g., definition of the graph and graph Laplacian estimation.

If we trained a different DNN for each of the 58 behavioral measures, a proper hyperparameter tuning would not be computationally feasible. Thus, a single FNN (or BrainNetCNN or GCNN) was trained for all 58 behavioral measures. We note that the joint prediction of multiple behavioral measures might not be a disadvantage for the DNNs and might potentially even improve prediction accuracies (Rahim et al., 2017). Furthermore, we tried to tune each DNN (FNN, BrainNetCNN or GCNN) for only fluid intelligence, but the performance for fluid intelligence prediction was not better than predicting all 58 behavioral measures simultaneously.

Furthermore, a proper inner-loop 20-fold cross-validation would involve tuning the hyperparameters for each DNN 20 times (once for each split of the data into training-test folds), which was computationally prohibitive. Thus, for each DNN (FNN, BrainNetCNN and GCNN), we tuned the hyperparameters once, using the first split of the data into training-test folds, and simply re-used the optimal hyperparameters for the remaining training-test splits of the data. Such a procedure biases the prediction performance in favor of the DNNs (relative to kernel regression), so the results should be interpreted accordingly (see Discussion). Such a bias is avoided in the UK Biobank dataset (see below). Further details about DNN hyperparameters are found in Appendix A2.

As is common in the FC-based prediction literature (Finn et al., 2015), model performance was evaluated based on the correlation between predicted and actual behavioral measures across subjects within each test fold. Furthermore, since certain behavioral measures were correlated with motion (Siegel et al., 2017), age, sex, and motion (FD) were regressed from the behavioral measures from the training and test folds (Li et al., 2018; Kong et al., 2018). Regression coefficients were estimated from the training folds and applied to the test folds.

UK Biobank training, validation and testing

The large UK Biobank dataset allowed us the luxury of splitting the 8868 subjects into training (N = 6868), validation (N = 1000) and test (N = 1000) sets, instead of employing an inner-loop cross-validation procedure like in the HCP dataset. Care were taken so that the distributions of various attributes (sex, age, fluid intelligence and pairs matching) were similar across training, validation and test sets.

Hyperparameters were tuned using the training and validation sets. The test set was only utilized to evaluate the final prediction performance. A separate DNN was trained for each of the four behavioral and demographic measures. Thus, the hyperparameters were tuned independently for each behavioral/demographic measure. Further details about DNN hyperparameters are found in Appendix A2. Initial experiments using a single neural network to predict all four measures simultaneously (like in the HCP dataset) did not appear to improve performance and so was not further pursued.

Like before, prediction accuracies for age, fluid intelligence and pairs matching were evaluated based on the correlation between predicted and actual measures across subjects within the test set. Since the age prediction literature often used mean absolute error (MAE) as an evaluation metric (Liem et al., 2017; Cole et al., 2018; Varikuti et al., 2018), we also included MAE as an evaluation metric. In the case of sex, accuracy was defined as the fraction of participants whose sex was correctly predicted. Like before, we regressed age, sex and motion from fluid intelligence and pairs matching measures in the training set and apply the regression coefficients to the validation and test sets. When predicting age and sex, no regression was performed.

Deep neural network implementation

The DNNs were implemented using Keras (Chollet, 2015) or PyTorch (Paszke et al., 2017) and run on NVIDIA Titan Xp GPU using CUDA. Our implementation of BrainNetCNN and GCNN were based on Github code from the original papers (Kawahara et al., 2017; Kipf and Welling, 2017). Our implementation achieved similar results for the experiments provided in the original Github implementations. More details can be found in Appendix A2.

Statistical tests

For the HCP dataset, we performed 20-fold cross-validation, yielding a prediction accuracy for each test fold. To compare two algorithms, the corrected resampled t-test was

performed (Nadeau and Bengio, 2003; Bouckaert and Frank, 2004). The corrected resampled t-test corrects for the fact that the accuracies across test folds were not independent.
In the case of the UK Biobank, there was only a single test fold, so the corrected resampled t-test could not be applied. Instead, when comparing correlations from two algorithms, the Steiger's Z-test was utilized (Steiger, 1980). When comparing prediction errors for age (MAE; mean absolute error), a two-tailed paired sample t-test was performed. When comparing prediction accuracies for sex, the McNemar's test was utilized (McNemar, 1947).

Data and code availability

This study utilized publicly available data from the HCP (https://www.humanconnectome.org/) and UK Biobank (https://www.ukbiobank.ac.uk/). The 400 cortical ROIs (Schaefer et al., 2018) can be found here (https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/brain_parcellation/Schaefer2018_LocalGlobal). The code utilized in this study can be found here: https://www.dropbox.com/sh/iq2d4gttxe3qvct/AAAVw7YJnVSwtOjouZDhhyPGa?dl=0 (note to readers/reviewers: we are in the midst of pushing our code to github. The dropbox link will be replaced by a github link).

Results

Three DNNs, fully connected neural network (FNN), BrainNetCNN and Graph Convolution Neural Network (GCNN), were compared with kernel regression in FC-based behavioral prediction using the HCP and UK Biobank datasets.

HCP behavioral prediction

Figure 4 shows the prediction accuracy (correlation) averaged across 58 HCP behavioral measures and 20 test folds. FNN achieved the highest average prediction accuracy of r = 0.121 ± 0.063 (mean ± std). On the other hand, kernel regression achieved an average prediction accuracy of r = 0.115 ± 0.036 (mean ± std). However, there was no statistical difference between FNN and kernel regression (p = 0.60; see Methods).

Interestingly, BrainNetCNN (r = 0.110 ± 0.043) and GCNN (r = 0.072 ± 0.034) did not outperform FNN, even though the two DNNs were designed for neuroimaging data. For completeness, Figures 5, S1, and S2 show the behavioral prediction accuracies for all 58 behavioral measures.

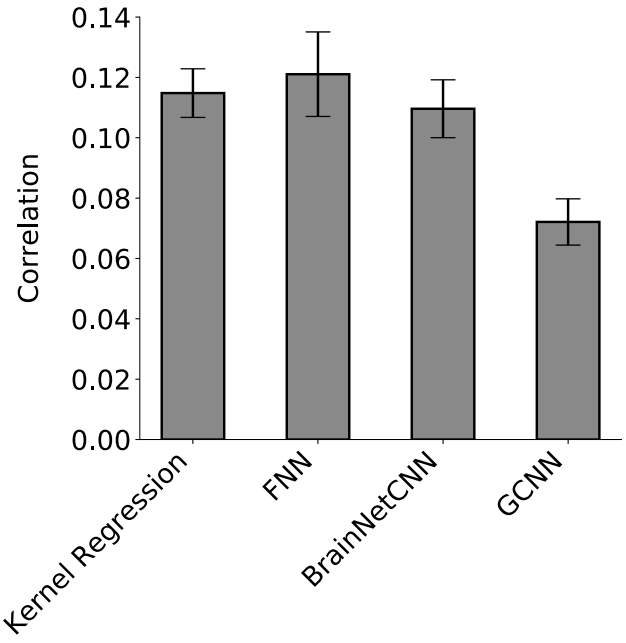

Figure 4. Prediction accuracy (correlation) averaged across 58 HCP behavioral measures and 20 test folds. Correlation was computed for each test fold and each behavior, and then averaged across the 58 behaviors. Bars show mean across test folds. Error bars show standard error of model performance across cross-validation folds. Kernel regression and FNN performed the best. There was no statistical difference (p = 0.60) between kernel regression and FNN.

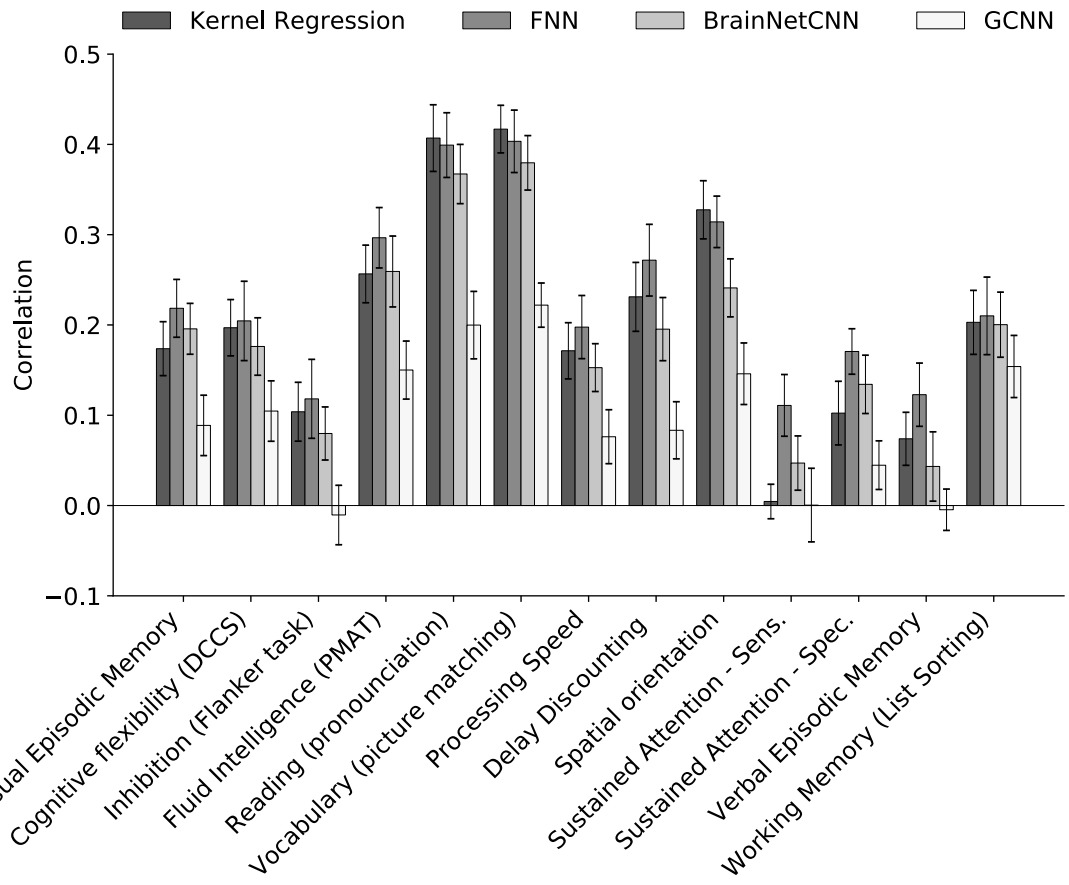

Figure 5. Prediction accuracies (correlations) in a curated set of 13 HCP cognitive measures averaged across 20 test folds. Correlation was computed for each test fold and each behavior. Bars show mean across test folds. Error bars show standard errors of model performance across cross-validation folds. Prediction accuracies of the remaining 45 behavioral measures are found in Figures S1 and S2.

UK Biobank behavioral and demographics prediction

Table 1 and Figure 6 show the prediction performances of sex, age, pairs matching and fluid intelligence. Kernel regression performed the best for age and fluid intelligence. BrainNetCNN performed the best for sex and pairs matching.

Statistical tests were performed between kernel regression and the three DNNs (see Methods). False discovery rate ($q < 0.05$) was applied to correct for multiple comparisons correction. For age (MAE), kernel regression was statistically better than GCNN ($p = 1.8e\text{-}6$). For fluid intelligence, kernel regression was statistically better than GCNN ($p = 5.5e\text{-}5$).

On the other hand, there was no statistical difference between kernel regression and BrainNetCNN in the case of sex and pairs matching, even though BrainNetCNN achieved a nominally higher accuracy.

Interestingly, the FNN achieved poor performance in the case pairs matching ($r = -0.0006$). Upon further investigation, we found that FNN achieved an accuracy of $r = 0.079$ in the UK Biobank validation set. Without any hyperparameter tuning (i.e., using the default set of hyperparameters), FNN achieved accuracies of $r = 0.046$ and $r = 0.031$ in the validation and test sets respectively. Overall, this suggests that the hyperparameter tuning overfitted the validation set, despite the rather large sample size.

Model	Sex	Age		Pairs matching	Fluid intelligence
	Accuracy	Correlation	MAE	Correlation	Correlation
Kernel Regression	0.916	**0.600**	**4.826**	0.061	**0.239**
FNN	0.908	0.598	4.896	-0.0006	0.239
BrainNetCNN	**0.917**	0.596	4.836	**0.063**	0.236
GCNN	0.908	0.577	5.110*	0.030	0.155*

Table 1. Prediction performance of four behavioral and demographic measures in the UK Biobank. For age (MAE), lower values imply better performance. For all the other measures, larger values imply better performance. **Bold** indicates best performance, although it does not imply statistical significance. Statistical tests were performed to compare kernel regression with each of the three DNNs. * indicates statistical significance after FDR ($q < 0.05$) correction.

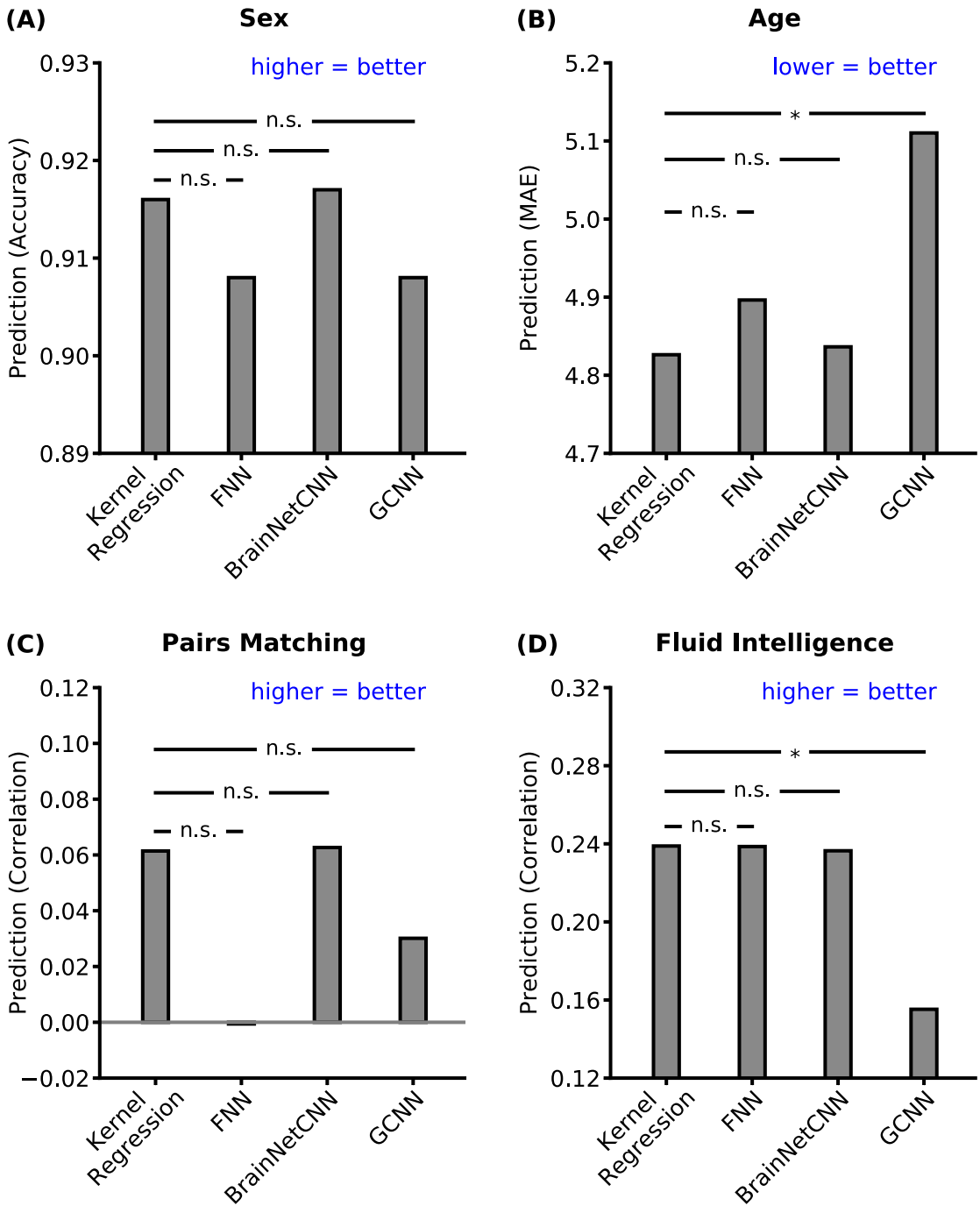

Figure 6. Prediction performance of four behavioral and demographic measures in the UK Biobank. For age (MAE), lower values imply better performance. For all the other measures, larger values imply better performance. The horizontal lines represent statistical tests between kernel regression and the DNNs. “n.s” stands for not significant. “*” implies statistical significance after FDR ($q < 0.05$) correction.

Computational costs

Kernel regression has a close-form solution (Appendix A1) and only one hyperparameter, so the computational cost is extremely low. For example, kernel regression training and grid search of 32 hyperparameter values in the UK Biobank validation set took about 20 minutes (single CPU core) for one behavioral measure. This is one reason why we considered kernel regression instead of other slower classical approaches (e.g., support vector regression or elastic net) requiring iterative optimization. On the other hand, FNN training and tuning of hyperparameters in the UK Biobank validation set took around 80 hours (single GPU) for one behavioral measure, excluding the manhours necessary for the manual tuning.

Discussion

In this study, we showed that DNNs did not outperform kernel regression in RSFC-based prediction of a wide range of behavioral and demographic measures across two large-scale datasets totaling almost 10,000 participants. Furthermore, FNN performed as well as the two DNNs that were specifically designed for connectome data[2]. Given comparable performance between kernel regression and the DNNs and the significantly greater computational costs associated with DNNs, our results suggest that DNNs should be more critically evaluated in the neuroimaging literature despite their promise.

Potential reasons why DNNs did not outperform kernel regression for RSFC-based prediction

There are several potential reasons why DNNs did not outperform kernel regression in our experiments on RSFC-based behavioral prediction. First, given the much larger datasets used in computer vision and natural language processing (Chelba et al., 2014; Russakovsky et al., 2015), it is possible that there was not enough neuroimaging data (even in the UK Biobank) to fully exploit DNNs.

Second, while the human brain is nonlinear and hierarchically organized (Deco et al., 2011; Breakspear, 2017), such a structure might not be reflected in the RSFC matrix in a way that was exploitable by the DNNs we considered. This could be due to the measurements themselves (Pearson's correlations of rs-fMRI timeseries), the particular representation (N x N connectivity matrices) or particular choices of DNNs, although we again note that BrainNetCNN and GCNN were specifically developed for connectome data.

Third, it is well-known that hyper-parameter settings and architectural details can impact the performance of DNNs. Thus, it is possible that the benchmark DNNs we implemented in this work can be further optimized. However, we do not believe this would alter our conclusions for two reasons. First, for some measures (e.g., sex classification in the UK Biobank), we were achieving performance at or near the state-of-the-art. Second, experiments with an automatic algorithm for tuning DNN hyperparameters (Ilievski et al., 2017) did not yield better performance than our hand-tuned hyperparameters (results not shown).

[2] FNN did seem to perform the worst for pairs matching in the UK Biobank, but the difference was not statistically significant. Furthermore, no approach seems to be able to predict pairs matching well.

Improving future DNNs research in neuroimaging

Given the exciting DNN results published in the top neuroimaging journals, we started this project with the expectation that DNNs would significantly outperform kernel regression. However, the results of this study suggest potential lessons for future DNN research in neuroimaging.

First, many DNN papers in neuroimaging do not utilize strong baseline algorithms for comparisons. In the case of RSFC-based behavioral prediction, our results suggest that kernel regression is a good baseline to be considered in future studies. Furthermore, in many (if not all) applications, a simple, but powerful baseline would be to replace the nonlinear activation functions (used in the DNN) with linear ones (Huang et al., 2018; Nguyen et al., 2018).

Second, the sample sizes of many DNN neuroimaging studies are often too small. In the case of behavioral prediction or disease classification, where the sample size is equal to the number of participants, we recommend at least a minimum of several hundred participants, since our results suggest that DNNs can achieve comparable performance with kernel regression. Thousands of participants would be better. Yet, given the results of this study, studies should perhaps aspire to even more participants. It is worth noting that what constitutes sample size depends on the problem. In the case of dense anatomical segmentation, the training data might involve manual segmentation of millions of voxels in a relatively small number of participants. In this scenario, the effective sample size might be closer to the number of labeled voxels than the number of labeled subjects. Consequently, this might explain the success of DNNs in segmentation challenges (Kamnitsas et al., 2017a; Hongwei Li et al., 2018).

Third, there are significantly more hyperparameters in DNNs compared with classical machine learning approaches. For example, for a fixed kernel (e.g., correlation metric in our study), kernel regression has one single regularization parameter. Even with a nonlinear kernel (e.g. radial basis function), there would only be two hyperparameters. This is in contrast to DNNs, where there can easily be more than ten hyperparameters. As such, it is important that studies spelled out clearly how those hyperparameters are tuned. In our experience, tuning large number of hyperparameters within a k-fold inner-loop (nested) cross-validation framework is difficult for two reasons. First, tuning so many hyperparameters k times (once for each fold) is prohibitively expensive. Second, if manual tuning is performed, information from tuning one fold will inevitably leak to another fold (via the person tuning the hyperparameters). Consequently, if the dataset is sufficiently large (e.g., UK Biobank), we recommend the data be divided into training, validation and test sets, just

like in our experiments. Hyperparameter tuning should be performed only using the training and validation sets, with the test set only be utilized in the final evaluation. In smaller datasets (e.g., HCP), an inner-loop k-fold cross-validation might unfortunately be necessary to ensure stability of results (Varoquaux, 2018).

Finally, we encourage studies to make their code publicly available. Publicly available code makes it significantly easier for other researchers to perform comparisons with the published algorithms. The current evaluation study is made possible due to generous code sharing by various authors (Kawahara et al., 2017; Parisot et al., 2017, 2018). Furthermore, there are simply too many DNN hyperparameters (and design choices) to be listed in a paper. In fact, there were hyperparameters too complex to completely specify in this paper. However, we have made our publicly available, so researchers can refer to the code for the exact hyperparameters.

Limitations and caveats

Although the current study suggests that DNNs do not outperform kernel regression of RSFC-based behavioral prediction, it is possible that other DNNs (we have not considered) might outperform kernel regression. Furthermore, our study focused on the use of N x N RSFC matrices for behavioral prediction. Other RSFC features in combination with DNNs might potentially yield better performance (Hongming Li et al., 2018; Khosla et al., 2018). Furthermore, the final UK Biobank dataset will include 100,000 participants with neuroimaging data, which is ten times the number of participants used in the current study. The larger quantity of data might strongly benefit deep learning approaches.

Given the success of DNNs in many fields and at various MICCAI predictive modeling challenges, we strongly believe that DNN remains a promising tool for neuroimaging. However, researchers should carefully consider whether and how their applications would benefit from DNNs' advantages over classical alternatives, rather than simply assume that deep learning is a panacea for all problems.

Conclusion

By using a combined sample of nearly 10,000 participants, we showed that three DNNs did not outperform kernel regression in RSFC-based prediction of a wide range of behavioral and demographic measures. Although we believe that deep learning remains a promising tool for neuroimaging data analysis, this suggests that DNNs should be more critically evaluated in the neuroimaging literature. Deep learning research in neuroimaging applications would benefit from comparisons with stronger baseline algorithms, large sample sizes, transparency in hyperparameter tuning and code availability.

Acknowledgment

This work was supported by Singapore MOE Tier 2 (MOE2014-T2-2-016), NUS Strategic Research (DPRT/944/09/14), NUS SOM Aspiration Fund (R185000271720), Singapore NMRC (CBRG/0088/2015), NUS YIA and the Singapore National Research Foundation (NRF) Fellowship (Class of 2017). Our research also utilized resources provided by the Center for Functional Neuroimaging Technologies, P41EB015896 and instruments supported by 1S10RR023401, 1S10RR019307, and 1S10RR023043 from the Athinoula A. Martinos Center for Biomedical Imaging at the Massachusetts General Hospital. Our computational work was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg). The Titan Xp GPUs used for this research were donated by the NVIDIA Corporation. This research has been conducted using the UK Biobank resource under application 25163 and 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.

References

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L.R., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D.C., Zhang, H., Dragonu, I., Matthews, P.M., Miller, K.L., Smith, S.M., 2018. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. Neuroimage 166, 400–424. https://doi.org/10.1016/j.neuroimage.2017.10.034

Assaf, M., Jagannathan, K., Calhoun, V.D., Miller, L., Stevens, M.C., Sahl, R., O'Boyle, J.G., Schultz, R.T., Pearlson, G.D., 2010. Abnormal functional connectivity of default mode sub-networks in autism spectrum disorder patients. Neuroimage 53, 247–256. https://doi.org/10.1016/j.neuroimage.2010.05.067.Assaf

Bahrami, K., Shi, F., Rekik, I., Shen, D., 2016. Convolutional Neural Network for Reconstruction of 7T-like Images from 3T MRI Using Appearance and Anatomical Features, in: MICCAI 2016 DL Workshop. pp. 39–47. https://doi.org/10.1007/978-3-319-46976-8_5

Barch, D.M., Burgess, G.C., Harms, M.P., Petersen, S.E., Schlaggar, B.L., Corbetta, M., Glasser, M.F., Curtiss, S., Dixit, S., Feldt, C., Nolan, D., Bryant, E., Hartley, T., Footer, O., Bjork, J.M., Poldrack, R., Smith, S., Johansen-Berg, H., Snyder, A.Z., Van Essen, D.C., 2013. Function in the human connectome: Task-fMRI and individual differences in behavior. Neuroimage 80, 169–189. https://doi.org/10.1016/j.neuroimage.2013.05.033

Beckmann, C.F., Smith, S.M., 2004. Probabilistic Independent Component Analysis for Functional Magnetic Resonance Imaging. IEEE Trans. Med. Imaging 23, 137–152. https://doi.org/10.1109/TMI.2003.822821

Bertolero, M.A., Yeo, B.T.T., D'Esposito, M., 2015. The modular and integrative functional architecture of the human brain. Proc. Natl. Acad. Sci. 112, E6798–E6807. https://doi.org/10.1073/pnas.1510619112

Biswal, B., FZ, Y., VM, H., JS, H., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med 34, 537–541. https://doi.org/10.1002/mrm.1910340409

Blumberg, S.B., Tanno, R., Kokkinos, I., Alexander, D.C., 2018. Deeper image quality transfer: Training low-memory neural networks for 3D images. Int. Conf. Med. Image Comput. Comput. Interv. 118–125. https://doi.org/10.1007/978-3-030-00928-1_14

Bouckaert, R.R., Frank, E., 2004. Evaluating the Replicability of Significance Tests for

Comparing Learning Algorithms. Adv. Knowl. Discov. data Min. 3–12. https://doi.org/10.1007/978-3-540-24775-3

Breakspear, M., 2017. Dynamic models of large-scale brain activity. Nat. Neurosci. 20, 340–352. https://doi.org/10.1038/nn.4497

Bronstein, M.M., Bruna, J., Lecun, Y., Szlam, A., Vandergheynst, P., 2017. Geometric Deep Learning: Going beyond Euclidean data. IEEE Signal Process. Mag. 34, 18–42. https://doi.org/10.1109/MSP.2017.2693418

Buckner, R.L., Krienen, F.M., Yeo, B.T.T., 2013. Opportunities and limitations of intrinsic functional connectivity MRI. Nat. Neurosci. 16, 832–837. https://doi.org/10.1038/nn.3423

Burgess, G.C., Kandala, S., Nolan, D., Laumann, T.O., Power, J.D., Adeyemo, B., Harms, M.P., Petersen, S.E., Barch, D.M., 2016. Evaluation of Denoising Strategies to Address Motion-Correlated Artifacts in Resting-State Functional Magnetic Resonance Imaging Data from the Human Connectome Project. Brain Connect. 6, 669–680. https://doi.org/10.1089/brain.2016.0435

Bzdok, D., Yeo, B.T.T., 2017. Inference in the age of big data: Future perspectives on neuroscience. Neuroimage 155, 549–564. https://doi.org/10.1016/j.neuroimage.2017.04.061

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., Robinson, T., 2014. One billion word benchmark for measuring progress in statistical language modeling. Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH 2635–2639. https://doi.org/10.1016/j.csl.2015.07.001

Choi, Y., Kwon, Y., Lee, H., Kim, B.J., Paik, M.C., Won, J.-H., 2016. Ensemble of Deep Convolutional Neural Networks for Prognosis of Ischemic Stroke, in: Crimi, A., Menze, B., Maier, O., Reyes, M., Winzeck, S., Handels, H. (Eds.), International MICCAI Brainlesion Workshop BrainLes 2016: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer International Publishing, Cham, pp. 231–243.

Chollet, F., 2015. Keras [WWW Document]. GitHub. URL https://github.com/fchollet/keras

Cole, J.H., Ritchie, S.J., Bastin, M.E., Valdés Hernández, M.C., Muñoz Maniega, S., Royle, N., Corley, J., Pattie, A., Harris, S.E., Zhang, Q., Wray, N.R., Redmond, P., Marioni, R.E., Starr, J.M., Cox, S.R., Wardlaw, J.M., Sharp, D.J., Deary, I.J., 2018. Brain age predicts mortality. Mol. Psychiatry 23, 1385–1392. https://doi.org/10.1038/mp.2017.62

Dalca, A. V., Balakrishnan, G., Guttag, J., Sabuncu, M.R., 2018. Unsupervised learning for fast probabilistic diffeomorphic registration. Int. Conf. Med. Image Comput. Comput.

Interv. 729–738. https://doi.org/10.1007/978-3-030-00928-1_82

Deco, G., Jirsa, V.K., McIntosh, A.R., 2011. Emerging concepts for the dynamical organization of resting-state activity in the brain. Nat. Rev. Neurosci. 12, 43–56. https://doi.org/10.1038/nrn2961

Defferrard, M., Bresson, X., Vandergheynst, P., 2016. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering, in: Advances in Neural Information Processing Systems. pp. 3844–3852.

Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., Anderson, J.S., Assaf, M., Bookheimer, S.Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D.A., Gallagher, L., Kennedy, D.P., Keown, C.L., Keysers, C., Lainhart, J.E., Lord, C., Luna, B., Menon, V., Minshew, N.J., Monk, C.S., Mueller, S., Müller, R.A., Nebel, M.B., Nigg, J.T., O'Hearn, K., Pelphrey, K.A., Peltier, S.J., Rudie, J.D., Sunaert, S., Thioux, M., Tyszka, J.M., Uddin, L.Q., Verhoeven, J.S., Wenderoth, N., Wiggins, J.L., Mostofsky, S.H., Milham, M.P., 2014. The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry 19, 659–667. https://doi.org/10.1038/mp.2013.78

Dubois, J., Adolphs, R., 2016. Building a Science of Individual Differences from fMRI. Trends Cogn. Sci. 20, 425–443. https://doi.org/10.1016/j.tics.2016.03.014

Eickenberg, M., Gramfort, A., Varoquaux, G., Thirion, B., 2017. Seeing it all: Convolutional network layers map the function of the human visual system. Neuroimage 152, 184–194. https://doi.org/10.1016/j.neuroimage.2016.10.001

Eickhoff, S.B., Yeo, B.T.T., Genon, S., 2018. Imaging-based parcellations of the human brain. Nat. Rev. Neurosci. 19, 672–686. https://doi.org/10.1038/s41583-018-0071-7

Elliott, P., Peakman, T.C., 2008. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. Int. J. Epidemiol. 37, 234–244. https://doi.org/10.1093/ije/dym276

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

Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., Dale, A.M., 2002. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron 33, 341–355. https://doi.org/10.1016/S0896-

6273(02)00569-X

Fox, M.D., Raichle, M.E., 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat. Rev. Neurosci. 8, 700–711. https://doi.org/10.1038/nrn2201

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Yacoub, E., Ugurbil, K., Andersson, J.L., Beckmann, C.F., Jenkinson, M., Smith, S.M., Essen, D.C. Van, 2017. A Multi-Modal Oarcellation of Human Cerebral Cortex. Nature 536, 171–178. https://doi.org/10.1038/nature18933.A

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., 2013. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage 80, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Gordon, E.M., Laumann, T.O., Adeyemo, B., Huckins, J.F., Kelley, W.M., Petersen, S.E., 2016. Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. Cereb. Cortex 26, 288–303. https://doi.org/10.1093/cercor/bhu239

Griffanti, L., Salimi-Khorshidi, G., Beckmann, C.F., Auerbach, E.J., Douaud, G., Sexton, C.E., Zsoldos, E., Ebmeier, K.P., Filippini, N., Mackay, C.E., Moeller, S., Xu, J., Yacoub, E., Baselli, G., Ugurbil, K., Miller, K.L., Smith, S.M., 2014. ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. Neuroimage 95, 232–247. https://doi.org/10.1016/j.neuroimage.2014.03.034

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H., 2017. Brain tumor segmentation with Deep Neural Networks. Med. Image Anal. 35, 18–31. https://doi.org/10.1016/j.media.2016.05.004

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? 2018 Int. Work. Pattern Recognit. Neuroimaging, PRNI 2018 6–9. https://doi.org/10.1109/PRNI.2018.8423958

Huang, H., Hu, X., Zhao, Y., Makkie, M., Dong, Q., Zhao, S., 2018. Modeling Task fMRI Data Via Deep Convolutional Autoencoder. IEEE Trans. Med. Imaging 37, 1551–1561. https://doi.org/10.1109/TMI.2017.2715285

Ilievski, I., Akhtar, T., Feng, J., Shoemaker, C.A., 2017. Efficient Hyperparameter Optimization of Deep Learning Algorithms Using Deterministic RBF Surrogates, in: 31st AAAI Conference on Artificial Intelligence (AAAI-17).

Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M.,

Lee, M., Kainz, B., Rueckert, D., Glocker, B., 2017a. Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation, in: International MICCAI Brainlesion Workshop BrainLes 2017: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. pp. 450–462.

Kamnitsas, K., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017b. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78. https://doi.org/10.1016/j.media.2016.10.004

Kawahara, J., Brown, C.J., Miller, S.P., Booth, B.G., Chau, V., Grunau, R.E., Zwicker, J.G., Hamarneh, G., 2017. BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment. Neuroimage 146, 1038–1049. https://doi.org/10.1016/j.neuroimage.2016.09.046

Kell, A.J.E., Yamins, D.L.K., Shook, E.N., Norman-Haignere, S. V., McDermott, J.H., 2018. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron 98, 630–644.e16. https://doi.org/10.1016/j.neuron.2018.03.044

Khaligh-Razavi, S.M., Kriegeskorte, N., 2014. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput. Biol. 10. https://doi.org/10.1371/journal.pcbi.1003915

Khosla, M., Jamison, K., Kuceyeski, A., Sabuncu, M.R., 2018. Ensemble learning with 3D convolutional neural networks for functional connectome-based prediction. arXiv Prepr. arXiv1806.04209.

Kingma, D.P., Ba, J.L., 2015. Adam: a Method for Stochastic Optimization. Int. Conf. Learn. Represent. 2015 1–15. https://doi.org/http://doi.acm.org.ezproxy.lib.ucf.edu/10.1145/1830483.1830503

Kipf, T.N., Welling, M., 2017. Semi-Supervised Classification with Graph Convolutional Networks. Int. Conf. Learn. Represent. 1–14. https://doi.org/10.1051/0004-6361/201527329

Kong, R., Li, J., Orban, C., Sabuncu, M.R., Liu, H., Schaefer, A., Sun, N., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.T., 2018. Spatial Topography of Individual-Specific Cortical Networks Predicts Human Cognition, Personality, and Emotion. Cereb. Cortex 213041. https://doi.org/10.1093/cercor/bhy123

Ktena, S.I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., Rueckert, D., 2018. Metric learning with spectral graph convolutions on brain connectivity networks.

Neuroimage 169, 431–442. https://doi.org/10.1016/j.neuroimage.2017.12.052

Lecun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Li, H., Jiang, G., Zhang, J., Wang, R., Wang, Z., Zheng, W., Menze, B., 2018. Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. Neuroimage 183, 650–665. https://doi.org/10.1016/j.neuroimage.2018.07.005

Li, H., Satterthwaite, T.D., Fan, Y., 2018. Brain Age Prediction Based on Resting-State Functional Connectivity Patterns Using Convolutional Neural Networks, in: IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). pp. 101–104.

Li, J., Kong, R., Liegeois, R., Orban, C., Sun, N., Holmes, A.J., Sabuncu, M.R., Ge, T., Yeo, B.T.T., 2018. Global Signal Regression Strengthens Association between Resting-State Functional Connectivity and Behavior. Under Review.

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Kharabian Masouleh, S., Huntenburg, J.M., Lampe, L., Rahim, M., Abraham, A., Craddock, R.C., Riedel-Heller, S., Luck, T., Loeffler, M., Schroeter, M.L., Witte, A.V., Villringer, A., Margulies, D.S., 2017. Predicting brain-age from multimodal imaging data captures cognitive impairment. Neuroimage 148, 179–188. https://doi.org/10.1016/j.neuroimage.2016.11.005

Liu, D., Lin, X., Ghosh, D., 2007. Semiparametric Regression of Multidimensional Genetic Pathway Data : Least-Squares Kernel Machines and Linear Mixed Models. Biometrics 1079–1088. https://doi.org/10.1111/j.1541-0420.2007.00799.x

Maas, A.L., Hannun, A.Y., Ng, A.Y., 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models. Proc. 30 th Int. Conf. Mach. Learn. 28, 6.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12, 153–157. https://doi.org/10.1007/BF02295996

Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L.R., Griffanti, L., Douaud, G., Okell, T.W., Weale, P., Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., Matthews, P.M., Smith, S.M., 2016. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. Nat. Neurosci. 19, 1523–1536. https://doi.org/10.1038/nn.4393

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective, MIT Press. https://doi.org/10.1007/978-3-642-21004-4_10

Nadeau, C., Bengio, Y., 2003. Inference for the generalization error. Mach. Learn. 52, 239–

281. https://doi.org/10.1023/A:1024068626366

Nguyen, M., Sun, N., Alexander, D.C., Feng, J., Thomas Yeo, B.T., 2018. Modeling Alzheimer's disease progression using deep recurrent neural networks. 2018 Int. Work. Pattern Recognit. Neuroimaging, PRNI 2018 1–4. https://doi.org/10.1109/PRNI.2018.8423955

Nie, D., Trullo, R., Petitjean, C., Ruan, S., Shen, D., 2017. Medical Image Synthesis with Context-Aware Generative Adversarial Networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 417–425. https://doi.org/10.1007/978-3-319-66179-7_48

Parisot, S., Ira, S., Ferrante, E., Lee, M., Guerrero, R., Glocker, B., 2018. Disease Prediction using Graph Convolutional Networks : Application to Autism Spectrum Disorder and Alzheimer ' s Disease. Med. Image Anal. 1–26. https://doi.org/10.1016/j.media.2018.06.001

Parisot, S., Ktena, S.I., Ferrante, E., Lee, M., Moreno, R.G., Glocker, B., Rueckert, D., 2017. Spectral Graph Convolutions for Population-Based Disease Prediction, in: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017. Springer International Publishing, Cham, pp. 177–185. https://doi.org/10.1007/978-3-319-66179-7_21

Paszke, A., Chanan, G., Lin, Z., Gross, S., Yang, E., Antiga, L., Devito, Z., 2017. Automatic differentiation in PyTorch. Adv. Neural Inf. Process. Syst. 30 1–4.

Pinto, A., Alves, V., Silva, C.A., 2016. Brain Tumor Segmentation using Convolutional Neural Networks in MRI Images. IEEE Trans. Med. Imaging 35, 1240–1251. https://doi.org/10.1109/TMI.2016.2538465

Plis, S.M., Hjelm, D.R., Slakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H., Paulsen, J., Turner, J., Calhoun, V.D., 2014. Deep learning for neuroimaging: A validation study. Front. Neurosci. 8, 1–11. https://doi.org/10.3389/fnins.2014.00229

Power, J.D., Cohen, A.L., Nelson, S.M., Wig, G.S., Barnes, K.A., Church, J.A., Vogel, A.C., Laumann, T.O., Miezin, F.M., Schlaggar, B.L., Petersen, S.E., 2011. Functional Network Organization of the Human Brain. Neuron 72, 665–678. https://doi.org/10.1016/j.neuron.2011.09.006

Rahim, M., Thirion, B., Bzdok, D., Buvat, I., Varoquaux, G., 2017. Joint prediction of multiple scores captures better individual traits from brain images. Neuroimage 158, 145–154. https://doi.org/10.1016/j.neuroimage.2017.06.072

Raz, G., Svanera, M., Singer, N., Gilam, G., Bleich, M., Lin, T., Admon, R., Gonen, T., Thaler, A., Granot, R.Y., Goebel, R., Benini, S., Valente, G., 2017. Robust inter-subject audiovisual decoding in functional magnetic resonance imaging using high-dimensional regression. Neuroimage 163, 244–263. https://doi.org/10.1016/j.neuroimage.2017.09.032

Reinen, J.M., Chén, O.Y., Hutchison, R.M., Yeo, B.T.T., Anderson, K.M., Sabuncu, M.R., Öngür, D., Roffman, J.L., Smoller, J.W., Baker, J.T., Holmes, A.J., 2018. The human cortex possesses a reconfigurable dynamic network architecture that is disrupted in psychosis. Nat. Commun. 9, 1–15. https://doi.org/10.1038/s41467-018-03462-y

Rosenberg, M.D., Finn, E.S., Scheinost, D., Papademetris, X., Shen, X., Constable, R.T., Chun, M.M., 2016. A neuromarker of sustained attention from whole-brain functional connectivity. Nat. Neurosci. 19, 165–171. https://doi.org/10.1038/nn.4179

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vis. 115, 211–252. https://doi.org/10.1007/s11263-015-0816-y

Salimi-Khorshidi, G., Douaud, G., Beckmann, C.F., Glasser, M.F., Griffanti, L., Smith, S.M., 2014. Automatic denoising of functional MRI data: Combining independent component analysis and hierarchical fusion of classifiers. Neuroimage 90, 449–468. https://doi.org/10.1016/j.neuroimage.2013.11.046

Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X., Holmes, A.J., Eickhoff, S.B., Yeo, B.T.T., 2018. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cereb. Cortex 3095–3114. https://doi.org/10.1093/cercor/bhx179

Shen, X., Tokoglu, F., Papademetris, X., Constable, R.T., 2013. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. Neuroimage 82, 403–415. https://doi.org/10.1016/j.neuroimage.2013.05.081

Siegel, J.S., Mitra, A., Laumann, T.O., Seitzman, B.A., Raichle, M., Corbetta, M., Snyder, A.Z., 2017. Data quality influences observed links between functional connectivity and behavior. Cereb. Cortex 27, 4492–4502. https://doi.org/10.1093/cercor/bhw253

Smith, S.M., Beckmann, C.F., Andersson, J., Auerbach, E.J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D.A., Griffanti, L., Harms, M.P., Kelly, M., Laumann, T., Miller, K.L., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A.Z., Vu, A.T., Woolrich, M.W., Xu, J., Yacoub, E., Uğurbil, K., Van Essen, D.C., Glasser, M.F., 2013.

Resting-state fMRI in the Human Connectome Project. Neuroimage 80, 144–168. https://doi.org/10.1016/j.neuroimage.2013.05.039

Smith, S.M., Fox, P.T., Miller, K.L., Glahn, D.C., Fox, P.M., Mackay, C.E., Filippini, N., Watkins, K.E., Toro, R., Laird, A.R., Beckmann, C.F., 2009. Correspondence of the brain's functional architecture during activation and rest. Proc. Natl. Acad. Sci. 106, 13040–13045. https://doi.org/10.1073/pnas.0905267106

Smith, S.M., Nichols, T.E., Vidaurre, D., Winkler, A.M., Behrens, T.E.J., Glasser, M.F., Ugurbil, K., Barch, D.M., Van Essen, D.C., Miller, K.L., 2015. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nat. Neurosci. 18, 1565–1567. https://doi.org/10.1038/nn.4125

Steiger, J.H., 1980. Tests for comparing elements of a correlation matrix. Psychol. Bull. https://doi.org/10.1037/0033-2909.87.2.245

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T., Collins, R., 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 12, 1–10. https://doi.org/10.1371/journal.pmed.1001779

van der Burgh, H.K., Schmidt, R., Westeneng, H.J., de Reus, M.A., van den Berg, L.H., van den Heuvel, M.P., 2017. Deep learning predictions of survival based on MRI in amyotrophic lateral sclerosis. NeuroImage Clin. 13, 361–369. https://doi.org/10.1016/j.nicl.2016.10.008

Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K., 2013. The WU-Minn Human Connectome Project: An overview. Neuroimage 80, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E.J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., Della Penna, S., Feinberg, D., Glasser, M.F., Harel, N., Heath, A.C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S.E., Prior, F., Schlaggar, B.L., Smith, S.M., Snyder, A.Z., Xu, J., Yacoub, E., 2012. The Human Connectome Project: A data acquisition perspective. Neuroimage 62, 2222–2231. https://doi.org/10.1016/j.neuroimage.2012.02.018

Varikuti, D.P., Genon, S., Sotiras, A., Schwender, H., Hoffstaedter, F., Patil, K.R., Jockwitz, C., Caspers, S., Moebus, S., Amunts, K., Davatzikos, C., Eickhoff, S.B., 2018. Evaluation of non-negative matrix factorization of grey matter in age prediction. Neuroimage 173, 394–410. https://doi.org/10.1016/j.neuroimage.2018.03.007

Varoquaux, G., 2018. Cross-validation failure : Small sample sizes lead to large error bars. Neuroimage 180, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

Vieira, S., Pinaya, W.H.L., Mechelli, A., 2017. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neurosci. Biobehav. Rev. 74, 58–75. https://doi.org/10.1016/j.neubiorev.2017.01.002

Wachinger, C., Reuter, M., Klein, T., 2018. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. Neuroimage 170, 434–445. https://doi.org/10.1016/j.neuroimage.2017.02.035

Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J., 2014. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. 111, 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yang, X., Kwitt, R., Styner, M., Niethammer, M., 2017. Quicksilver: Fast predictive image registration – A deep learning approach. Neuroimage 158, 378–396. https://doi.org/10.1016/j.neuroimage.2017.07.008

Yeo, B.T.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zollei, L., Polimeni, J.R., Fischl, B., Liu, H., Buckner, R.L., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106, 1125–1165. https://doi.org/10.1152/jn.00338.2011.

Zhang, X., He, L., Chen, K., Luo, Y., Zhou, J., Wang, F., 2018. Multi-View Graph Convolutional Network and Its Applications on Neuroimage Analysis for Parkinson's Disease. arXiv Prepr. arXiv1805.08801. https://doi.org/arXiv:1805.08801v2

Zhao, G., Liu, F., Oler, J.A., Meyerand, M.E., Kalin, N.H., Birn, R.M., 2018. Bayesian convolutional neural network based MRI brain extraction on nonhuman primates. Neuroimage 175, 32–44. https://doi.org/10.1016/j.neuroimage.2018.03.065

Zhao, X., Wu, Y., Song, G., Li, Z., Zhang, Y., Fan, Y., 2018. A deep learning model integrating FCNNs and CRFs for brain tumor segmentation. Med. Image Anal. 43, 98–111. https://doi.org/10.1016/j.media.2017.10.002

Zhu, X., Thung, K.-H., Adeli, E., Zhang, Y., Shen, D., 2017. Maximum Mean Discrepancy Based Multiple Kernel Learning for Incomplete Multimodality Neuroimaging Data, in: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer-Assisted Intervention − MICCAI 2017. Springer International Publishing, Cham, pp. 72–80.

Appendix

A1. Kernel Regression

In this section, we describe kernel regression in detail (Liu et al., 2007; Murphy, 2012). The kernel matrix K encodes the similarity between pairs of subjects. Motivated by Finn and colleagues (2015), the i-th row and j-th column of the kernel matrix is defined as the Pearson's correlation between the i-th subject's vectorized RSFC and j-th subject's vectorized RSFC (considering only the lower triangular portions of the RSFC matrices). The behavioral measure y_i of subject i can be written as:

$$y_i = \sum_{j=1}^{M} \alpha_j K(c_i, c_j) + e_i \tag{1}$$

where c_i is the vectorized RSFC of the i-th subject, $K(c_i, c_j)$ is the element at i-th row and j-th column of kernel matrix, M is the total number of training subjects, e_i is the noise term and α_j is the trainable weight. The goal of kernel regression is to find an optimal set of α. To achieve this goal, we maximize the penalized likelihood function:

$$J = -\frac{1}{2}\sum_{i=1}^{M}\{y_i - \sum_{j=1}^{M} \alpha_j K(c_i, c_j)\}^2 \tag{2}$$

with respect to $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_M]^T$. To avoid overfitting, a l_2 regularization (i.e., kernel ridge regression) can be added, so the resulting optimization problem becomes:

$$\boldsymbol{\alpha} = \underset{\alpha}{\operatorname{argmin}} \frac{1}{2}(\boldsymbol{y} - \mathbb{K}\boldsymbol{\alpha})^T(\boldsymbol{y} - \mathbb{K}\boldsymbol{\alpha}) + \frac{\lambda}{2}\boldsymbol{\alpha}^T\mathbb{K}\boldsymbol{\alpha} \tag{3}$$

where $\mathbb{K}$ is the $M \times M$ kernel matrix, $\boldsymbol{y} = [y_1, y_2, \ldots, y_M]^T$ and λ is a hyperparameter that controls the l_2 regularization. By solving equation (3) with respect to α, we can predict a test subject's behavioral measure y_s as:

$$y_s = \boldsymbol{K_s}\boldsymbol{\alpha} = \boldsymbol{K_s}(\mathbb{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y} \tag{4}$$

where $\boldsymbol{K_s} = [K(c_s, c_1), K(c_s, c_2), \ldots, K(c_s, c_M)]$.

In the case of the HCP, λ was selected via inner-loop cross-validation. In the case of the UK biobank, λ was tuned on the validation set.

A2. More details of deep neural networks

In this section, we describe further details of our DNN implementation. In the case of the HCP dataset:

- For all three DNNs, all behavioral measures were z-normalized based on training data. The loss function was mean squared error (MSE). Optimizer was stochastic gradient descent (SGD). With the MSE loss, the output layer has 58 nodes (FNN and BrainNetCNN) or filters (GCNN).
- Final FNN structure is shown in table 2. Dropout of 0.6 was added before each fully-connected layer. L2 regularization of 0.02 was added for layer 2.
- Final BrainNetCNN structure is shown in table 3. Dropout of 0.5 was added after E2N layer. LeakyReLU (Maas et al., 2013) with alpha of 0.1 was used as the activation function for the first three layers.
- Final GCNN structure is shown in table 4. Dropout of 0.3 was added for each layer. L2 regularization of 8e-4 was added for layer 1. The nodes of the graph corresponded to subjects. Edges were constructed based on Pearson's correlation between subjects' vectorized RSFC. The graph was thresholded by only retaining edges with top 5% correlation (across the entire graph). However, this might result in a disconnected graph. Therefore, the top five correlated edges of each node were also retained (even if these edges were not among the top 5% correlated edges). The graph convolution filters were estimated using a 5-degree Chebyshev polynomial (Defferrard et al., 2016).

In the case of the UK Biobank:

- For all three DNNs, model ensemble was used to improve final test result: for each DNN and each behavior, five models were trained separately. The prediction results were then averaged across the five models. All four behavioral measures were z-normalized based on training data. The loss function for sex prediction was cross entropy, i.e., the output layer for sex prediction have 2 nodes (FNN and BrainNetCNN) or filters (GCNN). The loss function was MSE for the other three measures. The output layer for these three measures have 1 node (FNN and BrainNetCNN) or filter (GCNN). Adam (Kingma and Ba, 2015) or SGD were used. See details in Tables 2, 3 and 4.

- For all DNNs, model was tuned for each behavior separately. Tables 2, 3 and 4 show the final DNN structures
- Final FNN structure is shown in table 2. For FNN, dropout of 0.2/0.3/0.4/0.4 (for sex/age/pairs matching/fluid intelligence respectively) was added before each fully-connected layer. L2 regularization of 0.02 was added for layer 2. Weight decay of 0.01/0.01/0.001/0.016 (for sex/age/pairs matching/fluid intelligence respectively) were applied to the weights of all fully connected layers.
- Final BrainNetCNN structure is shown in table 3. For BrainNetCNN, dropout of 0.21/0.6/0.25/0.54 (for sex/age/pairs matching/fluid intelligence respectively) was added after the E2E, E2N, and N2G layers. LeakyReLU was replaced by linear activation for all four models.
- Final GCNN structure is shown in table 4. Dropout of 0.3/0.6/0.6/0.7 (for sex/age/pairs matching/fluid intelligence respectively) was added for each layer. L2 regularization of 2e-5/2e-4/2e-4/2e-6 (for sex/age/pairs matching/fluid intelligence respectively) was added for layer 1. The nodes of the graph corresponded to subjects. Edges were constructed based on Pearson's correlation between subjects' vectorized RSFC. Thresholding of the graph was tuned separately for each behavior or demographic measure. For sex prediction, the top five correlated edges of each node were retained. For age, pairs matching and fluid intelligence prediction, the graph was thresholded by only retaining edges with top 5% correlation (across the entire graph). Furthermore, the top five correlated edges of each node were also retained (even if these edges were not among the top 5% correlated edges). The graph convolution filters for all four GCNNs were estimated by a 1-degree Chebyshev polynomial (Defferrard et al., 2016).

Dataset	Predicting	Model structure	Optimizer
HCP	58 behaviors	224, 128, 192, 58	SGD
UK Biobank	Sex	8, 32, 2	SGD
	Age	8, 8, 1	SGD
	Pairs matching	16, 384, 1	SGD
	Fluid intelligence	32, 32, 1	SGD

Table 2. FNN structure and hyperparameter settings for HCP and UK Biobank. Under "Model structure", the numbers represent the number of nodes at each fully connected layer. For example, "256, 96, 256, 58" represents a 4-layer FNN with 256, 96, 256 and 58 nodes.

Dataset	Predicting	Model structure	Optimizer
HCP	58 behaviors	16, 128, 26, 58	SGD
UK Biobank	Sex	15, 93, 106, 2	SGD
	Age	32, 92, 24, 1	SGD
	Pairs matching	30, 72, 96, 1	SGD
	Fluid intelligence	37, 40, 34, 1	SGD

Table 3. BrainNetCNN structure and hyperparameter settings for HCP and UK Biobank. Under "Model structure", the numbers represent the number of filters or nodes at each layer. For example, "15, 93, 106, 2" represents a BrainNetCNN with 15 filters for the E2E layer, 93 filters for the E2N layer, 106 filters (nodes) for the N2G layer and 2 nodes in the final fully connected layer. All BrainNetCNNs follow the same layer order: E2E, E2N, N2G and then a final fully connected layer.

Dataset	Predicting	Model structure	Optimizer
HCP	58 behaviors	256, 58	SGD
UK Biobank	Sex	6, 2	Adam
	Age	64, 1	SGD
	Pairs matching	20, 1	Adam
	Fluid intelligence	64, 1	Adam

Table 4. GCNN structure and hyperparameter settings for HCP and UK Biobank. Under "Model structure", the numbers represent the number of filters for each graph convolutional layer. For example, "64, 1" represents a 2-layer GCNN with 64 and 1 filters respectively.

Supplementary Materials

Description	HCP field
Visual Episodic Memory	PicSeq_Unadj
Cognitive Flexibility (DCCS)	CardSort_Unadj
Inhibition (Flanker Task)	Flanker_Unadj
Fluid Intelligence (PMAT)	PMAT24_A_CR
Vocabulary (Pronunciation)	ReadEng_Unadj
Vocabulary (Picture Matching)	PicVocab_Unadj
Processing Speed	ProcSpeed_Unadj
Delay Discounting	DDic_AUC_40K
Spatial Orientation	VSPLOT_TC
Sustained Attention – Sens.	SCPT_SEN
Sustained Attention – Spec.	SCPT_SPEC
Verbal Episodic Memory	IWRD_TOT
Working Memory (List Sorting)	ListSort_Unadj
Cognitive Status (MMSE)	MMSE_Score
Sleep Quality (PSQI)	PSQI_Score
Walking Endurance	Endurance_Unadj
Walking Speed	GaitSpeed_Unadj
Manual Dexterity	Dexterity_Unadj
Grip Strength	Strength_Unadj
Odor Identification	Odor_Unadj
Pain Interference Survey	PainInterf_Tscore
Taste Intensity	Taste_Unadj
Contrast Sensitivity	Mars_Final
Emotional Face Matching	Emotion_Task_Face_Acc
Arithmetic	Language_Task_Math_Avg_Difficulty_Level
Story Comprehension	Language_Task_Story_Avg_Difficulty_Level
Relational Processing	Relational_Task_Acc
Social Cognition – Random	Social_Task_Perc_Random
Social Cognition – Interaction	Social_Task_Perc_TOM

Table S1. Table showing original HCP variable names and corresponding descriptive labels used in the manuscript.

Description	HCP field
Working Memory (N-back)	WM_Task_Acc
Agreeableness (NEO)	NEOFAC_A
Openness (NEO)	NEOFAC_O
Conscientiousness (NEO)	NEOFAC_C
Neuroticism (NEO)	NEOFAC_N
Extraversion (NEO)	NEOFAC_E
Emot. Recog. – Total	ER40_CR
Emot. Recog. – Angry	ER40ANG
Emot. Recog. – Fear	ER40FEAR
Emot. Recog. – Happy	ER40HAP
Emot. Recog. - Neutral	ER40NOE
Emot. Recog. – Sad	ER40SAD
Anger – Affect	AngAffect_Unadj
Anger – Hostility	AngHostil_Unadj
Anger – Aggression	AngAggr_Unadj
Fear – Affect	FearAffect_Unadj
Fear – Somatic Arousal	FearSomat_Unadj
Sadness	Sadness_Unadj
Life Satisfaction	LifeSatisf_Unadj
Meaning & Purpose	MeanPurp_Unadj
Positive Affect	PosAffect_Unadj
Friendship	Friendship_Unadj
Loneliness	Loneliness_Unadj
Perceived Hostility	PercHostil_Unadj
Perceived Rejection	PercReject_Unadj
Emotional Support	EmotSupp_Unadj
Instrument Support	InstruSupp_Unadj
Perceived Stress	PercStress_Unadj
Self-Efficacy	SelfEff_Unadj

Table S1 (cont.). Table showing original HCP variable names and corresponding descriptive labels used in the manuscript.

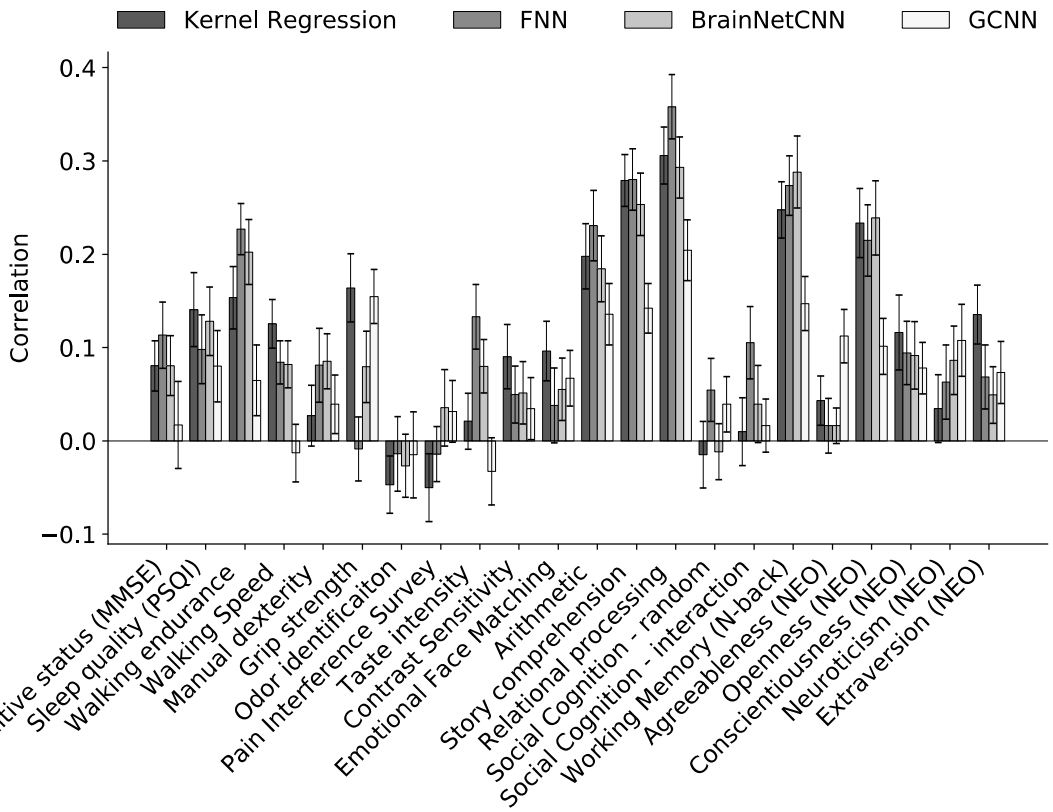

Figure S1. Prediction accuracy (correlation) of 22 HCP measures averaged across 20 test folds. Correlation was computed for each test fold and each behavior. Bars show mean across test folds. Error bars show standard errors.

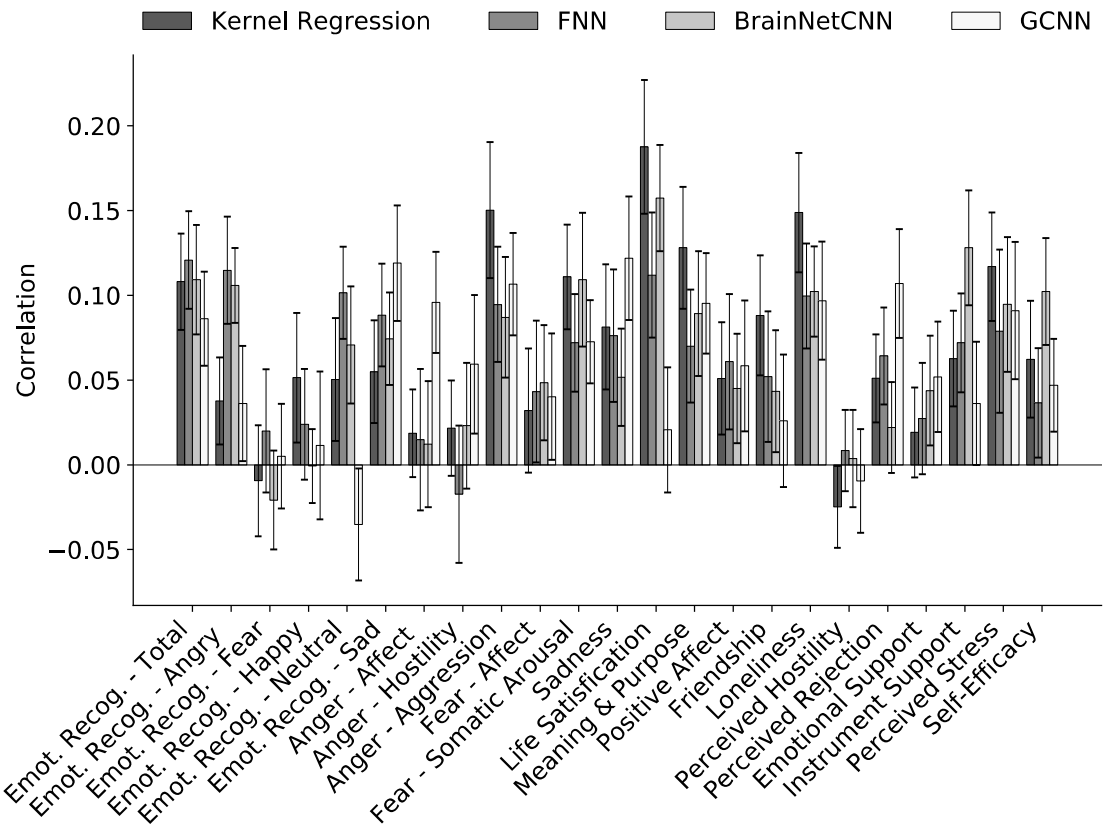

Figure S2. Prediction accuracy (correlation) of 23 HCP cognitive measures averaged across 20 test folds. Correlation was computed for each test fold and each behavior. Bars show mean across test folds. Error bars show standard errors.