Title Integrating Task-Based Functional MRI Across Tasks Markedly Boosts Prediction and Reliability of Brain-Cognition Relationship

Abbreviated Title: Task-Based Functional MRI and Cognitive Abilities

Alina Tetereva[a] http://orcid.org/0000-0003-2077-628X
Jean Li[b] https://orcid.org/0000-0002-5545-8102
Jeremiah Deng[b] https://orcid.org/0000-0003-3727-4403
Argyris Stringaris[c] http://orcid.org/0000-0002-6264-8377
Narun Pat[a] http://orcid.org/0000-0003-1459-5255

[a]Department of Psychology, University of Otago, New Zealand, 9016
[b]Department of Information Science, University of Otago, New Zealand, 9016
[c]Section on Clinical and Computational Psychiatry, National Institute of Mental Health, USA, 20892

Corresponding Authors:
Alina Tetereva, MS and
Narun Pat, PhD, also known as Narun Pornpattananangkul
Department of Psychology, University of Otago
William James Building
275 Leith Walk
Dunedin 9016, New Zealand
Email: alina.tetereva@postgrad.otago.ac.nz and narun.pat@otago.ac.nz

Number of pages: 27
Number of figures: 6
Number of tables: 5
Number of words for abstract: 248, introduction: 699, and discussion: 1,498

Conflict of interest statement: The authors declare no competing interests.

Acknowledgements
Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The authors thank Sam Harrison, PhD, for his inputs on data analyses. A.T. and N.P. were supported by Health Research Council Funding (21/618) and by University of Otago.

Abstract

Capturing individual differences in cognitive abilities is central to human neuroscience. Yet our ability to estimate cognitive abilities via brain MRI is still poor in both prediction and reliability. Our study tested if this inability was partly due to the over-reliance on 1) non-task MRI modalities and 2) single modalities. We directly compared predictive models comprising of different sets of MRI modalities (e.g., task vs. non-task). Using the Human Connectome Project (n=873 humans, 473 females, after exclusions), we integrated task-based functional MRI (tfMRI) across seven tasks along with other non-task MRI modalities (structural MRI, resting-state functional connectivity) via a machine-learning "stacking" approach. The model integrating all modalities provided unprecedented prediction (r=.581) and excellent test-retest reliability (ICC>.75) in capturing general cognitive abilities. Importantly, comparing to the model integrating among non-task modalities (r=.367), integrating tfMRI across tasks led to significantly higher prediction (r=.544) while still providing excellent test-retest reliability (ICC>.75). The model integrating tfMRI across tasks was driven by areas in the frontoparietal network and by tasks that are cognition-related (working-memory, relational processing, and language). This result is consistent with the parieto-frontal integration theory of intelligence. Accordingly, our results sharply contradict the recently popular notion that tfMRI is not appropriate for capturing individual differences in cognition. Instead, our study suggests that tfMRI, when used appropriately (i.e., by drawing information across the whole brain and across tasks and by integrating with other modalities), provides predictive and reliable sources of information for individual differences in cognitive abilities, more so than non-task modalities.

Keywords: reliability, prediction, task-based functional MRI, general cognitive abilities, individual differences

Significance Statement

Studies of individual differences in the brain-cognition relationship is dominated by a single, non-task MRI modality, e.g., structural MRI and resting-state functional connectivity. Task-based functional MRI (tfMRI) has come under scrutiny and deemed non-suitable for capturing individual differences. Here we demonstrate that tfMRI, when used appropriately, can provide unique and important sources of information about individual differences in cognitive abilities. Using machine-learning, we propose an approach to draw information from tfMRI across regions from the whole brain and across tasks and to combine this information with other modalities. This results in an interpretable, brain-based predictive model of general cognitive abilities with unprecedented levels of prediction and test-retest reliability. This facilitates the improvement of our ability to capture the brain-cognition relationship.

Introduction

Relating individual differences in cognitive abilities with the brain has been focal to human neuroscience (Deary et al., 2010). Yet, we still cannot use brain data to capture individual differences in cognitive abilities with high prediction and reliability (Pohl et al., 2019; Sui et al., 2020). Here, prediction refers to an ability to estimate cognitive abilities of unseen individuals (outside of the model-building process, aka out-of-sample) (Yarkoni & Westfall, 2017). Reliability refers to test-retest stability of measurement (Noble et al., 2021). This failure has led to a headline, such as 'Scanning the Brain to Predict Behavior, a Daunting 'Task' for MRI' (APS, 2020). Having highly predictive and reliable brain-based biomarkers could aid-in our studies of mental-illness mechanisms (Morris & Cuthbert, 2012).

Predicting out-of-sample individual-differences in cognitive abilities from neuroimaging has predominately been focused on non-task MRI modalities. Earlier studies associating general cognitive abilities and structural MRI (sMRI; reflecting brain volume/morphology) showed a weak association at *r* .1-.3 (McDaniel, 2005; Pietschnig et al., 2015). Because these associations were often done within-sample (not tested on unseen individuals), these weak associations may have already been biased upward (i.e., overfitting). Indeed, a recent competition showed *r* as low as .03 for out-of-sample prediction of general cognitive abilities via sMRI in children (Mihalik et al., 2019). Recently, researchers have turned to resting-state functional connectivity (resting-state FC) for prediction. Resting-state FC reflects the functional-connectivity between different brain areas, intrinsically occurring while resting. Using resting-state FC, researchers have found moderate out-of-sample prediction of general cognitive abilities at *r* .2-.4 (Dubois et al., 2018; Rasero et al., 2021; Sripada et al., 2020). Still, there is a large room for improvement.

Here we examined two potential solutions: 1) task-based functional MRI (tfMRI) and 2) stacking. First, tfMRI reflects the changes in BOLD induced by certain events while performing cognitive tasks. One study (Sripada et al., 2020) tested the prediction of specific tfMRI tasks using the Human Connectome Project (HCP) (Barch et al., 2013). Here, tfMRI from some tasks (e.g., taping working memory, relational skills and language) predicted cognitive abilities very well, at out-of-sample $r>.4$, which is higher than prediction from resting-state FC. This suggests that task-based activation during certain cognitive processes is a better candidate for capturing individual differences in cognitive abilities, compared to more commonly used modalities, e.g., sMRI and resting-state FC.

Nonetheless, tfMRI has recently come under intense scrutiny for its low reliability (Elliott et al., 2020). Researchers often quantify reliability using intraclass correlation (ICC), where low ICC reflects poor reliability (Cicchetti & Sparrow, 1981). Elliot and colleagues (2020) examined ICC of task-based activation at different regions and tasks using the HCP and showed poor ICC (<.4) across the regions and tasks. This is very different from sMRI's ICC, which was at the excellent range (ICC>.75). Accordingly, while providing better prediction, tfMRI may not be stable across time. Thus, this calls for research to boost reliability of tfMRI, e.g., via machine-learning (Kragel et al., 2021).

The second solution involves a machine-learning technique called "stacking"(Wolpert, 1992). Most studies rely on a single MRI modality to predict cognitive abilities (Sui et al., 2020). Stacking allows scientists to combine different MRI modalities (Engemann et al., 2020; Rasero et al., 2021). For instance, Rasero and colleagues (2021) used the HCP and combined many non-task MRI modalities (e.g., sMRI and resting-state FC) via stacking and showed enhanced predictive performance. However, potentially partly due to not including tfMRI into their stacked model, they only found R^2=.078, or roughly estimated r=.28, from their stacked model. Thus, the question arises, can integrating tfMRI across different tasks and/or with other modalities via stacking improve performance?

We asked whether the lack of prediction and reliability in previous studies attempting to capture general cognitive abilities was partly due to the over-reliance on 1) non-task MRI modalities and 2) single modalities. We directly compared predictive models comprising of different sets of MRI modalities (e.g., task vs. non-task) and a single modality. We expected to see high prediction and reliability from stacked, tfMRI models. Beyond prediction and reliability, we designed our machine-learning pipeline to be interpretable, such that we could examine the role of frontoparietal regions in keeping with the parieto-frontal integration theory of intelligence (Jung & Haier, 2007).

Materials and Methods

Participants

We used the Human Connectome Project's (HCP) S1200 release (Van Essen et al., 2013; WU-Minn Consortium Human Connectome Project, 2018). This release included multimodal-MRI and cognitive-performance data from 1,206 healthy participants (not diagnosed with psychiatric and neurological disorders). We discarded participants whose data were flagged with having quality control issues by the HCP (n=91): either having the "A" (anatomical anomalies) or "B" (segmentation and surface) flag or having known major issues (Elam, 2021). We also removed participants who had missing values in any of the multimodal-MRI (n= 233) or cognitive-ability (n=9) variables. This left 873 participants (473 females, *M*=28.7 (*SD*=3.7) years old). They are from 414 families as many participants are from the same family. We provided participants' ID on our GitHub repository (see below). Participants provided informed consent, including consent to share de-identified data. The Institutional Review Board at Washington University oversighted the HCP's study procedure.

To examine test-retest reliability of our predictive models, we also used the HCP Retest Dataset. This dataset included 45 participants who completed the HCP protocol for the second time (*M*=139.029 (SD=67.31) days apart). We had 34 participants whose data were complete across the two visits and were not flagged with having quality control issues.

Features: Multimodal MRI

The HCP provided complete details of the scanning parameters and preprocessing pipeline elsewhere (Barch et al., 2013; Glasser et al., 2013; Van Essen et al., 2013). Here, we used MRI data with the MSMAll alignment (Glasser et al., 2016; Robinson et al., 2018) and with extensive

processing (e.g., for task-based functional MRI, we obtained the general linear contrasts, see below). In total, the MRI data can be organized into 12 different modalities (i.e., sets of features):

1-7) Task-based functional MRI (tfMRI) from seven different tasks.

The HCP collected tfMRI from seven tasks (Barch et al., 2013), giving rise to seven sets of features in our model. The study scanned participants in each of the task twice with different phase encodings: right-to-left (RL) and left-to-right (LR). The HCP described preprocessing steps for tfMRI elsewhere (Glasser et al., 2013). Briefly, they included B_0 distortion correction, motion correction, gradient unwrap, boundary-based co-registration to T_1-weighted image, non-linear registration to MNI152 space, grand-mean intensity normalization and surface generation (see https://github.com/Washington-University/HCPpipelines). We parcellated tfMRI into 379 regions of interest (ROIs) using Glasser cortical atlas (360 ROIs) (Glasser et al., 2016) and Freesurfer subcortical segmentation (19 ROIs) (Fischl et al., 2002) and extracted the average value from each ROI. We treated general-linear model contrasts between standard experimental vs. control conditions for each tfMRI task as different modalities:

First, in the working memory task, we used the 2-back vs. 0-back contrast. Here, participants had to indicate whether the stimulus currently shown is the same as the stimulus shown two trials prior [2-back] or as the target stimulus shown in the beginning of that block [0-back]. Second, in the language task, we used the story vs. math contrast. Here, participants responded to questions about Aesop's fables [story] or math problems [math]. Third, in the relational processing task, we used the relational vs. match contrast. Here participants reported if two pairs of objects differed in the same dimension [relational] or matched with a given dimension [match]. Forth, in the motor task, we used the averaged movement vs. cue contrast. Here participants were prompted [cue] to subsequently execute a movement [movement] with their fingers, toes, and tongue. Fifth, in the emotion processing task, we used the face vs. shape contrast. Here participants decided which two of the bottom objects matched the top object, and all objects in each trial can either be (emotional) faces [face] or shapes [shape]. Sixth, in the social cognition task, we used the theory of mind vs. random contrast. Here participants saw movie clips of objects interacting with each other either socially [theory of mind] or randomly [random]. Seventh, in the gambling task, we used the reward vs. punishment contrast. Here, participants had to guess if a number was higher or lower than 5, and the correct guess was associated with winning (vs. losing) money. They mostly won in certain blocks [reward] and mostly lost in others [punishment].

8) Resting-state functional connectivity (resting-state FC)

The HCP collected resting-state FC from four 15-min runs, resulting in one-hour-long data (Glasser et al., 2013; Smith et al., 2013). Half of the runs were right-to-left phase encoding, and the other half were left-to-right phase encoding. The HCP applied similar a preprocessing pipeline (Glasser et al., 2013) with tfMRI (see https://github.com/Washington-University/HCPpipelines). We used ICA-FIX denoised resting-state FC data (Glasser et al., 2016) and parcellated them into 379 ROIs using the same atlases with the tfMRI (Fischl et al., 2002; Glasser et al., 2016). After the parcellation, we extracted each ROI's time series from each of the four runs and concatenated them into one. We then computed Pearson's correlation between concatenated time series of each ROI pair, resulting in a table of 71,631 non-overlapping resting-state FC connectivity indices.

Thereafter, we applied *r-to-z* transformation to the whole table. To reduce the number of features in the model, we applied a univariate feature filtering approach (Dubois et al., 2018; Finn et al., 2015). Here, we correlated each non-overlapping resting-state FC connectivity index with the target, general cognitive abilities, and removed all indices that demonstrated Pearson correlation with $p > 0.01$. This left 2,908 resting-state FC connectivity indices as the final features for resting-state FC. We used the same set of resting-state FC connectivity indices for the test-retest reliability.

9-12) Structural MRI (sMRI)

The HCP provided preprocessing pipeline for sMRI elsewhere (Glasser et al., 2013). Please see the preprocessing scripts here https://github.com/Washington-University/HCPpipelines. We separated sMRI data into four different modalities: cortical thickness, cortical surface area, subcortical volume and total brain volume. For cortical thickness and cortical surface area, we used Destrieux parcellation (148 ROIs) from FreeSurfer's aparc.stats file (Destrieux et al., 2010; Fischl, 2012). As for subcortical volume, we used subcortical segmentation (19 gray matter ROIs) from FreeSurfer's aseg.stats file (Fischl et al., 2002). As for total brain volume, we included five features calculated by FreeSurfer: estimated intra-cranial volume (FS_IntraCranial_Vol), total cortical gray matter volume (FS_TotCort_GM_Vol), total cortical white matter volume (FS_Tot_WM_Vol), total subcortical gray matter volume (FS_SubCort_GM_Vol) and ratio of brain segmentation volume to estimated total intracranial volume (FS_BrainSegVol_eTIV_Ratio).

Target: General Cognitive Abilities

We trained our models to predict general cognitive abilities, reflected by the average score of cognition assessments in the NIH Toolbox (Weintraub et al., 2014), as provided by the HCP (CogTotalComp_Unadj). The assessments included picture sequence memory, Flanker, list sorting, dimensional change card sort, pattern comparison, reading tests and picture vocabulary. Note we used the age-unadjusted average score since we controlled for age in the models themselves (see below).

Confound Correction

We first controlled for age (Dosenbach et al., 2010; Geerligs et al., 2015) and gender (Ruigrok et al., 2014; Trabzuni et al., 2013) in our models by linearly residualising them from both MRI data and cognitive abilities. We additionally residualised in-scanner movements from tfMRI and resting-state FC, given their sensitivity to motion artifacts (Power et al., 2012; Satterthwaite et al., 2013). More specifically, we defined in-scanner movements as the average of relative displacement (Movement_RelativeRMS_mean) across all available runs for each modality separately.

Predictive Modeling Pipeline: Stacking

For our predictive modelling pipeline (Figure 1), we used nested cross-validation (CV) to build the stacked models and evaluate their predictive performance. Since the HCP recruited many participants from the same family (Van Essen et al., 2013; WU-Minn Consortium Human Connectome Project, 2018), we first controlled the influences of the family structure by splitting

the data into eight folds based on families. In each of the folds, there were members of ~50 families, prohibiting members of the same family to be in different folds.

The nested CV involved two loops, nested with each other. In each CV "outer" loop, one of the eight folds that included ~50 families (~105 participants) was held-out. The rest was further split into 60% and 40% for the first- and second-layer training layers, respectively. Within the CV "inner" loops, we separately fit the first-layer data from each modality to predict general cognitive abilities. Here we applied a five-fold CV to tune hyperparameters of the models. This stage allowed us to create 12 modality-specific models. Using the second-layer data, we then computed predicted values for each of the 12 modalities based on the modality-specific models, and fit these predicted values across modalities to predict general cognitive abilities. Same as the first-layer data, we also applied a five-fold CV to tune hyperparameters of the models here. This stage allowed us to create three stacked models: 1) all-modality stacked model (i.e., combination of 12 modalities), 2) task stacked model (i.e., combination of seven different tfMRI tasks) and 3) non-task stacked models (i.e., combination of resting-state FC and four sMRI modalities).

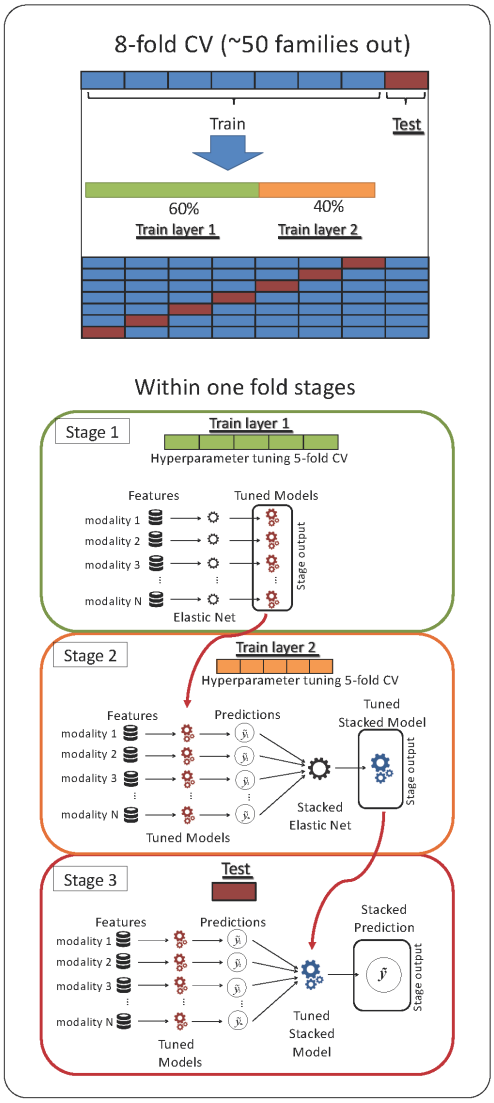


Figure 1. Predictive modelling pipeline. The diagram shows how we built stacked models and evaluated their predictive performance. CV = cross validation.

Predictive Modeling Algorithm: Elastic Net

Similar to previous work (Dubois et al., 2018), we used Elastic Net (Zou & Hastie, 2005) as the model-fitting algorithm via the scikit-learn package (Pedregosa et al., 2011). Elastic Net is a general form of penalized regression, allowing us to draw information across different brain indices simultaneously to predict one target variable. Compared to the classical, ordinary least squares regression, Elastic Net allows us to have more parameters (e.g., number of brain indices) than the number of observations (e.g., participants in each training set). Resting-state FC, for instance, often has the number of parameters (each pair of brain regions) higher than the number of participants. Compared to other more complicated algorithms, Elastic Net has the benefit of being easier to interpret (Molnar, 2019). Researchers can directly interpret the magnitude of each coefficient as the importance of each feature (e.g., brain indices).

Elastic Net fits a plane that minimises the squared distance between itself and the data points (James et al., 2021; Kuhn & Johnson, 2013). When strongly correlated features are present, the classical, ordinary least squares regression tends to give very unstable estimates of coefficients and extremely high estimates of model uncertainty (Alin, 2010; Graham, 2003; Monti, 2011; P. Vatcheva & Lee, 2016). To address this, Elastic Net simultaneously minimises the weighted sum of the features' slopes. For example, if the features are tfMRI from different regions, Elastic Net will shrink the contribution of some regions closer towards zero. The degree of penalty to the sum of the feature's slopes is determined by a shrinkage hyperparameter a: the greater the a, the more the slopes shrunk, and the more regularised the model becomes. Elastic Net also includes another hyperparameter, the 'ℓ_1 ratio', which determines the degree to which the sum of either the squared (known as 'Ridge'; ℓ_1 ratio=0) or absolute (known as 'Lasso'; ℓ_1 ratio=1) slopes is penalised (Zou & Hastie, 2005). The objective function of Elastic Net as implemented by sklearn is defined as:

$$argmin_w \left(\frac{\left|\left|y - Xw\right|\right|_2^2}{2 \times n_{samples}} + \alpha \times \ell_1_ratio \times \left|\left|w\right|\right|_1 + 0.5 \times \alpha \times (1 - \ell_1_ratio) \times \left|\left|w\right|\right|_2^2 \right)$$

where X is the predictor variable, y is the target variable, and w is the coefficient.

To find the appropriate hyperparameters for each training layer, we applied grid-search, 5-fold cross-validation (Efron & Gong, 1983; Hawkins et al., 2003; Koul et al., 2018) separately on each layer. In our grid, we searched for a ranged between 10^{-6} and 10^{4}, sampled on log scale with 500 values between -6 and 4, whereas we used 100 numbers in a linear space for the ℓ_1 ratio, ranging from 0 and 1.

Predictive Modeling Performance: Prediction

To evaluate models' prediction, we used the eight held-out folds across the outer CV loops. Using these held-out folds, we first computed predicted values from each model. We then tested how similar these predicted values were to the real, observed values, using four measures (Poldrack

et al., 2020). First, Pearson's *r* is defined as $\frac{cov(y,\hat{y})}{\sigma_y \sigma_{\hat{y}}}$, where cov is the covariance, σ is the standard deviation, y is the observed value and ŷ is the predicted value. Pearson's *r* ranges from -1 to 1. The high positive Pearson's *r* reflects high predictive accuracy, regardless of scale. Negative *r* reflects that no predictive information is present in the model. Second, coefficient of determination (R^2) is defined using the sum-of-squared formulation, $1 - \frac{\sum_i(\hat{y}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2}$, where $\bar{y}$ is the mean of the observed value. R^2 is often interpreted as variance explained, with the value closer to 1 reflecting high predictive accuracy. Like Pearson's *r,* R^2 can be negative in case of no predictive information in the model. Note that we did not use the squared Pearson's *r* definition of R^2*,* which is not appropriate in the context of out-of-sample prediction (Poldrack et al., 2020), given that it wrongly converts negative Pearson's *r* into a positive R^2. Third, the mean squared error (MSE) is defined as $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. MSE is sensitive to scaling and is often used to compare models across different algorithms/features. Lower MSE reflects high predictive accuracy. Forth, the mean absolute error (MAE) is defined as $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$. MAE is similar to MSE, but given the use of absolute (as opposed to squared) values, MAE can be more robust to outliers. Using multiple measures of predictive performance is highly recommended to reveal different aspects of the models (Poldrack et al., 2020).

To statistically compare measures of predictive performance across models, we combined predicted and observed values across the eight held-out folds and computed the four measures of predictive performance. We then created bootstrapped distributions (Efron & Tibshirani, 1993) of the difference between each pair of models in their prediction. If the 95% confidence interval of the distributions did not include zero, we concluded that the two models were significantly different from each other.

Predictive Modeling Feature Importance: Elastic Net Coefficients

We used the Elastic Net coefficients to locate 1) which of the modalities contributed highly to the prediction of the stacked models and 2) which of the brain indices contributed highly to the prediction of the modality-specific models. More specifically, once we showed which of the 12 modalities contributed highly to the prediction of all-modality stacked model, we would then investigate the brain indices of the top-performing modality-specific models that contributed highly to the prediction. In addition to plotting the coefficients on the brain images, we also provided a list of top-20 brain indices for each top-performing modality-specific model. We evaluated contribution of each brain index based on the magnitude of their Elastic Net coefficient. In this list, we identified brain networks associated with each brain region using the Cole-Anticevic Brain Network Atlas (Glasser et al., 2016; Ji et al., 2019). We also provided MNI coordinates for each region, obtained by transforming voxel coordinates (based on https://neuroimaging-core-docs.readthedocs.io/en/latest/pages/atlases.html) to the MNI space via nilearn.image.coord_transform() using the standard FSL template, MNI152_T1_1mm, as a reference.

To represent contributing areas across tfMRI tasks, we combined the magnitude of Elastic Net coefficients from all tasks at each brain area, weighted by the overall magnitude of Elastic Net

coefficients of the task stacked model. Mapping the contributing areas across tfMRI tasks allowed us to demonstrate brain activity across different cognitive domains (represented by tasks) that are related to general cognitive abilities, thereby examining the role of brain areas implicated in theories such as the parieto-frontal integration theory of intelligence (Jung & Haier, 2007). This is akin to previous meta-analyses that combined tfMRI studies with different tasks that were associated with general cognitive abilities to demonstrate brain activity across different cognitive domains (Basten et al., 2015; Jung & Haier, 2007; Santarnecchi et al., 2017). Yet most meta-analyses focused on the consistency in mass-univariate associations, while ours focused on multivariate associations via Elastic Net. Despite the differences in the methodologies, we still demonstrated the overlapping areas between ours and those found in a recent meta-analysis of cognitive abilities (Santarnecchi et al., 2017). Here we downloaded the Activation Likelihood Estimate (ALE) map of significant foci that showed associations with various cognitive abilities (Gf_net.nii from http://www.tmslab.org/netconlab-fluid.php) in MNI, volumetric space. We then converted this ALE map to surface space and overlaid the ALE map on top of the magnitude of Elastic Net coefficients from all tasks using Connectome Workbench (Marcus et al., 2011).

Predictive Modeling Performance: Test-retest reliability

To evaluate models' reliability, we used the data from participants who were scanned twice as a test set (as opposed to using 50 families held out in Figure 1). This allowed us to test how stable the predicted values from different models were across the two-time points, using intraclass correlation (ICC) (Shrout & Fleiss, 1979). ICC is generally defined as $\frac{\sigma_p^2}{\sigma_t^2}$, where σ_p^2 is the between participant variance, and σ_t^2 is the within participant variant. There are two commonly used types of ICC for test-retest reliability in MRI (Noble et al., 2021). First, ICC(2,1) reflects an absolute agreement with random sources of error. ICC(2,1) is defined as $\frac{MS_p - MS_e}{MS_p + (k-1)MS_e + \frac{k}{n}(MS_t - MS_e)}$ where MS_p is mean square for participants, MS_e is mean square for error, MS_t is mean square for time points (i.e., measurements), n is the number of participants, k is the number of time points. Second, ICC (3,1) reflects a consistency with fixed sources of error. ICC (3,1) is defined as $\frac{MS_p - MS_e}{MS_p + (k-1)MS_e}$. We computed both types using the Pingouin package (https://pingouin-stats.org/). Based on an established criterion (Cicchetti & Sparrow, 1981), we considered ICC less than .4 as poor, between .4 and .59 as fair, between .6 and .74 as good and over .75 as excellent.

Code Accessibility

The shell and python scripts used in the analyses are made available here: https://github.com/alinatet/HCP_stacked_ML_cognition_retest

Results

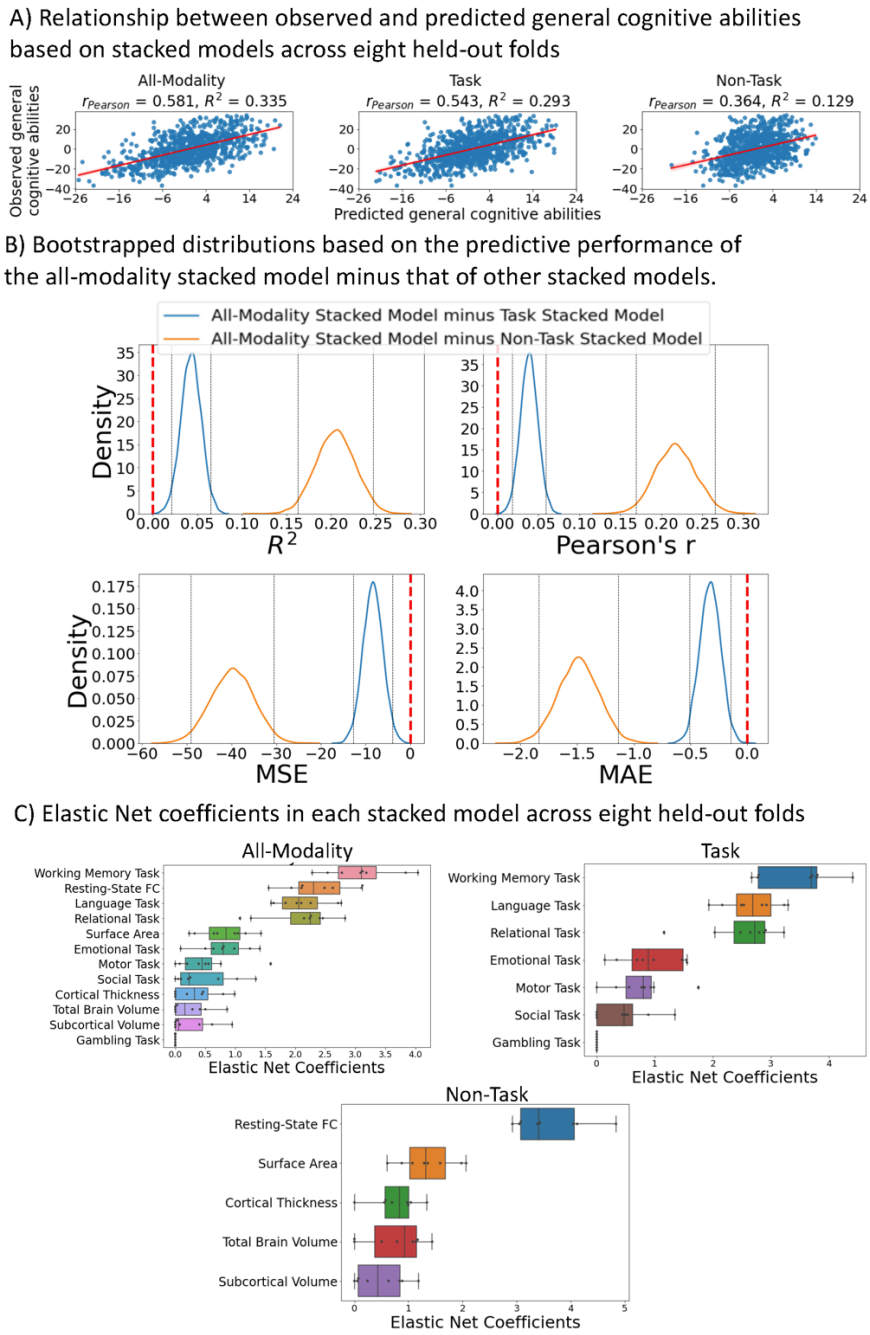
Prediction and Feature Importance for Stacked Models

The all-modality stacked model had the highest predictive performance, compared to other stacked and modality-specific models, reflected by the highest *r* (*M*=.582, *SD*= .045) and R^2

(*M*=.335, *SD*= .048) and lowest MSE (*M*=128.621, *SD*= 9.208) and MAE (*M*=9.079, *SD*= .444) (Figure 2a, 3a). Based on bootstrapping analyses (Figure 2b, 3b), the all-modality stacked model was significantly better in prediction than any other stacked and modality-specific models. Examining the all-modality stacked model's Elastic Net coefficients (Figure 2c) reveals the four main contributing modalities: working memory tfMRI, resting-state FC, language tfMRI and relational tfMRI, respectively.

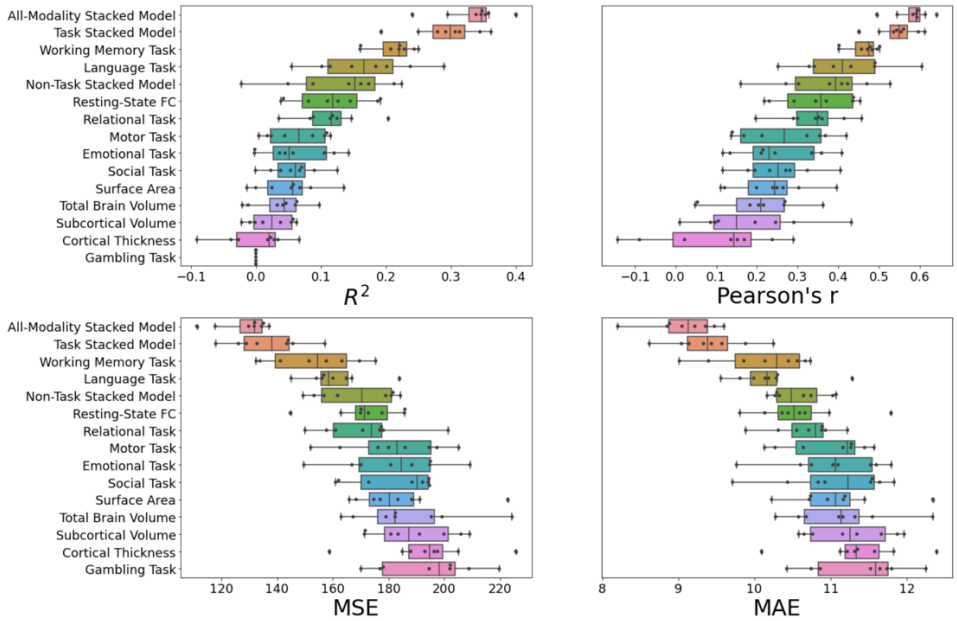
The task stacked model that combined tfMRI from seven different tasks was the second-best model in prediction, reflected by high *r* (*M*=.544, *SD*= .052), R^2 (*M*=.293, *SD*= .054) and low MSE (*M*=136.989, *SD*= 12.740) and MAE (*M*=9.408, *SD*= .507) (Figure 2a, 3a). Based on bootstrapping analyses (Figure 2b), the task stacked model was significantly better in prediction than the non-task stacked model. Based on Elastic Net coefficients (Figure 2c), the three tasks that contributed highest to the task stacked model were working memory, language and relational tasks, respectively.

The non-task stacked model was relatively poorer in prediction with *r* (*M*=.367, *SD*= .118), R^2 (*M*=.129, *SD*= .085) MSE (*M*=168.360, *SD*= 14.492) and MAE (*M*=10.567, *SD*= .356) (Figure 2a, 3a). Numerically, the non-task stacked model was worse in prediction, not only than the two other stacked models, but also than two modality-specific models from tfMRI: working memory and language tasks. Based on Elastic Net coefficients (Figure 2c), the non-task stacked model was mainly driven by resting-state FC.

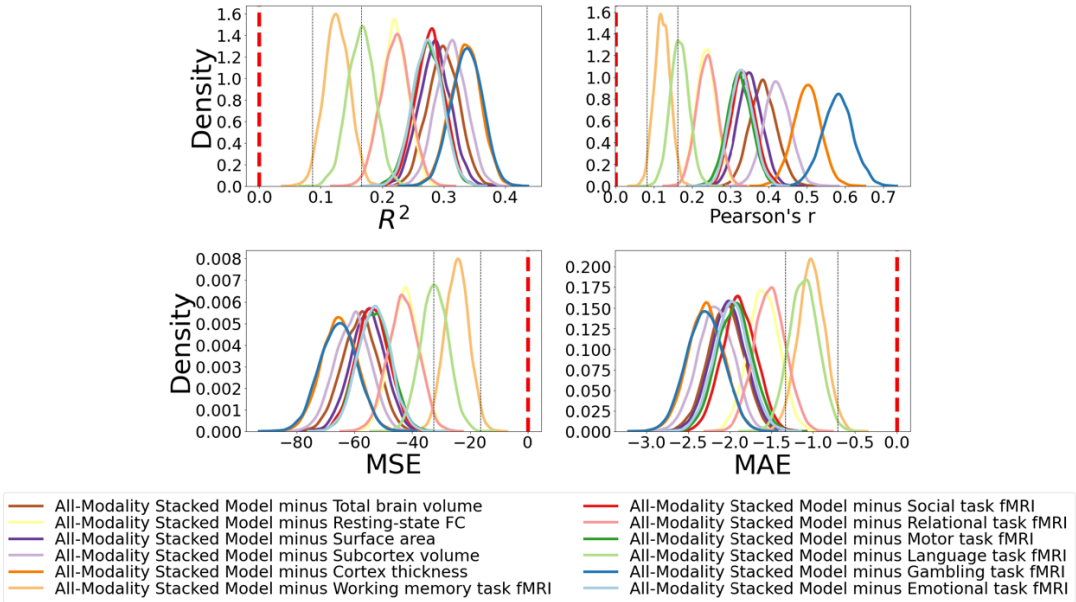


<u>Figure 2. Predictive performance of the three stacked models</u>. 2a shows scatter plots depicting the relationships between predicted and observed values of general cognitive abilities across the eight held-out folds based on each stacked model. 2b shows bootstrapped distribution depicting the differences in the predictive performance between the all-modality stacked model and the other two stacked models. The grey lines indicate 95%CI. If a model's 95%CI does not include zero (red lines), the predictive performance of that model is significantly different from that of all-modality stacked model. R^2 = coefficient of determination; MSE = mean squared error; MAE = mean absolute error. 2c shows Elastic Net coefficients of the three stacked models, indicating which modalities contributed highly to the model's prediction. Each dot represents the Elastic Net coefficient from each of the eight held-out folds.

A) Predictive performance of stacked and modality-specific models across eight held-out folds



B) Bootstrapped distributions based on predictive performance of the all-modality stacked model minus that of other modality-specific models.



<u>Figure 3. Predictive performance of stacked and modality-specific models.</u> 3a shows predictive performance of each model across eight held-out folds. Each dot represents predictive performance from each of the eight held-out folds. 3b shows bootstrapped distribution depicting the differences in the predictive performance between the all-modality stacked model and each modality-specific model. The grey lines indicate 95%CI of the best performing modality-specific model (working-memory tfMRI). If the model's 95%CI does not include zero (red lines), the predictive performance of that model is significantly different from that of all-modality stacked

model. R^2=coefficient of determination; MSE = mean squared error; MAE = mean absolute error.

Among the 12 modality-specific models, working-memory tfMRI, language tfMRI, resting-state FC and relational tfMRI had the highest prediction, respectively (Figure 3a). On the contrary, models based on sMRI and gambling tfMRI had relatively poor prediction (R^2 $M \leq .0512$).

Feature Importance for Modality-Specific Models

Figure 4 shows feature importance of each modality-specific model, as reflected by its Elastic Net coefficients. Table 1 – 4 list 20 features (brain regions/connectivity pair) with the highest Elastic Net magnitude for each of the top-four modalities. We provided a full list of feature importance for all modalities on our GitHub repository.

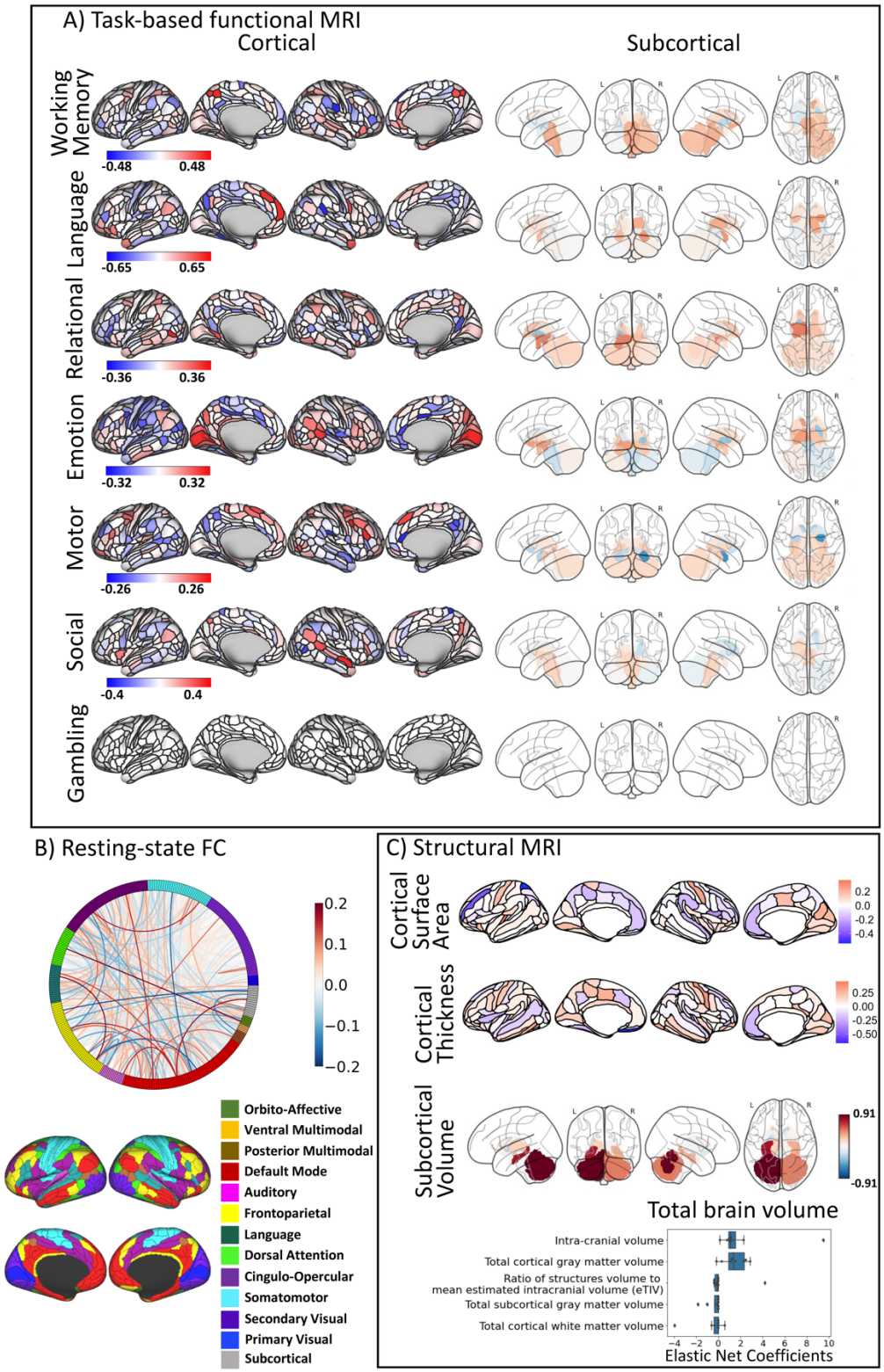


Figure 4. Feature importance of each modality-specific model, as reflected by Elastic Net coefficients. 4a, 4b and 4c show Elastic Net coefficients for task-based functional MRI, resting-state functional MRI and structural MRI, respectively.

For working-memory tfMRI (Figure4, Table 1), we found highly contributing areas from the fronto-parietal, dorsal attention, and default mode networks. These included areas such as superior parietal, inferior frontal, and dorsolateral prefrontal cortices.

Table 1. Top-20 most contributing brain regions for working memory tfMRI. The x,y,z coordinates are in MNI space.

Glasser Label	Brain Region	Network	x	y	z	Coeff M	Coeff SD
L_7Pm	Superior Parietal	Frontoparietal	-5	-68	49	0.44	0.24
R_7Pm	Superior Parietal	Frontoparietal	5	-67	50	0.44	0.15
R_IP2	Inferior Parietal	Frontoparietal	40	-47	45	0.34	0.11
R_AVI	Insular and Frontal Opercular	Frontoparietal	32	25	-4	0.30	0.19
R_d32	Anterior Cingulate and Medial Prefrontal	Frontoparietal	6	39	27	0.28	0.22
R_LIPd	Superior Parietal	Dorsal Attention	30	-55	44	0.31	0.16
L_MIP	Superior Parietal	Dorsal Attention	-26	-67	47	0.29	0.10
R_7PL	Superior Parietal	Dorsal Attention	12	-73	58	0.29	0.24
L_7PL	Superior Parietal	Dorsal Attention	-15	-74	56	0.29	0.14
R_EC	Medial Temporal	Default	20	-11	-27	0.33	0.16
L_PGs	Inferior Parietal	Default	-42	-77	38	-0.28	0.22
L_31pv	Posterior Cingulate	Default	-10	-44	33	-0.28	0.29
R_8BL	Dorsolateral Prefrontal	Default	11	43	48	-0.44	0.19
L_FOP3	Insular and Frontal Opercular	Cingulo-Opercular	-36	3	12	-0.29	0.23
R_IFSa	Inferior Frontal	Cingulo-Opercular	48	39	2	-0.31	0.13
R_PFop	Inferior Parietal	Cingulo-Opercular	62	-20	23	-0.50	0.16
L_PCV	Posterior Cingulate	Posterior Multimodal	-6	-50	48	0.48	0.14
R_PCV	Posterior Cingulate	Posterior Multimodal	5	-52	50	0.40	0.19
L_PIT	Ventral Stream Visual	Visual2	-47	-77	-11	-0.29	0.21
L_6mp	Paracentral Lobular and Mid Cingulate	Somatomotor	-10	-15	69	-0.30	0.24

As for resting-state FC (Figure 4, Table 2), we found that the highly contributing connectivity pairs involved brain regions in the frontoparietal and default-mode networks.

Table 2. Top-20 most contributing connectivity pairs for resting-state FC.

Glasser Label	Network 1	Network 2	Coeff M	Coeff SD
R_PBelt-L_10pp	Auditory	Default	0.15	0.19
R_TE2a-L_PI	Default	Cingulo-Opercular	0.15	0.13
L_31pv-L_PGi	Default	Default	0.12	0.10
L_23d-L_11l	Default	Frontoparietal	-0.12	0.16
R_TE2a-R_TE1m	Default	Frontoparietal	-0.12	0.26
R_STSvp-R_TPOJ3	Default	Posterior Multimodal	0.12	0.12
L_10pp-Diencephalon_ventral_right	Default	subcortex	0.17	0.14
R_6a-L_pOFC	Dorsal Attention	Orbito-Affective	0.20	0.16
R_13l-L_5mv	Frontoparietal	Cingulo-Opercular	-0.13	0.16

L_11l-L_31a	Frontoparietal	Default	-0.17	0.22
R_a10p-L_10pp	Frontoparietal	Default	0.15	0.17
R_13l-R_s32	Frontoparietal	Default	0.12	0.16
L_a10p-L_13l	Frontoparietal	Frontoparietal	0.13	0.14
L_a10p-L_Pir	Frontoparietal	Orbito-Affective	-0.12	0.12
L_AVI-Putamen_left	Frontoparietal	subcortex	-0.18	0.17
R_TPOJ1-L_10pp	Language	Default	0.14	0.13
L_44-L_13l	Language	Frontoparietal	0.19	0.16
L_PCV-Pallidum_left	Posterior Multimodal	subcortex	-0.14	0.13
R_7PC-L_43	Somatomotor	Cingulo-Opercular	0.15	0.16
L_3a-L_s32	Somatomotor	Default	-0.16	0.15

For language tfMRI (Figure 4, Table 3), we found highly contributing areas from frontoparietal and default-mode networks. These included areas such as anterior cingulate, medial prefrontal, insular, orbital and polar frontal, inferior frontal, lateral temporal and dorsolateral prefrontal cortices.

Table 3. Top-20 most contributing brain regions for language tfMRI. The x,y,z coordinates are in MNI space.

Glasser label	Brain Region	Network	x	y	z	Coeff M	Coeff SD
L_8BM	Anterior Cingulate and Medial Prefrontal	Frontoparietal	-6	33	44	0.63	0.43
L_AVI	Insular and Frontal Opercular	Frontoparietal	-31	25	-4	0.40	0.23
L_13l	Orbital and Polar Frontal	Frontoparietal	-23	28	-21	0.26	0.19
R_p10p	Orbital and Polar Frontal	Frontoparietal	23	61	1	-0.26	0.22
L_IFSa	Inferior Frontal	Frontoparietal	-47	33	9	-0.27	0.22
L_7Pm	Superior Parietal	Frontoparietal	-5	-68	49	-0.28	0.07
L_IP1	Inferior Parietal	Frontoparietal	-32	-71	39	-0.28	0.10
L_POS2	Posterior Cingulate	Frontoparietal	-9	-73	37	-0.38	0.12
L_9m	Anterior Cingulate and Medial Prefrontal	Default	-7	54	22	0.65	0.48
R_TGd	Lateral Temporal	Default	35	14	-37	0.48	0.22
L_TGd	Lateral Temporal	Default	-37	12	-37	0.36	0.20
L_47l	Inferior Frontal	Default	-47	29	-12	0.31	0.22
L_PGi	Inferior Parietal	Default	-49	-65	27	0.30	0.21
L_7m	Posterior Cingulate	Default	-6	-62	34	0.29	0.17
L_9p	Dorsolateral Prefrontal	Default	-19	47	38	0.27	0.15
R_6v	Premotor	Somatomotor	58	7	31	0.55	0.26
R_1	Somatosensory and Motor	Somatomotor	48	-22	54	-0.28	0.23
	Amygdala right	Subcortical				0.28	0.08
L_PHA3	Medial Temporal	Dorsal Attention	-34	-35	-21	-0.50	0.30
R_PSL	Temporo-Parieto-Occipital Junction	Cingulo-Opercular	64	-37	27	-0.52	0.22

For relational tfMRI (Figure 4, Table 4), we found highly contributing areas from many networks, e.g., default-mode, visual 2, language, subcortical, frontoparietal and dorsal attention. These included areas such as posterior cingulate, medial temporal, orbital and polar frontal, visual, inferior frontal, dorsolateral prefrontal, and superior parietal cortices.

Table 4. Top-20 most contributing brain regions for relational tfMRI. The x,y,z coordinates are in MNI space.

Glasser Label	Brain Region	Network	x	y	z	Coeff M	Coeff SD
R_7m	Posterior Cingulate	Default	5	-60	33	0.22	0.18
L_23d	Posterior Cingulate	Default	-4	-20	38	-0.19	0.25
L_PHA2	Medial Temporal	Default	-31	-36	-14	-0.21	0.14
R_POS1	Posterior Cingulate	Default	11	-57	15	-0.26	0.21
L_47m	Orbital and Polar Frontal	Default	-37	31	-17	-0.35	0.31
L_FST	MT+ Complex and Neighboring Visual Areas	Visual2	-48	-68	5	0.36	0.21
R_PH	MT+ Complex and Neighboring Visual Areas	Visual2	49	-61	-11	0.18	0.10
L_PIT	Ventral Stream Visual	Visual2	-47	-77	-11	-0.26	0.19
L_PSL	Temporo-Parieto Occipital Junction	Language	-60	-48	25	0.18	0.10
R_45	Inferior Frontal	Language	50	27	0	-0.22	0.19
L_45	Inferior Frontal	Language	-52	27	2	-0.22	0.12
	Diencephalon ventral	subcortex				0.18	0.12
	Amygdala left	subcortex				0.18	0.22
R_IP2	Inferior Parietal	Frontoparietal	40	-47	45	0.23	0.14
R_p9-46v	Dorsolateral Prefrontal	Frontoparietal	46	33	25	0.19	0.08
R_6a	Premotor	Dorsal Attention	26	-2	53	0.21	0.29
L_AIP	Superior Parietal	Dorsal Attention	-40	-39	41	0.21	0.13
L_ProS	Posterior Cingulate	Visual1	-23	-55	3	0.30	0.30
L_FOP2	Insular and Frontal Opercular	Somatomotor	-44	-5	14	-0.22	0.12
R_AAIC	Insular and Frontal Opercular	Orbito-Affective	35	15	-12	0.22	0.07

Figure 5 and Table 5 show contributing brain regions across tfMRI tasks, reflected by the magnitude of Elastic Net coefficients from all tasks at each brain area, weighted by the magnitude of Elastic Net coefficients of the task stacked model. This figure shows the contribution of the areas in the frontoparietal, followed by default, networks to the prediction of general cognitive abilities across tfMRI tasks. Additionally, overlaying the ALE map from a previous meta-analysis of cognitive abilities (Santarnecchi et al., 2017) on top of the contributing brain regions across tasks shows the overlapping regions in the frontoparietal network, in areas such as the left inferior frontal, left anterior cingulate and medial prefrontal and left superior parietal cortices.

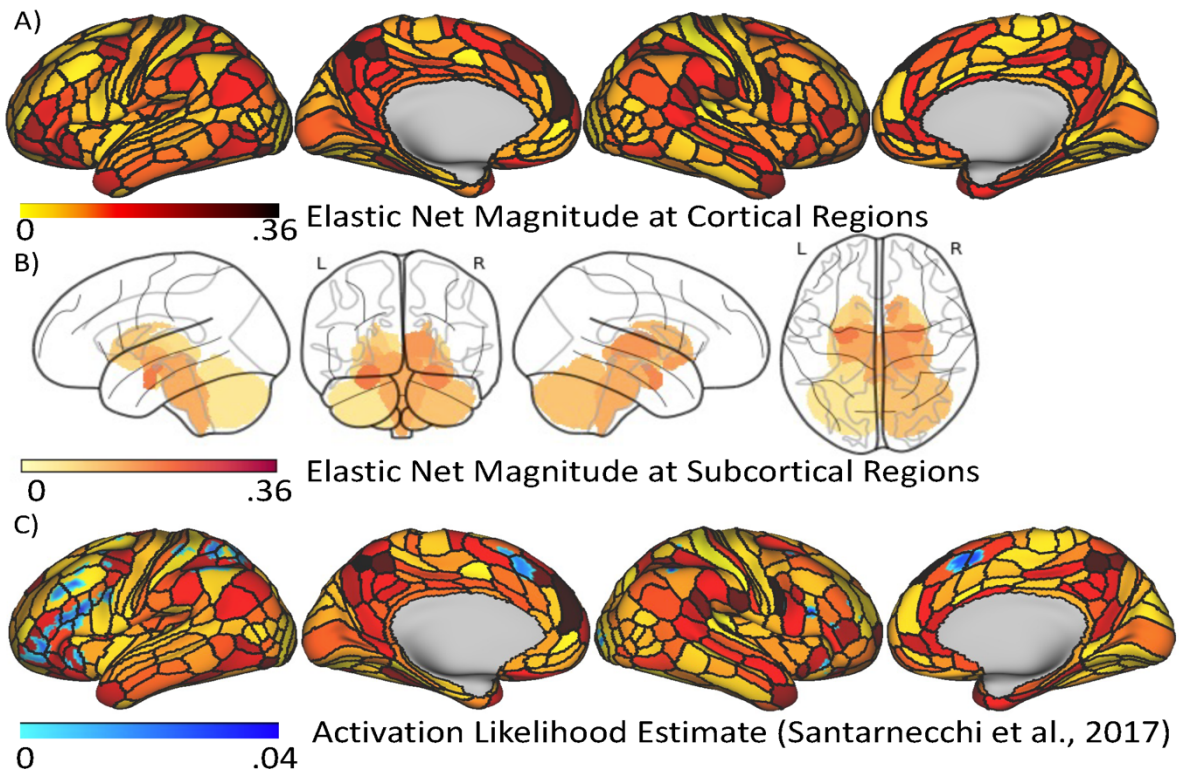


Figure 5. Feature importance of task-based functional MRI (tfMRI) across seven tasks. Here we combined the magnitude of Elastic Net coefficients from all seven tfMRI tasks at each brain area, weighted by the magnitude of Elastic Net coefficients of the task stacked model. A higher value indicates a stronger contribution to the prediction of general cognitive abilities, regardless of the directionality. 5A shows the magnitude at cortical regions while 5B shows the magnitude at subcortical regions. 5C overlays the Activation Likelihood Estimate (ALE) map of the mass-univariate associations with cognitive abilities from a previous meta-analysis (Santarnecchi et al., 2017) on top of the magnitude at the cortical regions.

Table 5. Top-20 contributing brain areas across all tfMR tasks. Here we combined the magnitude of Elastic Net coefficients from all seven tfMRI tasks, weighted by the magnitude of Elastic Net coefficients of the task stacked model.

Glasser Label	Brain Region	Network	x	y	z	Magnitude
L_7Pm	Superior Parietal	Frontoparietal	-5	-68	49	0.36
L_8BM	Anterior Cingulate and Medial Prefrontal	Frontoparietal	-6	33	44	0.30
R_IP2	Inferior Parietal	Frontoparietal	40	-47	45	0.28
R_7Pm	Superior Parietal	Frontoparietal	5	-67	50	0.28
L_IP1	Inferior Parietal	Frontoparietal	-32	-71	39	0.24
L_p47r	Inferior Frontal	Frontoparietal	-45	43	0	0.24
R_p10p	Orbital and Polar Frontal	Frontoparietal	23	61	1	0.24
L_AVI	Insular and Frontal Opercular	Frontoparietal	-31	25	-4	0.24
R_8BL	Dorsolateral Prefrontal	Default	11	43	48	0.36

L_9m	Anterior Cingulate and Medial Prefrontal	Default	-7	54	22	0.34
R_TGd	Lateral Temporal	Default	35	14	-37	0.24
R_PCV	Posterior Cingulate	Posterior Multimodal	5	-52	50	0.33
L_PCV	Posterior Cingulate	Posterior Multimodal	-6	-50	48	0.32
L_PHA3	Medial Temporal	Dorsal Attention	-34	-35	-21	0.28
R_LIPd	Superior Parietal	Dorsal Attention	30	-55	44	0.24
R_PFop	Inferior Parietal	Cingulo-Opercular	62	-20	23	0.30
R_PSL	Temporo-Parieto-Occipital Junction	Cingulo-Opercular	64	-37	27	0.29
L_PIT	Ventral Stream Visual	Visual2	-47	-77	-11	0.29
R_6v	Premotor	Somatomotor	58	7	31	0.35
R_AAIC	Insular and Frontal Opercular	Orbito-Affective	35	15	-12	0.25

Test-retest Reliability

The four sMRI-based models had the highest ICC (> .94) across the two definitions: ICC(2,1) and ICC(3,1) (Figure 6). Similarly, the all-modality stacked and task stacked models had high test-retest reliability, reflected by the excellent level of ICC > .75. The non-task stacked model had the good level of ICC. Modality-specific tfMRI models had ICC varied from poor (gambling, emotional and motor), fair (social and relational) to good (working memory) and excellent (language). The resting-state FC had the fair level of ICC.

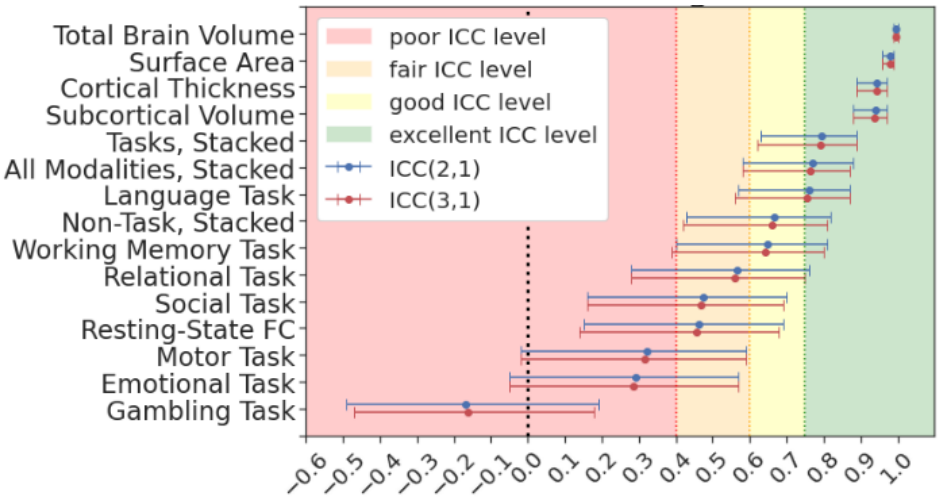


Figure 6. Test-retest reliability of stacked and modality-specific models. We computed test-retest reliability using two definitions of interclass correlation: ICC(2,1) and ICC(3,1). Lines indicates 95%CI.

Discussion

Using tfMRI and stacking, we aim to boost prediction and reliability for brain MRI to capture general cognitive abilities. We directly compared performance of stacked (with or without tfMRI) and modality-specific models. We found that combining tfMRI, resting-state FC and sMRI into the all-modality stacked model gave us the best, unprecedented level of prediction while providing

excellent test-retest reliability. Importantly, this prediction of the all-modality stacked model was primarily driven by many tfMRI tasks. Combining tfMRI across tasks gave us prediction that was closer to the all-modality stacked model and still provided excellent test-retest reliability, showing the importance of tfMRI as an information source for general cognitive abilities. Our use of an interpretable machine-learning algorithm (Elastic Net) allowed us to demonstrate the crucial role of frontoparietal regions across different tfMRI tasks in predicting general cognitive abilities, in line with the parieto-frontal integration theory of intelligence (Jung & Haier, 2007). Conversely, the non-task stacked model that combined sMRI and resting-state FC provided relatively poorer prediction and reliability.

The all-modality stacked model had the highest prediction, compared to other models, across the four measures (having the highest *r* and R^2 and lowest MSE and MAE). This level of prediction is higher relative to those shown in other studies to date (Dubois et al., 2018; McDaniel, 2005; Mihalik et al., 2019; Pietschnig et al., 2015; Rasero et al., 2021; Sripada et al., 2020). Indeed, this level of prediction (r=.58, R^2=.34) is much higher than the performance based on polygenic risk scores from genome-wide association (R^2=.10) (Allegrini et al., 2019). This suggests the potential use of multimodal MRI as a robust biomarker for general cognitive abilities. Note because the performance of tfMRI, resting-state FC and sMRI was consistent with earlier studies (Dubois et al., 2018; Greene et al., 2018; McDaniel, 2005; Pietschnig et al., 2015; Rasero et al., 2021; Sripada et al., 2020), our machine learning pipeline appears to be working as expected. Accordingly, future researchers who need a relatively high predictive and reliable brain-based biomarker for general cognitive abilities could employ our method that takes advantage of all MRI modalities available.

The second-best model in prediction was the task stacked model (r=.54) that combined tfMRI from seven different tasks. This confirms the superior performance of tfMRI shown in a recent study that separately investigated each tfMRI task (Sripada et al., 2020) . Moreover, our results extended this task-specific work (Sripada et al., 2020), such that combining tfMRI across tasks further boosted the prediction. We also showed that, when every modality was combined into the all-modality stacked model, tfMRI from several tasks together with resting FC drove the prediction. This confirms that tfMRI from certain tasks provided unique and important sources of information relevant to general cognitive abilities. The task stacked model was also superior than non-task stacked model even though non-task modalities (resting-state FC and sMRI) are much more commonly implemented in the literature on individual differences and cognition (Sui et al., 2020). Altogether, despite its superior performance, tfMRI has been ignored and downplayed in its importance for individual differences over non-task modalities, partly causing the unpopularity in using tfMRI as a predictive tool for cognition.

One of the main criticisms of tfMRI is its low reliability, compared to non-task modalities, such as sMRI (Elliott et al., 2020). Elliot and colleagues (2020) employed the same dataset (HCP) with ours and analysed ICC of tfMRI using a traditional univariate approach, i.e., separately at each prespecified region and task (Noble et al., 2021). They found poor ICC (<.4) of tfMRI signals across regions and tasks. Conversely, our predictive models drew tfMRI information across regions from the whole brain for each task and, for the task stacked model, further drew tfMRI information

across seven tasks. With this, we found ICC for the task stacked model at the excellent level (>.75) (Cicchetti & Sparrow, 1981), suggesting marked improvement over the classical univariate approach. Applying stacking to tfMRI gave us the best of both worlds: relatively high prediction and excellent reliability. This sharply contradicts the recommendations by Elliot and colleagues (2020) to rely on non-task modalities, such as sMRI which, although leading to high reliability (ICC>.75), provided very poor prediction (at $R^2 \leq .0512$).

Feature importance of the all-modality stacked model (Figure 2C) shows the important roles of three tfMRI tasks (working-memory, relational processing, and language) and resting-state FC. Given Elastic Net coefficients reflect unique contributions from each feature (Zou & Hastie, 2005), the three tfMRI tasks and resting-state FC appeared to provide non-overlapping variance in predicting the general cognitive abilities. This again reiterates the importance of including tfMRI in the predictive model of cognitive abilities.

Within the tfMRI stacked model (Figure 2C), certain tasks contributed highly to the model while other tasks did not provide much contribution (e.g., gambling and, to a lesser extent, social, motor and emotional). The three highly contributing tasks (working-memory, relational processing, and language) were tasks that are relevant to general cognitive abilities (Salthouse, 2004). Accordingly, the feature importance of the tfMRI stacked model seems to suggest domain specificality from each task (i.e., not just any tasks, but tasks related to the target of the model). Interestingly, brain activity of each of the three highly contributing tasks appeared to involve similar networks, dominated by the frontoparietal and default-mode networks, and to less extent, accounted for by the dorsal attention and cingulo-opercular, networks (Figure 4). When combing contribution of each region across all tasks (Figure 5), we then found a distributed network of brain regions in the frontoparietal network that drove the prediction of general cognitive abilities. In fact, our findings showed overlapping areas with those found in a meta-analysis of association studies (Santarnecchi et al., 2017) mainly at the frontoparietal network. This fits nicely with the parieto-frontal integration theory of intelligence (Jung & Haier, 2007), suggesting the important role of the frontoparietal network across cognitive contexts (i.e., tfMRI tasks).

Beyond providing a predictive and reliable method for capturing brain-cognition relationship, our work paves the way for developing a robust biomarker for cognitive abilities. According to the National Institute of Mental Health's Research Domain Criteria (RDoC), cognitive abilities are considered one of the six major transdiagnostic spectrums that cut across neuropsychiatric illnesses (Morris & Cuthbert, 2012). Following the RDoC, to understand neuropsychiatric illnesses, scientists need tools to examine the transdiagnostic spectrums (such as, cognitive abilities) at different units of analysis (such as, gene, brain to behaviours). Recent genome-wide association studies have brought out polygenic scores that quantify cognitive abilities at the genetic level (Allegrini et al., 2019). Having a cognitive brain-based biomarker as developed in this study can serve as a link between genetics (e.g., polygenic scores) and phenotypes (e.g., cognitive abilities). Examining this link can uncover the pathway between having genetic risks to developing neuropsychiatric symptoms (Gottesman & Gould, 2003). Next, neuroscientists can also apply the brain-based biomarker to examine interventions/behaviours that may alter

cognitive abilities. For instance, neuroscientists can implement the brain-based biomarker to investigate whether sleep (Taveras et al., 2017), exercise (Hötting & Röder, 2013) or extracurricular activities (Kirlic et al., 2021) improve brain processing involved in cognitive abilities, thereby deriving protective factors against many neuropsychiatric disorders. Accordingly, our biomarkers for cognitive abilities can play a vital role in the RDoC framework.

Our study is not without limitations. First, to demonstrate the benefits of the task over non-task modalities, we focused on the GLM contrasts of tfMRI that reflected changes in BOLD between experimental vs. control conditions for each task. While the GLM contrasts allowed us to focus on condition-specific variance of tfMRI, we may have missed condition-non-specific variance during the tfMRI scans that may also be related to general cognitive abilities. Recent studies (Elliott et al., 2019; Greene et al., 2018) have captured condition-non-specific variance using function-connectivity during tasks and found boosted prediction and reliability over those of resting-state FC. Accordingly, future studies may further blur the line between task vs non-task modalities by including condition-non-specific function-connectivity during both tasks and rest in the stacked models and examine their performance. Second, to ensure the interpretability of the machine-learning models (Molnar, 2019), we only applied Elastic Net (Zou & Hastie, 2005) that assumed additivity between brain features and linearity between brain features and the target. If interpretability is not the focus, future research may employ algorithms that allow interaction and non-linearity, such as support vector machine (Cortes & Vapnik, 1995) and random forest (Ho, 1995).

In conclusion, over the last decade, investigations of individual differences in the brain-cognition relationship have been dominated by non-task modalities (Sui et al., 2020). Here we show clearly that tfMRI, when used appropriately by 1) drawing information across regions from the whole brain and across tasks and by 2) combining with other MRI modalities, can provide unique and important sources of information about individual differences in cognitive abilities. This has led to an interpretable predictive model with unprecedentedly high prediction and excellent and reliability. Our research, thus, encourages the use of tfMRI in capturing individual differences in the brain-cognition relationship for general cognitive abilities and beyond.

References

Alin, A. (2010). Multicollinearity: Multicollinearity. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(3), 370–374. https://doi.org/10.1002/wics.84

Allegrini, A. G. … Plomin, R. (2019). Genomic prediction of cognitive traits in childhood and adolescence. *Molecular Psychiatry*, *24*(6), 819–827. https://doi.org/10.1038/s41380-019-0394-4

APS. (2020, June 3). *Scanning the Brain to Predict Behavior, a Daunting 'Task' for MRI*. Association for Psychological Science - APS. https://www.psychologicalscience.org/news/releases/scanning-the-brain-fmri.html

Barch, D. M. … 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

Basten, U. … Fiebach, C. J. (2015). Where smart brains are different: A quantitative meta-analysis of functional and structural brain imaging studies on intelligence. *Intelligence*, *51*, 10–27. https://doi.org/10.1016/j.intell.2015.04.009

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Deary, I. J. … Johnson, W. (2010). The neuroscience of human intelligence differences. *Nature Reviews Neuroscience*, *11*(3), 201–211. https://doi.org/10.1038/nrn2793

Destrieux, C. … Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15. https://doi.org/10.1016/j.neuroimage.2010.06.010

Dosenbach, N. U. F. … Schlaggar, B. L. (2010). Prediction of Individual Brain Maturity Using fMRI. *Science*, *329*(5997), 1358–1361. https://doi.org/10.1126/science.1194144

Dubois, J. … Adolphs, R. (2018). A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1756), 20170284. https://doi.org/10.1098/rstb.2017.0284

Efron, B., & Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. *The American Statistician*, *37*(1), 36–48. https://doi.org/10.2307/2685844

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman & Hall.

Elam, J. (2021, February 15). *HCP Data Release Updates: Known Issues and Planned fixes—Connectome Data Public—HCP Wiki*. https://wiki.humanconnectome.org/display/PublicData/HCP+Data+Release+Updates%3A+Known+Issues+and+Planned+fixes

Elliott, M. L. … Hariri, A. R. (2019). General functional connectivity: Shared features of resting-state and task fMRI drive reliable and heritable individual differences in functional brain networks. *NeuroImage*, *189*, 516–532. https://doi.org/10.1016/j.neuroimage.2019.01.068

Elliott, M. L. … Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, *31*(7), 792–806. https://doi.org/10.1177/0956797620916786

Engemann, D. A. … Gramfort, A. (2020). Combining magnetoencephalography with magnetic resonance imaging enhances learning of surrogate-biomarkers. *ELife*, *9*, e54055. https://doi.org/10.7554/eLife.54055

Finn, E. S. … Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*(11), 1664–1671. https://doi.org/10.1038/nn.4135

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Fischl, B. … Dale, A. M. (2002). Whole Brain Segmentation. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/S0896-6273(02)00569-X

Geerligs, L. … Lorist, M. M. (2015). A Brain-Wide Study of Age-Related Changes in Functional Connectivity. *Cerebral Cortex*, *25*(7), 1987–1999. https://doi.org/10.1093/cercor/bhu012

Glasser, M. F. … Van Essen, D. C. (2016). The Human Connectome Project's neuroimaging approach. *Nature Neuroscience*, *19*(9), 1175–1187. https://doi.org/10.1038/nn.4361

Glasser, M. F. … 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

Gottesman, I. I., & Gould, T. D. (2003). The Endophenotype Concept in Psychiatry: Etymology and Strategic Intentions. *American Journal of Psychiatry*, *160*(4), 636–645. https://doi.org/10.1176/appi.ajp.160.4.636

Graham, M. H. (2003). Confronting multicollinearity in ecological multiple regression. *Ecology*, *84*(11), 2809–2815. https://doi.org/10.1890/02-3114

Greene, A. S. … Constable, R. T. (2018). Task-induced brain state manipulation improves prediction of individual traits. *Nature Communications*, *9*(1), 2807. https://doi.org/10.1038/s41467-018-04920-3

Hawkins, D. M. … Mills, D. (2003). Assessing Model Fit by Cross-Validation. *Journal of Chemical Information and Computer Sciences*, *43*(2), 579–586. https://doi.org/10.1021/ci025626i

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282. https://doi.org/10.1109/ICDAR.1995.598994

Hötting, K., & Röder, B. (2013). Beneficial effects of physical exercise on neuroplasticity and cognition. *Neuroscience & Biobehavioral Reviews*, *37*(9, Part B), 2243–2257. https://doi.org/10.1016/j.neubiorev.2013.04.005

James, G. … Tibshirani, R. (2021). *An Introduction to Statistical Learning: With Applications in R*. Springer US. https://doi.org/10.1007/978-1-0716-1418-1

Ji, J. L. … Cole, M. W. (2019). Mapping the human brain's cortical-subcortical functional network organization. *NeuroImage*, *185*, 35–57. https://doi.org/10.1016/j.neuroimage.2018.10.006

Jung, R. E., & Haier, R. J. (2007). The Parieto-Frontal Integration Theory (P-FIT) of intelligence: Converging neuroimaging evidence. *Behavioral and Brain Sciences*, *30*(2), 135–154. https://doi.org/10.1017/S0140525X07001185

Kirlic, N. … Paulus, M. P. (2021). Extracurricular Activities, Screen Media Activity, and Sleep May Be Modifiable Factors Related to Children's Cognitive Functioning: Evidence From the ABCD Study®. *Child Development*, *n/a*(n/a). https://doi.org/10.1111/cdev.13578

Koul, A. … Cavallo, A. (2018). Cross-Validation Approaches for Replicability in Psychology. *Frontiers in Psychology*, *9*, 1117. https://doi.org/10.3389/fpsyg.2018.01117

Kragel, P. A. … Wager, T. D. (2021). Functional MRI Can Be Highly Reliable, but It Depends on What You Measure: A Commentary on Elliott et al. (2020). *Psychological Science*, *32*(4), 622–626. https://doi.org/10.1177/0956797621989730

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer New York. https://doi.org/10.1007/978-1-4614-6849-3

Marcus, D. … Van Essen, D. (2011). Informatics and Data Mining Tools and Strategies for the Human Connectome Project. *Frontiers in Neuroinformatics*, *5*, 4. https://doi.org/10.3389/fninf.2011.00004

McDaniel, M. A. (2005). Big-brained people are smarter: A meta-analysis of the relationship between in vivo brain volume and intelligence. *Intelligence*, *33*(4), 337–346. https://doi.org/10.1016/j.intell.2004.11.005

Mihalik, A. … Oxtoby, N. P. (2019). ABCD Neurocognitive Prediction Challenge 2019: Predicting Individual Fluid Intelligence Scores from Structural MRI Using Probabilistic Segmentation and Kernel Ridge Regression. In K. M. Pohl … M. G. Linguraru (Eds.), *Adolescent Brain Cognitive Development Neurocognitive Prediction* (pp. 133–142). Springer International Publishing. https://doi.org/10.1007/978-3-030-31901-4_16

Molnar, C. (2019). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/

Monti, M. (2011). Statistical Analysis of fMRI Time-Series: A Critical Review of the GLM Approach. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00028

Morris, S. E., & Cuthbert, B. N. (2012). Research Domain Criteria: Cognitive systems, neural circuits, and dimensions of behavior. *Dialogues in Clinical Neuroscience*, *14*(1), 29–37.

Noble, S. … Constable, R. T. (2021). A guide to the measurement and interpretation of fMRI test-retest reliability. *Current Opinion in Behavioral Sciences*, *40*, 27–32. https://doi.org/10.1016/j.cobeha.2020.12.012

P. Vatcheva, K., & Lee, M. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology: Open Access*, *06*(02). https://doi.org/10.4172/2161-1165.1000227

Pedregosa, F. … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Pietschnig, J. … Voracek, M. (2015). Meta-analysis of associations between human brain volume and intelligence differences: How strong are they and what do they mean? *Neuroscience & Biobehavioral Reviews*, *57*, 411–432. https://doi.org/10.1016/j.neubiorev.2015.09.017

Pohl, K. M. … Linguraru, M. G. (Eds.). (2019). *Adolescent Brain Cognitive Development Neurocognitive Prediction: First Challenge, ABCD-NP 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 13, 2019, Proceedings*. Springer International Publishing. https://doi.org/10.1007/978-3-030-31901-4

Poldrack, R. A. … Varoquaux, G. (2020). Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry*, *77*(5), 534–540. https://doi.org/10.1001/jamapsychiatry.2019.3671

Power, J. D. … Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018

Rasero, J. … Verstynen, T. (2021). Integrating across neuroimaging modalities boosts prediction accuracy of cognitive ability. *PLOS Computational Biology*, *17*(3), e1008347. https://doi.org/10.1371/journal.pcbi.1008347

Robinson, E. C. … Rueckert, D. (2018). Multimodal surface matching with higher-order smoothness constraints. *NeuroImage*, *167*, 453–465. https://doi.org/10.1016/j.neuroimage.2017.10.037

Ruigrok, A. N. V. … Suckling, J. (2014). A meta-analysis of sex differences in human brain structure. *Neuroscience & Biobehavioral Reviews*, *39*, 34–50. https://doi.org/10.1016/j.neubiorev.2013.12.004

Salthouse, T. A. (2004). Localizing age-related individual differences in a hierarchical structure. *Intelligence*, *32*(6), 541–561. https://doi.org/10.1016/j.intell.2004.07.003

Santarnecchi, E. … Pascual-Leone, A. (2017). Dissecting the parieto-frontal correlates of fluid intelligence: A comprehensive ALE meta-analysis study. *Intelligence*, *63*, 9–28. https://doi.org/10.1016/j.intell.2017.04.008

Satterthwaite, T. D. … Wolf, D. H. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *NeuroImage*, *64*, 240–256. https://doi.org/10.1016/j.neuroimage.2012.08.052

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Smith, S. M. … 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

Sripada, C. … Shedden, K. (2020). Toward a “treadmill test” for cognition: Improved prediction of general cognitive ability from the task activated brain. *Human Brain Mapping*, *41*(12), 3186–3197. https://doi.org/10.1002/hbm.25007

Sui, J. … Calhoun, V. (2020). Neuroimaging-based Individualized Prediction of Cognition and Behavior for Mental Disorders and Health: Methods and Promises. *Biological Psychiatry*, *88*(11), 818–828. https://doi.org/10.1016/j.biopsych.2020.02.016

Taveras, E. M. … Oken, E. (2017). Prospective Study of Insufficient Sleep and Neurobehavioral Functioning Among School-Age Children. *Academic Pediatrics*, *17*(6), 625–632. https://doi.org/10.1016/j.acap.2017.02.001

Trabzuni, D. … Ryten, M. (2013). Widespread sex differences in gene expression and splicing in the adult human brain. *Nature Communications*, *4*(1), 2771. https://doi.org/10.1038/ncomms3771

Van Essen, D. C. … 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

Weintraub, S. … Gershon, R. (2014). The Cognition Battery of the NIH Toolbox for Assessment of Neurological and Behavioral Function: Validation in an Adult Sample. *Journal of the International Neuropsychological Society*, *20*(6), 567–578. https://doi.org/10.1017/S1355617714000320

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

WU-Minn Consortium Human Connectome Project. (2018, April 10). *1200 Subjects Data Release—Connectome*. https://www.humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x