Cross-cohort replicability and generalizability of connectivity-based psychometric prediction patterns

Jianxiao Wu[a,b], Jingwei Li[a,b] Simon B. Eickhoff[a,b], Felix Hoffstaedter[a,b], Michael Hanke[a,b], B.T. Thomas Yeo[c,d,e,f,g] and Sarah Genon[a,b]

[a] Institute for Systems Neuroscience, Medical Faculty, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany

[b] Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Center Jülich, Jülich, Germany

[c] Department of Electrical and Computer Engineering, National University of Singapore, Singapore City, Singapore

[d] Centre for Sleep & Cognition & Centre for Translational Magnetic Resonance Research, Yong Loo Lin School of Medicine, Singapore City, Singapore

[e] N.1 Institute for Health & Institute for Digital Medicine, National University of Singapore, Singapore City, Singapore

[f] Integrative Sciences and Engineering Programme (ISEP), National University of Singapore, Singapore City, Singapore

[g] Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, Massachusetts

Correspondence and requests for materials should be addressed to Jianxiao Wu (j.wu@fz-juelich.de) or Sarah Genon (s.genon@fz-juelich.de).

Abstract

An increasing number of studies have investigated the relationships between inter-individual variability in brain regions' connectivity and behavioral phenotypes, making use of large population neuroimaging datasets. However, the replicability of brain-behavior associations identified by these approaches remains an open question. In this study, we examined the cross-dataset replicability of brain-behavior association patterns for fluid cognition and openness predictions using a previously developed region-wise approach, as well as using a standard whole-brain approach. Overall, we found moderate similarity in patterns for fluid cognition predictions across cohorts, especially in the Human Connectome Project Young Adult, Human Connectome Project Aging, and Enhanced Nathan Kline Institute Rockland Sample cohorts, but low similarity in patterns for openness predictions. In addition, we assessed the generalizability of prediction models in cross-dataset predictions, by training the model in one dataset and testing in another. Making use of the region-wise prediction approach, we showed that first, a moderate extent of generalizability could be achieved with fluid cognition prediction, and that, second, a set of common brain regions related to fluid cognition across cohorts could be identified. Nevertheless, the moderate replicability and generalizability could only be achieved in specific contexts. Thus, we argue that replicability and generalizability in connectivity-based prediction remain limited and deserve greater attention in future studies.

Keywords: behavior prediction, resting-state functional connectivity, generalizability, brain-behavior relationships, machine learning, fluid intelligence

1. Introduction

In recent years, the availability of population-based neuroimaging datasets (Nooner et al. 2012; Van Essen et al. 2013; Caspers et al. 2014; Holmes et al. 2015) has enabled many investigations into the relationships between functional connectivity (FC) and behavior. Resting-state functional connectivity (RSFC) has been employed in the predictions of various psychometric variables, ranging from cognitive measures to personality traits (Finn et al. 2015; Noble et al. 2017; Beaty et al. 2018; Dubois et al. 2018a; Dubois et al. 2018b; Jiang et al. 2018; Maglanoc et al. 2019; Avery et al. 2020; He et al. 2020; Jiang et al. 2020). In other words, by training a model to learn the relationships between RSFC and psychometric variables, the model can infer, to some extent, the values of these psychometric variables in a new sample, using the new sample's RSFC. These approaches can be overall referred to connectivity-based psychometric prediction (CBPP) approaches. Generally, when using these approaches, a strong focus is put on achieving good, or at least what could be deemed decent, prediction performance. To evaluate prediction performance, typically, the data from one dataset is repeatedly and randomly partitioned into training and test sets, where the model's performance is thus determined by its average prediction accuracies on the test set data. This is typically referred to as a (within cohort) cross-validation approach. In rare cases, prediction performance may also be computed using a held-out test set, where the accuracy measured is still within the same cohort (Maglanoc et al. 2019; Avery et al. 2020), or out-of-sample test data, where the accuracy is measured in a fully new cohort (Beaty et al. 2018; He et al. 2020; Jiang et al. 2020). Usually capitalizing on within-cohort cross-validation, many studies have further investigated the technical factors affecting the model performance and/or the neurobiological insights provided by predictive models (Li et al. 2019; Pervaiz et al. 2020; Wu et al. 2021; Kong et al. 2021).

CBPP approaches typically require relatively large sample size which, however, is a very scarce resource in the field. Based on recent surveys (Sui et al. 2020; Yeung et al. 2022), 38 CBPP studies with relatively large sample size can be identified ($N \geq 200$; see Supplemental Materials for a list of the studies), among which 71% made use of the Human Connectome Young Adult (HCP-YA; Van Essen et al. 2013) dataset, while 47% used only the HCP-YA dataset. This is expected, as the HCP-YA dataset was one of the first large population-based dataset which offers high-quality resting-state scans and extensive psychometric characterization. Nevertheless, this brings forth the issue of generalizability, as the HCP-YA data were distinct from other datasets in multiple aspects. First, the subjects involved had a specific age range of 22 to 35. Second, most subjects involved were family members. Finally, the demographic characteristics, psychometric tools, scanning protocols and image processing of the HCP-YA were also different from most other datasets. The last point could be generalized to most existing datasets. As many datasets were based on different initiatives to fulfill different research questions, they would be different from each other in terms of sample characteristics, psychometric tools, scanning protocols and image processing protocols. Consequently, results and insights obtained by a research study performed using a single dataset would be inherently affected and limited by the idiosyncrasies of that specific dataset. As more and more population-based datasets become available, it is necessary to investigate how replicable the brain-behavior association patterns identified could be and to which extent the prediction model learnt based on one cohort is generalizable to other cohorts.

The motivation of this study is two-fold. First, it is necessary to investigate the replicability of CBPP results, both in terms of prediction performance and the derived brain-behavior association patterns. Importantly, the brain-behavior association patterns derived from a prediction model (or brain prediction patterns, for short) can help to interpret the prediction model from a neurobiological perspective. Hence, the patterns' replicability across cohorts could limit the usefulness of the model. Second, the generalizability of prediction models is also a crucial aspect of CBPP model validity. In order for the CBPP models to achieve practical utility, they must be generalizable to unseen data. In particular, the brain prediction patterns and the interpretation of the model should remain consistent in unseen data.

To investigate the cross-cohort replicability of brain prediction patterns, we first made use of a previously proposed region-wise CBPP framework (Wu et al. 2021) for disentangling brain-behavior relationships, by building a prediction model for each brain region (or parcel) separately. Under this framework, an accuracy distribution map could be constructed for a psychometric variable, illustrating the contribution of each brain region's connectivity profile to the prediction of this psychometric variable. While such a local approach obviously simplifies the complexity of brain function (Horien et al. 2019), it allows us to identify relevant brain regional connectivity patterns for easier interpretation of the prediction models, as well as for future applications based on small sample sizes that would hence require significant features' reduction. By making use of the accuracy distribution map for a specific psychometric variable as a representation for the predictive brain pattern, we can assess the replicability of brain prediction patterns (i.e., brain-behavior association patterns derived from a prediction model) for similar behavioral measurements in different cohorts. In line with the trend in the field, we also implemented whole-brain CBPP, where all region-to-region connectivity values were used in one prediction model. As the regression weights from whole-brain CBPP models are not directly interpretably, the Haufe transformation (Haufe et al. 2014) was applied to transform these regression weights into values which can be associated with the predictive power of the functional connectivity edges. In this way, we could use the Haufe transformed patterns of different cohorts as prediction patterns of these whole-brain CBPP models. Both region-wise and whole-brain prediction patterns in this case will be referred to as 'within-dataset prediction patterns' for assessing the 'replicability of brain prediction patterns', as the patterns would be derived from, and thus specific to, a single dataset.

In addition, since prediction models trained on a specific dataset would be largely influenced by the idiosyncrasies of the dataset, its generalizability to a new dataset cannot be simply assumed. By training prediction models in one dataset and testing on other datasets, we could obtain the 'cross-dataset prediction patterns' for assessing the 'generalizability of prediction models'. If the within-dataset prediction pattern of a psychometric variable was similar to cross-dataset prediction patterns trained on the same data, we may infer that prediction models trained on this dataset could potentially be generalized to other datasets for this psychometric variable.

Intelligence and personality are core domains in differential psychology and hence in the study of interindividual variability in humans (Humphreys and Revelle 1984; Deary et al. 2011). Not surprisingly, among the most investigated psychometric variables in CBPP studies were fluid intelligence and personality traits. Using HCP-YA data and linear predictive models, the fluid intelligence and openness scores were both commonly investigated and generally among the best predicted psychometric variables; in particular, among the Big Five personality traits, only openness was reported to be predicted with statistical significance (Dubois et al. 2018a). In that context, the reported accuracies (Pearson correlation between predicted and observed scores) were in the range of 0.20 to 0.25 (Smith et al. 2016; Noble et al. 2017; Dubois et al. 2018a; Dubois et al. 2018b; Li et al. 2019; Pervaiz et al. 2020; Wu et al. 2021; Kong et al. 2021). Furthermore, the measure of fluid intelligence, as well as fluid cognition, could be related to various intelligence quotient (IQ) measures in other datasets and the openness trait from the Neuroticism/Extroversion/Openness Five Factor Inventory (NEO-FFI) inventory is a common measure in many datasets. Therefore, we selected these two measures as the best candidate psychometric variables for our replicability and generalizability investigations.

Accordingly, we selected four healthy adult datasets in which both fluid cognition and openness measures were available: the HCP-YA, the Human Connectome Project Aging (HCP-A), the Enhanced Nathan Kline Institute Rockland Sample (eNKI-RS) and the Brain Genomics Superstruct Project (GSP) cohorts, providing opportunities to examine cohort differences in terms of sample characteristics, image acquisition and psychometric test implementation (see a summary in table 1 and age distribution plots in figure S1). We first assessed the replicability of brain prediction patterns across these cohorts by producing the region-wise and whole-brain (within-dataset) brain spatial prediction patterns for each psychometric variable in each cohort. Then, we assessed the generalizability of prediction models for fluid cognition prediction by comparing the within-dataset and cross-dataset prediction patterns. Based on these prediction patterns, we demonstrated that a common set of brain regions related to fluid intelligence could be identified. Finally, we will discuss the implication of our results for the field and future studies.

Table 1. Summary of datasets used

	HCP-YA (Van Essen et al. 2013)	**HCP-A (Bookheimer et al. 2019)**	**eNKI-RS (Nooner et al. 2012)**	**GSP (Holmes et al. 2015)**
Number of subjects (N)	931	601 (fluid cognition) 715 (openness)	970 (fluid cognition) 820 (openness)	867
Age	28.81±3.70	58.11±13.88	39.70±23.15	21.59±2.84
Gender	497 female, 434 male	329 female, 242 male	575 female, 389 male	500 female, 367 male
Length of resting-state runs	14.4 min / 1200 frames	6 min / 488 frames	10 min / 900 frames	6 min / 120 frames
Repetition time (TR)	720 ms	720 ms	645 ms	3000 ms
Resolution of resting-state scans	2mm isotropic	2mm isotropic	3mm isotropic	3mm isotropic
Fluid cognition measures	1. fluid cognition composite score (*CogFluidComp_AgeAdj*) 2. fluid intelligence (*PMAT24_A_CR*)	fluid cognition composite score (*nih_fluidcogcomp_ageadjusted*)	Wechsler Abbreviated Scale of Intelligence (WASI-II; *FSIQ – 4 Composite Score*)	Shipley IQ (*EstIQ_Matrix_Int_Bin*)

Openness measures	NEO-FFI openness (*NEOFAC_O*)	NEO-FFI openness (*neo2_score_op*)	NEO-FFI openness (*O T-Score*)	NEO-FFI openness (*NEO_O*)
Confounding variables	Sex (*Gender*), age (*Age_in_Yrs*), age^2, sex*age, sex*age^2, handedness (*Handedness*), brain size (*FS_BrainSeg_Vol*), intracranial volume (*FS_IntraCranial_Vol*), and acquisition quarter (*Acquisition*)	Sex (*sex*), age (*interview_age*), age^2, sex*age, sex*age^2, handedness (*hcp_handedness_score*), brain size (*BrainSegVol*), intracranial volume (*EstimatedTotalIntraCranialVol*)	Sex (‘*What is your sex?*’), age (*Calculated Age*), age^2, sex*age, sex*age^2, handedness (*LATERALITY INDEX*), brain size (*BrainSegVol),* intracranial volume (*EstimatedTotalIntraCranialVol*)	Sex (*Sex*), age (*Age_Bin*), age^2, sex*age, sex*age^2, handedness (*Hand*), brain size (*BrainSegVol*), and intracranial volume (*ICV*)

2. Materials and Methods

2.1. Data and Preprocessing

The HCP-YA S1200 Release (Van Essen et al. 2013) includes phenotype and imaging data from over 1200 healthy young adults (aged 22 to 37), from families with twins and non-twin siblings. Imaging data were acquired using a customized Siemens 3T Skyra. Each subject visited in two consecutive days, during each of which two resting-state runs were acquired using different phase-encodings, left-right and right-left. Each run is 1200 frames (14.4 min) in length, with a repetition time (TR) of 720 ms. All resting-state functional Magnetic Resonance Imaging (fMRI) data were 2mm isotropic. We only considered subjects with all four runs completed (N = 931). All raw resting-state data were preprocessed by the HCP Minimal Processing Pipelines (Glasser et al. 2013), followed by ICA-FIX denoising (Smith et al. 2013; Griffanti et al. 2014; Salimi-Khorshidi et al. 2014).

The HCP-A Release 2.0 (Harms et al. 2018; Bookheimer et al. 2019) includes phenotype and imaging data from 725 healthy adults (ages 36 to 100+), as an extension for the HCP-YA cohort. Imaging data were acquired using a Siemens 3T Prisma. Similar to the HCP-YA cohort, two resting-state sessions each with two runs were acquired for each subject, using anterior-posterior and posterior-anterior phase encoding respectively. Each run is 488 frames in length, with a TR of 720 ms. All resting-state fMRI data were 2mm isotropic. We only considered subjects with all four runs completed (N=720). All raw resting-state data were preprocessed by the HCP Minimal Processing Pipelines (Glasser et al. 2013), followed by ICA-FIX denoising (Smith et al. 2013; Griffanti et al. 2014; Salimi-Khorshidi et al. 2014).

The eNKI-RS (Nooner et al. 2012) includes phenotype and imaging data from a lifespan sample of over 1000 participants (aged 6 to 85). Imaging data were acquired using a Siemens 3T Tim Trio. We made use of the fast repetition time (TR = 645 ms) resting-state scans each lasting 10 minutes (actual number of time point = 900) and with a resolution of 3mm isotropic (N=1309), which was anticipated to improve comparison with the HCP-YA data (Nooner et al. 2012). We processed all raw resting-state data with fMRIPrep (Esteban et al. 2019) with default parameters and additionally ICA-AROMA denoising (Pruim et al. 2015a; Pruim et al. 2015b); the details of the pipeline implementation can be found in Supplemental Materials.

The GSP initial data release (Holmes et al. 2015) includes phenotype and imaging data from young adults (aged 18 to 35; N=867). Imaging data were acquired using matched Siemens 3T Tim Trio scanners at two sites. The resting-state scans were 3mm isotropic, each with 120 frames and a TR of 3000 ms. These resting-state data were preprocessed with an in-house pipeline, which includes fieldmap correction, motion correction, slice-time correction, spatial normalization to the MNI152 standard space and ICA-FIX denoising.

For all four datasets, resting-state data in the MNI152 space were used. We applied nuisance regression to control for white matter signals, cerebrospinal fluid signals and their derivatives, as well as 24 motion parameters. As the HCP-YA and HCP-A datasets offer already preprocessed data, while the eNKI-RS and GSP data were preprocessed by us, we note that the data processing across the four cohorts may not be considered comparable. However, this is in line with our aim to highlight potential issues of generalizability and replicability in practical scenarios, where data processing procedures can hardly be standardized. In particular, we aim to avoid situations in which a unique preprocessing pipeline would be selected and could be optimal for one dataset, but not for the others. Accordingly, for the eNKI-RS and GSP datasets, preprocessing was done in the way that was deemed optimal for the respective dataset so that data quality should remain comparable across cohorts despite the difference in preprocessing pipelines.

In order to make sure that our results are not limited by the specificity of a single atlas, or a single granularity, we defined brain regions using parcels from two different atlases. We here selected the AICHA atlas (Figure 1A; Joliot et al. 2015) as an atlas that is independent from the datasets used in this study and derived in a volumetric space. Additionally, we used a combination of the Schaefer cortical atlas and the 3T version of Melbourne subcortex atlas (Figure 1B-E; Schaefer et al. 2018; Tian et al. 2020) as extensively evaluated and used atlases. The Schaefer cortical atlas and the Melbourne subcortex atlas were derived based on the GSP cohort and the HCP-YA cohort respectively, but were nonetheless useful in offering different levels of granularity. The AICHA atlas contains 384 parcels encompassing both cortical and subcortical regions. The Schaefer atlas and the Melbourne atlas were combined by the level of granularity. In other words, the 100-parcel Schaefer atlas was combined with the 16-parcel Melbourne atlas, the 200-parcel Schaefer atlas with the 32-parcel Melbourne atlas, the 300-parcel Schaefer atlas with the 50-parcel Melbourne atlas, and the 400-parcel Schaefer atlas with the 54-parcel Melbourne atlas. This hence allows us to examine the brain prediction patterns across 4 levels of granularity, with 116 parcels, 232 parcels, 350 parcels and 454 parcels respectively. Within each parcel, the mean time series across all voxels inside the parcel was computed. The FC profile for each parcel was then obtained by computing the Pearson correlation between the mean time series between that parcel and every other parcel. For the HCP-YA and HCP-A subjects, the average connectivity values across all four runs were used.

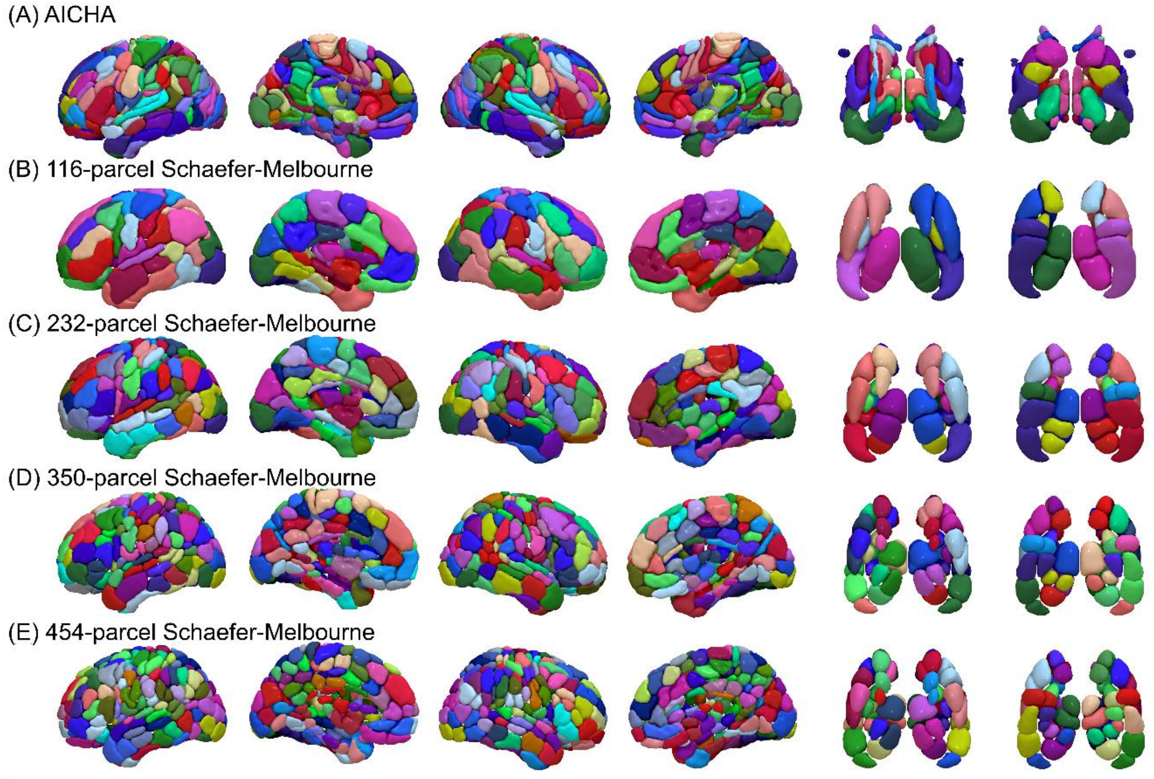
(A) AICHA
(B) 116-parcel Schaefer-Melbourne
(C) 232-parcel Schaefer-Melbourne
(D) 350-parcel Schaefer-Melbourne
(E) 454-parcel Schaefer-Melbourne

Figure 1. (A) AICHA atlas. (B) 116-parcel Schaefer-Melbourne combined atlas. (C) 232-parcel Schaefer-Melbourne combined atlas. (D) 350-parcel Schaefer-Melbourne combined atlas. (D) 454-parcel Schaefer-Melbourne combined atlas. From left to right, the columns reflect: the lateral view of left hemisphere, the medial view of left hemisphere, the lateral view of right hemisphere, the medial view of right hemisphere, the superior view of subcortical regions, and the inferior view of subcortical regions.

2.2. Psychometric Variables

Three psychometric variables were considered for the HCP-YA dataset, namely the fluid intelligence measure, the fluid cognition composite score and the NEO-FFI openness measure. Fluid intelligence was measured using Form A of an abbreviated version of the Raven's Progressive Matrices (Bilker et al. 2012). The Fluid cognition composite score is an age-adjusted summary score, comprising the Dimensional Change Card Sort Test for cognitive flexibility, the Flanker Inhibitory Control and Attention Test, the Picture Sequence Memory Test for non-verbal episodic memory, the List Sorting Working Memory Test, and the Pattern Comparison Processing Speed Test. Lastly, the openness score was based on the 12 specific items from the revised 60-item version of the NEO-FFI (McCrae and Costa 2004).

For the HCP-A dataset, the fluid cognition composite score and the NEO-FFI openness measure were selected, both of which were estimated in the same way as their counterpart in the HCP-YA dataset. As the two measures were only available in a subset of participants, we implemented predictions for them separately using data from all participants with the respective measure (N=623 for fluid cognition composite score, N=715 for openness). For the fluid cognition prediction, we further excluded subjects whose score values were 999 (final N=601).

For the eNKI-RS dataset, two psychometric variables were considered, the Wechsler Abbreviated Scale of Intelligence (WASI-II) measure and the NEO-FFI openness measure. The WASI is a general intelligence test designed to assess specific and overall cognitive capabilities, consisting of four subtests: vocabulary, block design, similarities, and matrix reasoning. Similar to HCP-YA, the openness measure was based on the revised 60-item NEO-FFI (McCrae and Costa 2004). As the NEO-FFI test was done only in a subset of participants who went through the WASI-II test, we implemented predictions for the two psychometric variables separately using data from all participants with the respective measure (N=970 for WASI-II intelligence; N=820 for openness).

For the GSP dataset, two psychometric variables were considered, the Shipley IQ measure and the NEO-FFI openness measure. The IQ measure was estimated from Shipley-Hartford age-corrected t-scores, which showed strong relation to WASI derived IQ in a subset of subjects

(Holmes et al. 2015). The openness measure was based on the 60-item NEO-FFI (Costa and McCrae 1992).

In order to focus on brain-behavior relationships, the influence of demographic factors, such as age and gender, need to be controlled. In our previous work, we also showed that head size estimates could affect the psychometric prediction profile of individual parcels (Wu et al. 2021). In line with our previous work and the HCP MegaTrawl analysis (Smith et al. 2016), we considered a set of eight confounding variables for all samples: sex, age, age^2, sex*age, sex*age^2, handedness, brain size, and intracranial volume. For the HCP-YA sample, we additionally included acquisition quarter as a confounding variable. As the multiband reconstruction algorithm used was different in earlier quarters, controlling for the acquisition quarter helps to mitigate the effect of this change in data collection protocol (Dubois et al. 2018a).

2.3. Whole-brain and Region-wise Connectivity-based Psychometric Prediction

In line with the global trend in the field, the whole-brain CBPP model uses all parcel-to-parcel connectivity values as input features for a linear machine learning algorithm to predict the target psychometric variable. In contrast, the region-wise model uses each parcel's FC profile separately. For instance, for a 300-parcel atlas, a subject's input feature for the whole-brain model is the upper triangular part of the FC matrix with dimensions of 300×300, excluding the diagonal values, resulting in a final number of features of 44850. For the region-wise model, a subject's input feature is the FC between the chosen region and all other regions, hence resulting in 299 features for a 300-parcel atlas.

For each model, the whole-brain FC matrices or parcel-wise FC profiles (the FC features) for all training subjects, as well as their psychometric variable values, are provided to a machine learning algorithm. A linear relationship is estimated between the FC features and the psychometric variables by the algorithm. This linear relationship can then be used to infer psychometric variable values in new subjects, using their FC features. Finally, the inferred (or predicted) values are compared with the actual observed values in these new subjects, to evaluate the prediction performance.

The psychometric prediction was carried out with 100 repeats of 10-fold cross-validation. For each fold, subjects inside that fold are considered the test set, while the remaining nine folds are considered the training set. The prediction model is learnt using the training set data and evaluated using the test set data. During each repeat, every fold is used as the test set fold once; such process is then repeated 100 times. For each test set fold, confounding variables were first regressed out from the remaining nine training folds. The same regression coefficients were used to remove confounding effects from the test set fold. To account for the family structure within the HCP-YA cohorts, family members were always kept within the same fold for the HCP-YA cohort.

We applied support vector regression (SVR; Boser et al. 1992; Cortes and Vapnik 1995) using Matlab's fitrlinear function. The hyperparameter determining the error tolerance during model fitting, epsilon, was set to the default value which is dependent on data variance. Specifically, epsilon was set to $\frac{IQR(Y)}{13.49}$, where the numerator is the interquartile range of the target variable in the training set. In order to accommodate the large feature space, especially in the whole-brain approach, ridge penalty is included in the objective function for regularization. The regularization strength, lambda, was also set to the default value of $\frac{1}{n}$, the inverse of the training sample size. For comparison's sake, we also applied elastic net (EN; Zou and Hastie 2005) using the glmnet package for Matlab (Qian et al. 2013). For each fold, the hyperparameter alpha, which determines the compromise between ridge and lasso, was first fixed, while the hyperparameter lambda, which determines the degree of regularization, was tuned with 10-fold inner-loop cross-validation using 8 of the training folds. The model with the best lambda was then validated on the last training fold, to determine the best value for alpha.

Prediction accuracy was assessed by computing the Pearson correlation between predicted and observed psychometric values, averaged across all test set folds and all repeats. For whole-brain approaches, this means that one accuracy value was computed for each atlas and each psychometric variable in each cohort. For region-wise approaches, one accuracy value was computed for each parcel in each atlas, and for each psychometric variable in each cohort.

2.4. Replicability of Brain Prediction Patterns

Interpreting the prediction patterns for the region-wise CBPP models is straightforward. For each psychometric variable, we could visually or numerically compare the prediction accuracy distribution across parcels (Wu et al. 2021). The prediction accuracy achieved by each parcel's region-wise model can be related to the contribution of that parcel's connectivity with other parcels in the corresponding behavioral function. For each psychometric variable and each atlas, permutation test was performed by shuffling the scores of the psychometric variables in 1000 repeats of 10-fold cross-validation (100 repeats for EN due to higher computational cost). Multiple comparisons across parcels were corrected using false discovery rate (FDR; Benjamini and Hochberg 1995) of $q < 0.05$.

The whole-brain CBPP models cannot be interpreted directly using the regression weights assigned to each connectivity edge. Since SVR (and most other machine learning models often employed in prediction studies) is a backward model, interpreting the weights can be drastically misleading; large weights can be assigned to features unrelated to the brain process of interest (Haufe et al. 2014). To solve this issue, the Haufe transformation (Haufe et al. 2014) can be used to turn these weights into weights of a corresponding forward model. These transformed weight values can then be related to the FC edge's predictive power, where a larger absolute value would suggest that the FC edge is more involved in the prediction of the target psychometric variable. For each psychometric variable and each atlas, permutation test was performed by shuffling the scores of the psychometric variables during the transformation, in 1000 repeats of 10-fold cross-validation. Multiple comparisons across parcels were corrected using false discovery rate (FDR; Benjamini and Hochberg 1995) of $q < 0.05$. Each set of transformed weight values were then z-score normalized to have zero mean and unit variance.

To numerically assess the cross-cohort replicability of brain prediction patterns, the Pearson correlation coefficient was computed between patterns derived for different psychometric variables for each atlas option. For the region-wise models, this means computing the correlation between the two arrays of parcel-specific prediction accuracies. For the whole-brain models, this means computing the correlation between the two sets of Haufe transformed weight values. Between two psychometric variables, the replicability is indicated by the average Pearson correlation value across different atlases.

2.5. Generalizability of Prediction Models

In practice, replicability of prediction patterns would not suffice to validate a prediction model, as the model's generalizability to novel data must be tested too. For any population level inference, the inference needs to be generalizable to other populations. Similarly, for clinical applications using machine learning, models trained on existing data need to be generalizable to new patients. Focusing on the prediction patterns derived from region-wise prediction models, we examined such cross-dataset generalizability by training region-wise CBPP models based on each single dataset and testing the models on the other datasets.

For the assessment of cross-dataset generalizability of prediction models, we focused on models where relatively higher prediction accuracies and consistent prediction patterns were observed across cohorts, i.e., the fluid cognition prediction for the HCP-YA, HCP-A, and eNKI-RS cohorts. Specifically, region-wise CBPP models were trained on the FC and psychometric data from one dataset and tested on the FC and psychometric data from another. We refer to this as 'cross-dataset predictions'. For each parcel in each atlas, one model was trained and evaluated separately. Consequently, we obtained one accuracy value for each parcel in each atlas for each test set (this is different from the replicability case where accuracy values were averaged across test sets in cross-validation).

We then visualize the prediction patterns as prediction accuracy distribution maps, thus comparing them to the prediction patterns of models trained and tested in the same dataset, which are referred to as 'within-dataset predictions'. For numerical comparison, we computed the Pearson correlation coefficients between cross-dataset and within-dataset prediction patterns. The generalizability is indicated by the correlation value between patterns derived from models trained and tested in one dataset, and patterns derived from models trained in the same dataset but tested in a different dataset (or models trained in a different dataset but tested in the same dataset). In other words, for each pair of patterns, we computed the correlation between the two arrays of parcel-specific prediction accuracies.

2.6. Data and Code Availability

All data were managed via version-controlled DataLad datasets (Halchenko, et al., 2021) that are either publicly available, or were provided by an institutional data management system when public sharing was prevented by the terms of the respective data usage agreements.

The HCP-YA imaging data were accessed via the public DataLad dataset provided at https://github.com/datalad-datasets/human-connectome-project-openaccess (2e2a8a70-3eaa-11ea-a9a5-b4969157768c@a33e528) which interfaces the HCP Open Access dataset (https://registry.opendata.aws/hcp-openaccess) on AWS S3. The unrestricted and restricted phenotype data were downloaded from the ConnectomeDB (https://db.humanconnectome.org) after accepting the Open Access Data User Terms and Restricted Access Data Use Terms respectively.

The HCP-A imaging and phenotype data were downloaded from the NIMH Data Archive (NDA; https://nda.nih.gov), after applying for the Data Use Certification. The associated study ID is 1376 (http://dx.doi.org/10.15154/1524254).

The eNKI-RS imaging data were downloaded from the COINS data exchange (https://coins.trendscenter.org). The phenotype data were downloaded after accepting the Data Usage Agreement.

The GSP imaging and phenotype data were downloaded from the LONI Imaging Data Archive (https://ida.loni.usc.edu) after accepting the GSP Data Use Terms and GSP Restricted Data Use Terms.

All codes are publicly available at https://github.com/inm7/cbpp.

3. Results

Before investigating the brain patterns supporting the prediction of our selected psychometric variables, we first examined the global prediction performance for these variables. We then examined the replicability of brain prediction patterns based on the within-dataset brain patterns related to the prediction of the two behavioral measures (intelligence and openness). This includes patterns obtained when using the region-wise CBPP model, as well as when using a whole-brain connectivity matrix with post-hoc evaluation of functional connectivity edges contribution. The consistency of these brain patterns across datasets was addressed by computing correlation. Finally, in a generalizability perspective, we investigated the similarity of the cross-dataset brain patterns derived from a region-wise CBPP when the model is trained on one dataset and tested in another dataset, in comparison to the within-dataset brain patterns trained or tested using the same dataset. We would here expect that high similarity across different train-test datasets pairs reflect a generalizable involvement or contribution of a set of regions for the prediction of a given behavioral aspect.

3.1. Prediction Performance for Fluid Intelligence and Openness

We first examine the whole-brain model performance and the best region-wise model performance for each psychometric variable (figure 2). Each box shows the distribution of prediction accuracies across the 5 different atlases used. For each atlas, the best region-wise model represents the highest prediction accuracy achieved by region-wise models using that atlas. Numerically, the best whole-brain model prediction accuracies were $r = 0.21, 0.24, 0.39, 0.31, 0.13$ for HCP-YA fluid intelligence, HCP-YA fluid cognition, HCP-A fluid cognition, eNKI-RS WASI-II intelligence and GSP Shipley IQ, respectively, while the best region-wise accuracies were $r = 0.20, 0.25, 0.30, 0.25, 0.14$. For the openness measure, the best whole-brain accuracies were $r = 0.18, 0.28, 0.04, 0.20$ for HCP-YA, HCP-A, eNKI-

RS and GSP respectively, while the best region-wise accuracies were $r = 0.20, 0.24, 0.15, 0.18$. With the exception of the GSP Shipley IQ and eNKI-RS openness, most of the psychometric variables were predicted with accuracies similar to existing studies (Smith et al. 2016; Noble et al. 2017; Dubois et al. 2018a; Dubois et al. 2018b; Li et al. 2019; Pervaiz et al. 2020; Wu et al. 2021; Kong et al. 2021).

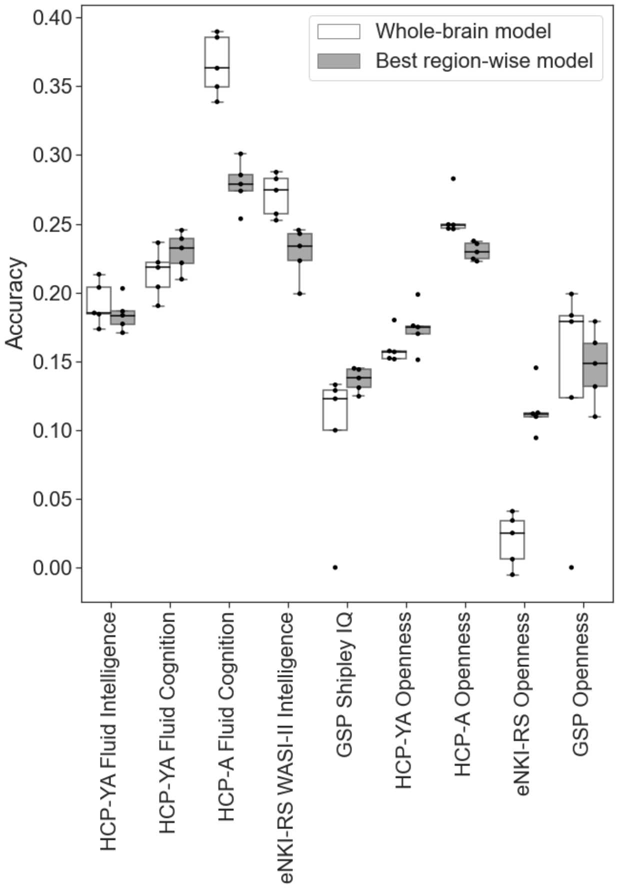

Figure 2. Prediction accuracies (Pearson's correlation between predicted and observed values) for each psychometric variable using the whole-brain model (white boxes) and best region-wise model (black boxes) with different atlases. For each psychometric variable and each atlas, only the region-wise model with the highest prediction accuracy (i.e., the best region-wise model) was included. Black bars inside the boxes represent the median accuracy value across different parcellations.

3.2. Region-wise CBPP Patterns

First, we present the prediction patterns for the region-wise CBPP models, shown as prediction accuracy distribution maps. Figures 3 and 4 shows the prediction accuracy distribution maps for the fluid intelligence and openness measures respectively. When analyzing these prediction patterns, we focus on the relative comparisons of prediction accuracies between different parcels, for each psychometric variable separately.

For the HCP-YA fluid intelligence measure, different prediction patterns were observed when different parcellations were used. For predictions using the AICHA atlas, better performing parcels were identified in the left precuneus and right anterior cingulate cortex. For predictions using the Schaefer-Melbourne atlas, across granularities, better performing parcels were mostly in the left supramarginal gyrus, right temporal cortex, and left and right precuneus. For the HCP-YA fluid cognition measure, better performing parcels were generally found in the left and right occipital lobe, left and right anterior insula, left and right anterior and posterior cingulate cortex, left and right supramarginal gyrus, and right prefrontal cortex. For the HCP-A fluid cognition measure, many parcels achieved predictions accuracies of $r > 0.2$, spanning across the prefrontal cortex, cingulate cortex, lateral temporal lobe, occipital lobe, supramarginal gyrus, precuneus, and anterior insula. For the eNKI-RS WASI-II intelligence measure, the prediction patterns were rather similar to the patterns for the HCP-YA fluid cognition measure, with better performing parcels additionally identified in the right temporal lobe and left hippocampus body. For the GSP Shipley IQ measure, prediction accuracies were generally low across the brain. Overall, by visual inspection, some similarities can be observed between the HCP-YA fluid cognition, HCP-A fluid cognition, and eNKI WASI-II intelligence measure.

To validate the robustness of the prediction patterns for fluid cognition, we derived the prediction patterns using EN as well (Figure S2). While some differences could be observed between patterns using SVR and EN, we could identify the same sets of better performing parcels in both patterns for each fluid cognition measure. The similarity between HCP-YA fluid cognition, HCP-A fluid cognition, and eNKI WASI-II intelligence measure remains.

Overall, we found the region-wise prediction patterns consistent across the two regression algorithms used, SVR and EN.

To validate the specificity of the prediction patterns for fluid cognition, we also derived the prediction accuracy distribution maps for the crystallized cognition composite score in the HCP-YA and HCP-A cohorts (Figure S3). For the HCP-YA fluid cognition and crystallized cognition, better performing parcels were found in the left and right anterior and posterior cingulate cortex, left and right supramarginal gyrus, and right prefrontal cortex, in both cases. However, some unique better performing parcels were also identified for fluid cognition and crystallized cognition. For the HCP-A cohort, many parcels achieved prediction accuracies of $r > 0.2$ for both fluid cognition and crystallized cognition.

Regarding the openness measures, first we note that the prediction accuracies were generally low for the eNKI-RS data. For the HCP-YA and GSP cohorts, the better performing parcels are few and sparse. Accordingly, very little similarity could be observed across datasets. For the HCP-YA openness measure, the relatively better performing parcels were found in the left and right prefrontal cortex, left and right anterior and posterior cingulate cortex, and left and right insula. For the HCP-A openness measure, the relatively better performing parcels were found in the left and right prefrontal cortex, right anterior cingulate cortex, left and right posterior cingulate cortex, left and right posterior temporal lobe, left and right temporal pole, and left and right precuneus. For GSP openness measure, the relatively better performing parcels were found in the left and right parietooccipital sulcus when using the AICHA atlas, and in the right ventral posterior temporal lobe when using the Schaefer-Melbourne atlas. Overall, little similarity was found across openness prediction patterns from different cohorts.

To quantitatively assess the similarity between the prediction patterns, we computed the correlation between the prediction accuracy distributions of each pair of psychometric variables, averaged across different parcellations (Figure 5). The correlation values were low to moderate ($r = 0 - 0.40$). Relatively, the higher consistencies were found between the two HCP-YA fluid intelligence measures, between HCP-YA and HCP-A fluid cognition measures, between these two (HCP-YA and HCP-A fluid cognition measures) and eNKI-RS WASI-II intelligence, as well as between HCP-A fluid cognition and openness measures. Overall, the highest similarity was found between HCP-YA and HCP-A fluid cognition measures, which are also the two most similar fluid cognition measures in terms of psychometric tools used.

To summarize, using a region-wise CBPP approaches within each dataset, a certain degree of replicability of the brain pattern was mainly observed for intelligence measures, while openness can hardly be predicted in some datasets and show poor replicability of brain patterns across datasets.

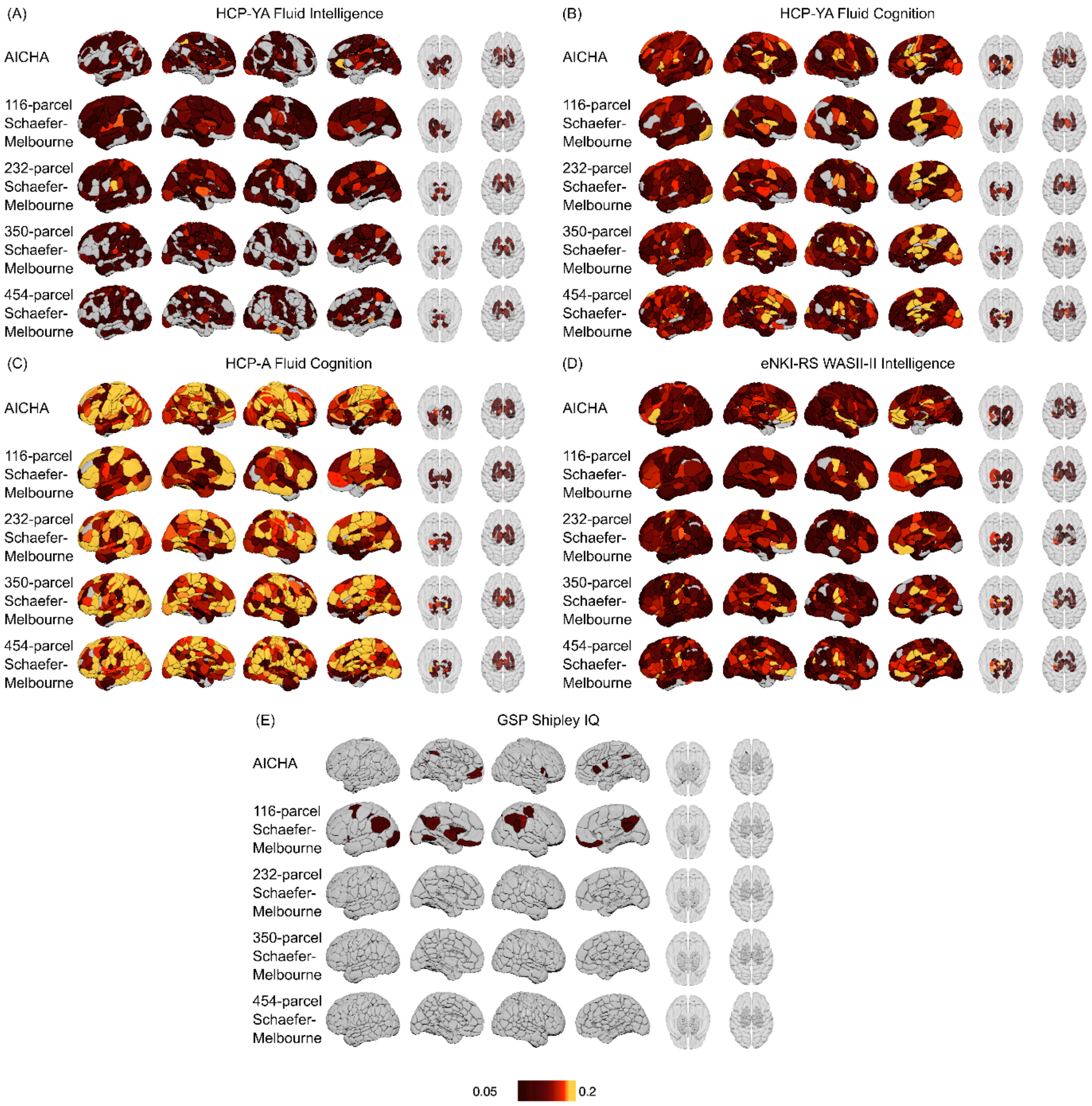
(A)
HCP-YA Fluid Intelligence
(B)
HCP-YA Fluid Cognition
(C)
HCP-A Fluid Cognition
(D)
eNKI-RS WASII-II Intelligence
(E)
GSP Shipley IQ
AICHA
116-parcel Schaefer-Melbourne
232-parcel Schaefer-Melbourne
350-parcel Schaefer-Melbourne
454-parcel Schaefer-Melbourne
0.05
0.2

Figure 3. Prediction accuracy distribution maps of (A) HCP-YA fluid intelligence, (B) HCP-YA fluid cognition, (C) HCP-A fluid cognition, (D) eNKI-RS WASI-II intelligence, and (E) GSP Shipley IQ. Within each section, each row shows the prediction accuracy distribution overlaid on a parcellation used in region-wise CBPP prediction, in lateral and medial views of the left and right cortical hemispheres, as well as the bottom and top views of the subcortical regions. Color represents the magnitude of the prediction accuracies (Pearson correlation between predicted and observed values). Accuracies below 0.05 and non-significant accuracies are shown in gray.

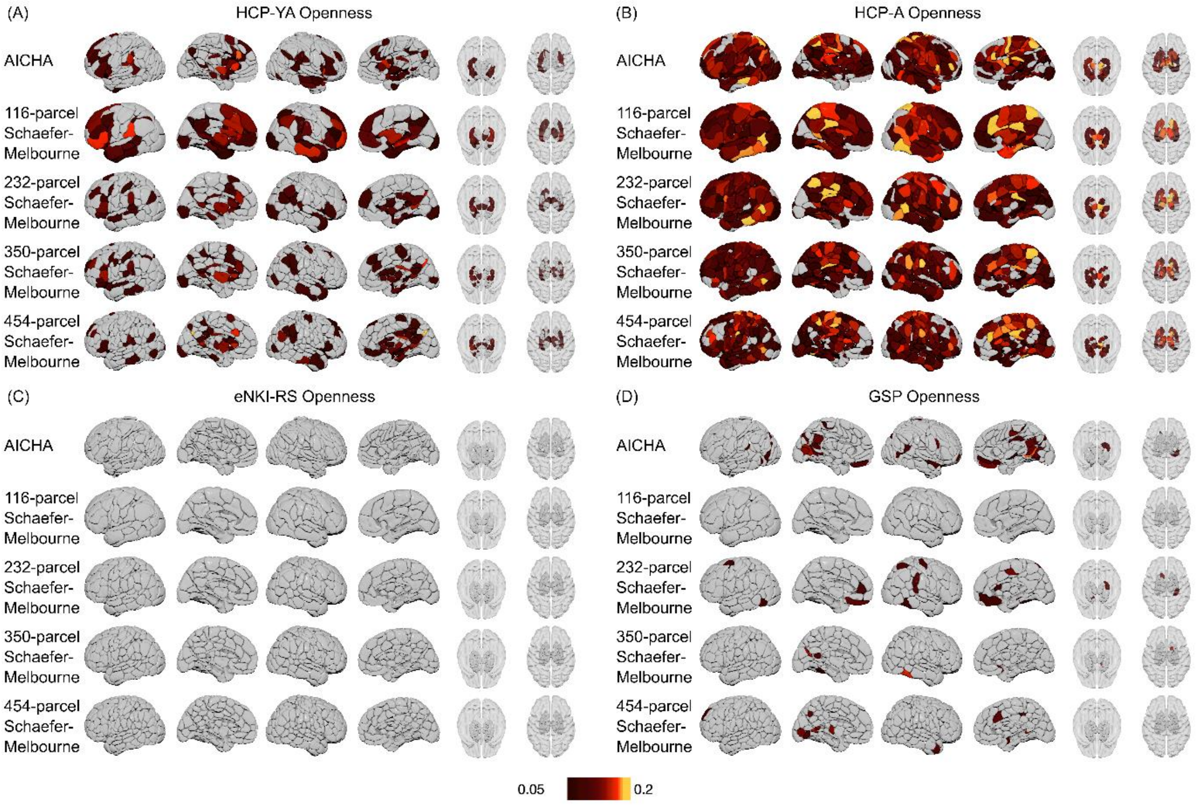
(A)
HCP-YA Openness
(B)
HCP-A Openness
AICHA
116-parcel Schaefer-Melbourne
232-parcel Schaefer-Melbourne
350-parcel Schaefer-Melbourne
454-parcel Schaefer-Melbourne
(C)
eNKI-RS Openness
(D)
GSP Openness
0.05
0.2

Figure 4. Prediction accuracy distribution maps of (A) HCP-YA openness, (B) HCP-A openness, (C) eNKI-RS openness, and (D) GSP openness. Within each section, each row shows the prediction accuracy distribution overlaid on a parcellation used in region-wise CBPP prediction, in lateral and medial views of the left and right cortical hemispheres, as well as the bottom and top views of the subcortical regions. Color represents the magnitude of the prediction accuracies (Pearson correlation between predicted and observed values). Accuracies below 0.05 and non-significant accuracies are shown in gray.

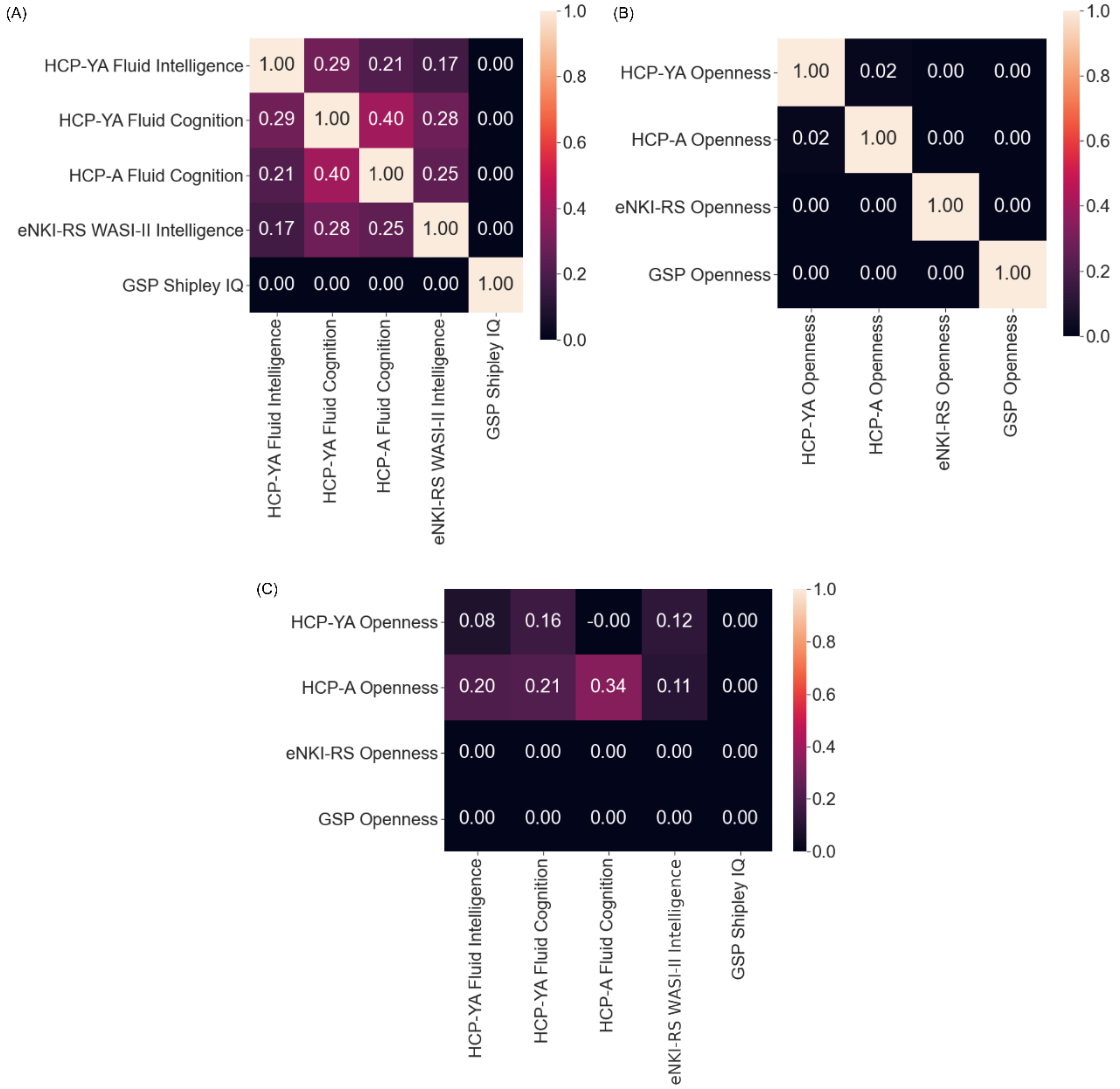

Figure 5. Correlation between prediction accuracy distributions of each pair of psychometric variables, averaged across parcellation choices.

3.3. Whole-brain CBPP Patterns

For the whole-brain CBPP models, we present the prediction patterns derived using the Haufe transformation. Figures 6 and 7 shows the Haufe transformed patterns for the fluid intelligence and openness measures respectively. To avoid visual clutter, only the top 0.05% edges by z-score normalized absolute values for each psychometric variable and each parcellation were shown. Visual inspection showed that little similarity could be found between Haufe transformed patterns between any pair of psychometric variables, or across different parcellations for the prediction of the same psychometric variable.

To quantitatively assess the similarity, or dissimilarity, between the Haufe transformed patterns, we computed the correlation between the activation values of each pair of psychometric variables, averaged across different parcellations (Figure 8). Very low similarity was found between any pair of psychometric variables, although a negative correlation was found between the activation values for GSP Shipley IQ measure and those for GSP openness measure.

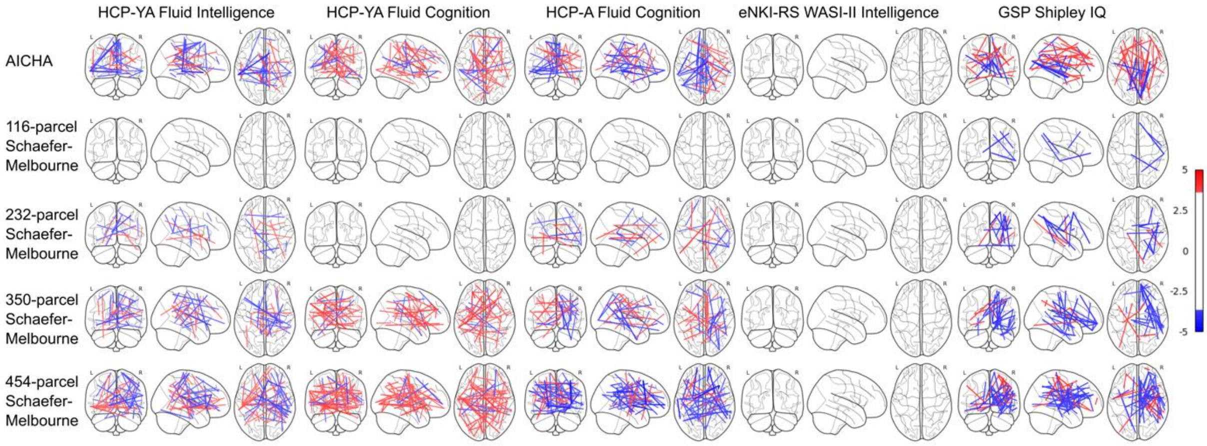
HCP-YA Fluid Intelligence
HCP-YA Fluid Cognition
HCP-A Fluid Cognition
eNKI-RS WASI-II Intelligence
GSP Shipley IQ
AICHA
116-parcel Schaefer-Melbourne
232-parcel Schaefer-Melbourne
350-parcel Schaefer-Melbourne
454-parcel Schaefer-Melbourne
5
2.5
0
-2.5
-5

Figure 6. Haufe transformed patterns of HCP-YA fluid intelligence, HCP-YA fluid cognition, HCP-A fluid cognition, eNKI-RS WASI-II intelligence and GSP Shipley IQ respectively. Each row shows the patterns for predictions using one specific parcellation. In each plot, only significant edges with the top 0.05% z-score normalized absolute values were shown, with positive edges shown in red and negative edges in blue.

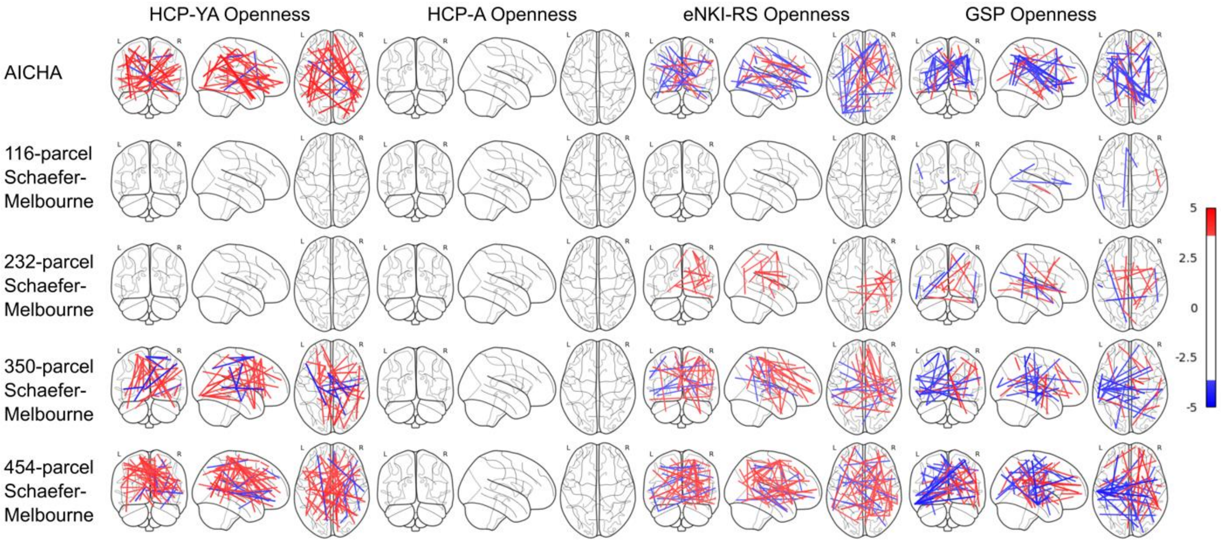
HCP-YA Openness
HCP-A Openness
eNKI-RS Openness
GSP Openness
AICHA
116-parcel Schaefer-Melbourne
232-parcel Schaefer-Melbourne
350-parcel Schaefer-Melbourne
454-parcel Schaefer-Melbourne
L
R
5
2.5
0
-2.5
-5

Figure 7. Haufe transformed patterns of HCP-YA, HCP-A, eNKI-RS and GSP openness measures respectively. Each row shows the patterns for predictions using one specific parcellation. In each plot, only significant edges with the top 0.05% z-score normalized absolute values were shown, with positive edges shown in red and negative edges in blue.

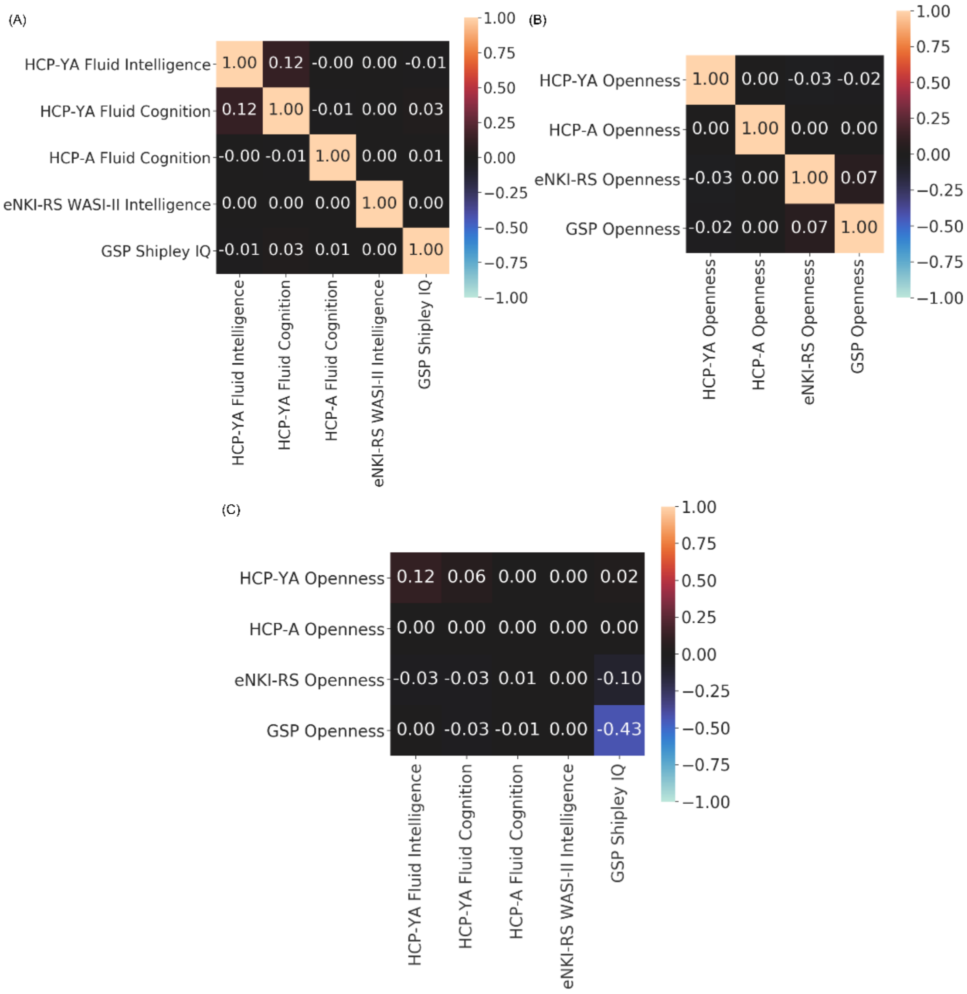

Figure 8. Correlation between Haufe transformed patterns of each pair of psychometric variables, averaged across parcellation choices. Colors were assigned based on the absolute correlation value, since the signs of the Haufe transformed values can be arbitrary.

3.4. Cross-dataset Generalizability

The prediction accuracy distribution maps for cross-dataset predictions from the 3 pairs of datasets are shown in figure 9 (using the AICHA atlas) and figure S4 (using the Schaefer-Melbourne atlases). The within-dataset prediction patterns are shown in the diagonal spots for comparison. In general, the prediction accuracies in cross-dataset predictions (with best region-wise prediction accuracies in the range $r = 0.17 - 0.25$) were lower compared to the within-dataset predictions (with best region-wise prediction accuracies in the range $r = 0.25 - 0.30$).

To facilitate comparison of patterns, the color scales were adjusted and hence are different from those for within-dataset prediction patterns.

For most cross-dataset prediction patterns, some similarity can be observed in comparison with the within-dataset prediction patterns using the same test set. For the HCP-YA fluid cognition measure (i.e., using the HCP-YA data as test set), both cross-dataset and within-dataset models (figure 9 top row) showed better performing parcels in the right anterior insula, left and right anterior cingulate cortex and right supramarginal gyrus. For the HCP-A fluid cognition measure (i.e. using the HCP-A data as test set), both cross-dataset and within-dataset models (figure 9 middle row) showed better performing parcels in the left lateral prefrontal cortex, left and right cingulate cortex, left and right lateral temporal lobe, left and right supramarginal gyrus, left and right precuneus and left and right anterior insula. For the eNKI-RS WASI-II intelligence measure (i.e., using the eNKI-RS data as test set), the cross dataset models trained on HCP-A data and the within-dataset models (figure 9 bottom row, middle and rightmost columns) showed better performing parcels in left medial prefrontal cortex and right middle cingulate cortex.

Figure 10 shows the correlation between the prediction accuracy distributions of both within-dataset and cross-dataset predictions for these fluid intelligence measures. Overall, cross-dataset prediction patterns were still similar, to some extent, to the within-dataset prediction patterns of each fluid intelligence measure respectively. For most cross-dataset prediction pattern, the correlation to the within-dataset prediction pattern using the same test set is higher than the correlation to the within-dataset prediction pattern using the same training set. Finally, the cross-dataset prediction patterns are more similar to other cross-dataset prediction patterns using the same pair of datasets than to any within-dataset prediction pattern. For instance, the most similar pair of prediction patterns are between the cross-dataset prediction patterns using the HCP-YA and HCP-A datasets, but swapping the training and test set ($r = 0.76$).

By evaluating region-wise CBPP models in out-of-sample test sets, a certain degree of generalizability of the prediction models could be observed ($r = 0.13 - 0.52$), especially when the brain prediction patterns were compared to the within-dataset patterns where the same test set was used ($r = 0.24 - 0.52$). Notably, the highest similarity between prediction patterns was observed when two models are trained and tested on the same pair of datasets ($r = 0.66, 0.50, 0.38$).

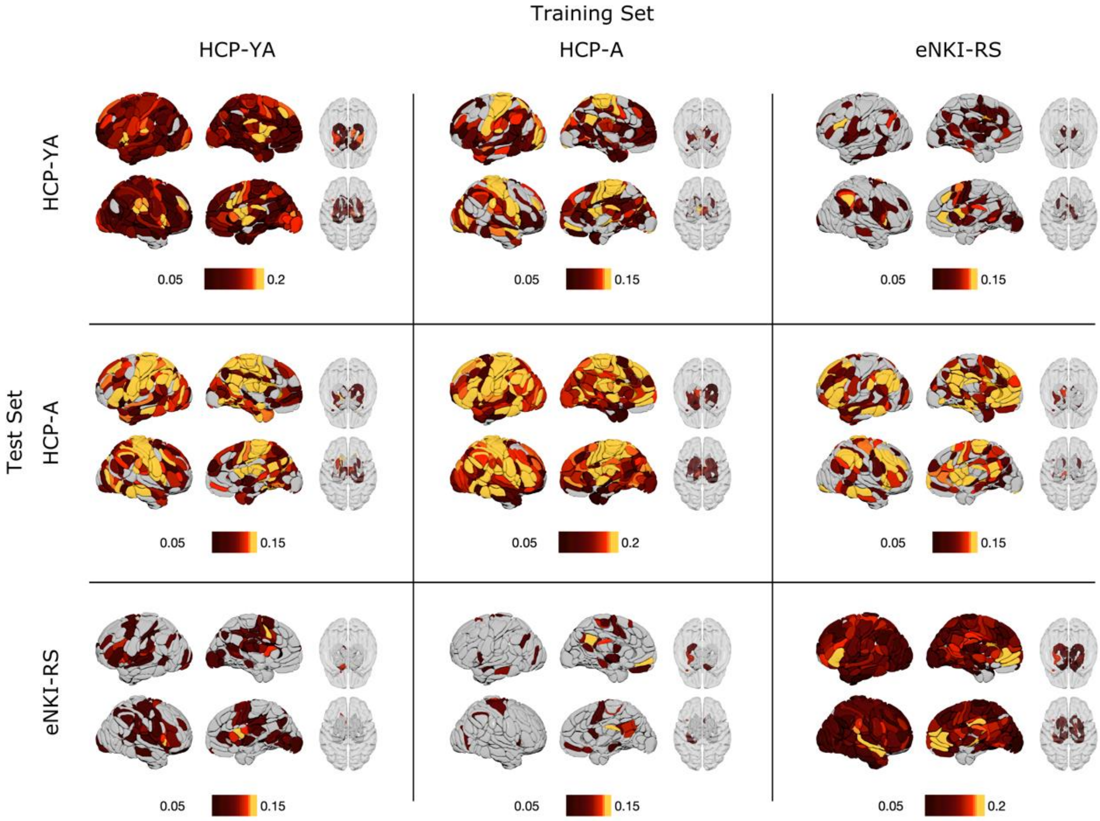

Figure 9. Prediction accuracy distribution maps of within-dataset and cross-dataset predictions using the AICHA atlas, arranged according to the training set and test set involved. Each plot shows the prediction accuracy distribution overlaid on the AICHA atlas, in lateral and medial views of the left and right cortical hemispheres, as well as the bottom and top views of the subcortical regions. Color represents the magnitude of the prediction accuracies (Pearson correlation between predicted and observed values). Accuracies below 0.05 and non-significant accuracies are shown in gray.

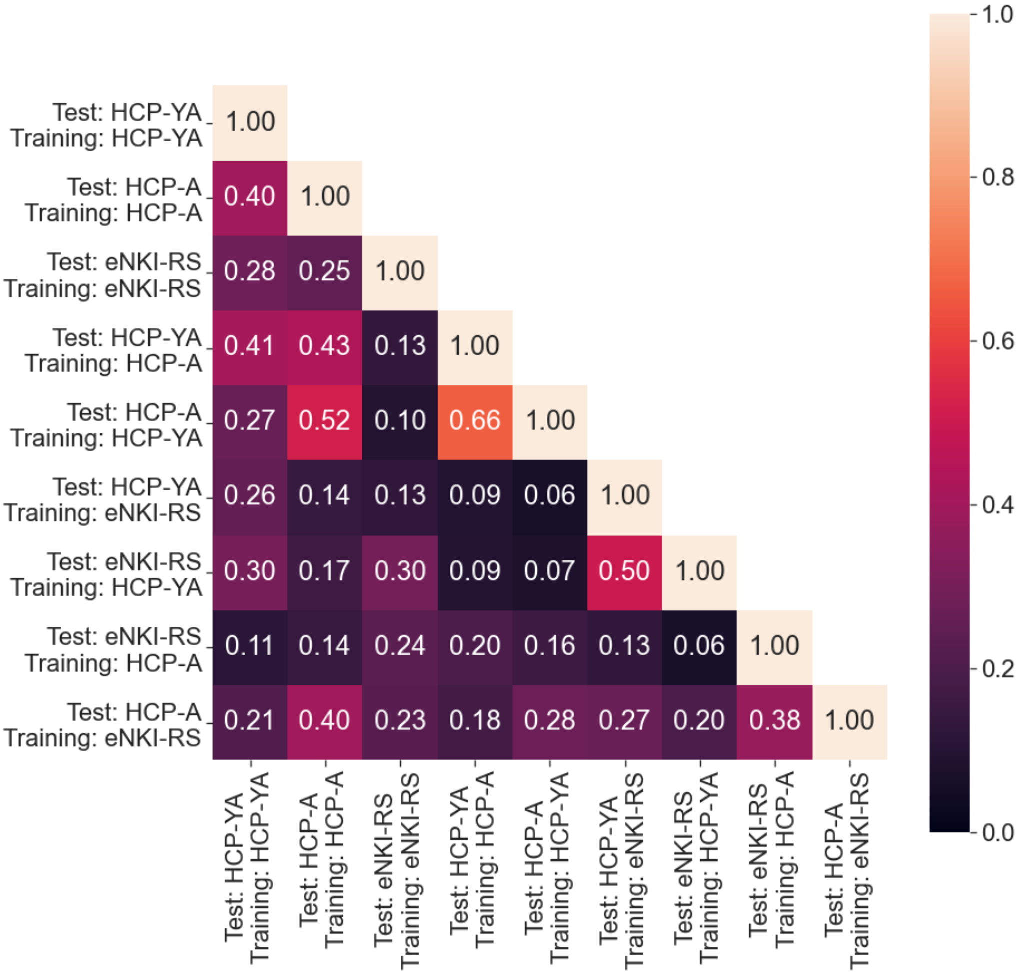

Figure 10. Correlation between within-dataset and cross-dataset prediction patterns of HCP-YA fluid cognition, HCP-A fluid cognition, and eNKI-RS WAISI-II intelligence. Each box shows the Pearson's correlation between one pair of prediction patterns, average across all 5 atlas options.

4. Discussion

To investigate the replicability of connectivity-based psychometric prediction (CBPP) patterns across distinct population neuroscience cohorts, we compared the prediction performance and prediction patterns based on whole-brain and region-wise CBPP models for two psychometric variables in four separate large datasets. Similar prediction accuracies can be achieved in most cases. However, low similarity in prediction patterns was observed across datasets, illustrating the difficulty in cross-cohort replicability of brain prediction patterns. Similarly, generalizability of prediction models trained on one dataset to a new dataset could only be achieved to a low to moderate extent. In our examination of replicability of prediction patterns, we noted higher similarity between the region-wise prediction patterns of the HCP-YA fluid cognition measure, the HCP-A fluid cognition measure, and the eNKI-RS WASI-II intelligence measure, suggesting potential replicability for intelligence measure. In addition, our analysis demonstrated the usefulness of the region-wise CBPP approach (Wu et al. 2021) for comparing prediction results based on brain-behavior association patterns. Due to the inherent low psychometric prediction accuracies based on RSFC, focusing on a set of parcels with relatively higher prediction accuracies may be helpful for both the power and the generalizability of an analysis.

4.1. The HCP-YA Cohort as an overoptimistic benchmark

Many existing studies showed similar prediction accuracies in the range of 0.2 to 0.25 for fluid intelligence and the openness trait in the HCP-YA cohort (Smith et al. 2016; Noble et al. 2017; Dubois et al. 2018a; Dubois et al. 2018b; Li et al. 2019; Pervaiz et al. 2020; Wu et al. 2021; Kong et al. 2021). However, our results demonstrated that similar accuracies should not be assumed to be achievable in other cohorts. As the HCP-YA dataset contains a large sample of high-quality imaging data (long scan durations, short TR and high resolution) which is often not available in other datasets, it may not be advisable to use the accuracy values reported for the HCP-YA cohort as benchmarks for other cohorts. Because the majority of existing CBPP studies made use of the HCP-YA cohort; the reported prediction accuracies from these studies may lead to overoptimistic expectations of prediction performance for future studies. Such expectations may indirectly lead to some form of hacking, in order to match such expectations, which in turn would maintain and further contribute to unrealistic expectations with regards to performance. We would hence here suggest that underperforming results may be reported in reference to the characteristics of the cohort(s).

4.2. Replicability of Brain Prediction Patterns for Fluid Cognition and Openness Predictions

Furthermore, similar prediction accuracies do not imply that similar interpretation can be made based on the prediction model. For instance, while personality traits scores are generally thought to be stable within individuals (Murray et al. 2003; McCrae and Costa 2004; Dubois et al. 2018a), the non-revised NEO-FFI openness score (Costa and McCrae 1992; used in the GSP cohort) may suffer from lower reliability and lack of congruency in distinct samples (Caruso 2000; Egan et al. 2000). In our case, even though we achieved similar prediction accuracies for HCP-YA openness, HCP-A openness and GSP openness, we did not find similar prediction patterns between predictions of these three measures. In contrast, for the fluid cognition measure, we have observed some similarity in prediction patterns between HCP-YA, HCP-A and eNKI-RS cohorts, with moderate correlations ($r = 0.31 - 0.35$). Overall, our results suggested better replicability from test performance measures like fluid cognition than self-reported measures like openness, when different versions or types of measurement were used across cohorts.

For fluid cognition prediction, for which some similarity was observed between the prediction patterns of HCP-YA fluid cognition, HCP-A fluid cognition and eNKI-RS WASI-II intelligence, we demonstrated that a subset of the within-dataset prediction patterns could also be replicated with cross-dataset predictions. This convergence may reveal the common brain regions related to fluid cognition, namely the right prefrontal cortex, the right anterior insula, left anterior cingulate cortex and right supramarginal gyrus, when considering all three cohorts. The prefrontal cortex, anterior cingulate cortex and supramarginal gyrus have been identified as neural correlates for fluid intelligence using task-based fMRI, possibly involved in attentional control, executive control, and visualization (Kane and Engle 2002; Gray et al. 2003; Preusse et al. 2011; Ebisch et al. 2012; Santarnecchi et al. 2021). Furthermore, the right prefrontal cortex and right inferior parietal cortex have been identified to be associated with fluid intelligence in lesion studies, possibly involved in working memory and spatial processing (Roca et al. 2010; Barbey et al. 2013). All four regions have been identified as core functional hubs in network analyses (van den Heuvel and Sporns 2013), and hence may be related to efficiency of information processing in general. These regions were also observed to have higher interindividual variability, with frequent reports of associations between their FC measures and individual differences across different cognitive domains (Mueller et al. 2012). Following these observations, it may be recommended to focus on these key regions when the number of predictive features has to be limited, not only specifically for fluid intelligence prediction, but also for behavioral phenotype in general, as these regions represent crucial hub for which interindividual variability importantly matters.

4.3. Generalizability of Prediction Models

Many studies assessed the generalizability of their prediction models by applying them to a new cohort (Rosenberg et al. 2016; Beaty et al. 2018; Jiang et al. 2018; Avery et al. 2020; Jiang et al. 2020; Speer et al. 2022). In most cases, the generalizability of the developed prediction model seems promising. However, our results suggest that the generalizability of models may be less optimistic when assessed using the region-wise prediction patterns.

Varying degree of similarity was observed between cross-dataset prediction patterns and their corresponding within-dataset prediction patterns ($r = 0.16 - 0.53$) in this study. Overall, the higher correlations and visual similarity were observed between cross-dataset prediction patterns and within-dataset prediction patterns using the same test set. In contrast to our initial speculation, prediction patterns seem to depend more on the brain-behavior association in the test data than in the training data. It is possible that, in cross-dataset predictions, the final prediction pattern indicates an overlap of the specific brain-behavior association patterns in the training set and those in the test set. As the final prediction pattern was generated in the test set, the brain-behavior association patterns specific to the test set may be more prominently observed, hence the higher similarity to the within-dataset prediction pattern generated using the same test set. Moreover, the highest similarity was observed between cross-dataset prediction patterns using the same pair of datasets but swapping the training and test set. In this case, both prediction patterns would be relating to the overlap of specific brain-behavior association patterns in the same two sets of data, hence the high similarity. Indeed, the most similar pair of cross-dataset prediction patterns used the HCP-YA and the HCP-A datasets, where most similar data collection protocols were used, where more overlap in brain-behavior association patterns may be expected.

Finally, for both the reliability and generalizability analyses, our findings are mostly consistent across granularity, except at the 116-parcel level where pattern similarities were generally lower (see figure S5). Thus, our results would tend to suggest that higher granularities (≥200) offer better representations of features for machine learning approaches, in line with previous studies (Arslan et al. 2018; Varikuti et al. 2018).

4.4. A Possible Effect of Data Collection and Processing Protocols on reliability and generalizability

From our findings of the highest similarity between the cross-dataset prediction patterns being observed for the HCP-YA and the HCP-A datasets, we suggest that data collection and processing protocols may be the most influential factor in cross-cohort replicability and generalizability. It should be acknowledged, however, that comparison between different datasets is not straightforward, and that differences in prediction accuracies or patterns could not attributed to any specific factor with certainty. As this work is focused on assessing the extent of replicability and generalizability of prediction patterns, our results were limited in finding the exact causes for the lack of replicability or generalizability. Future work with methods specifically designed would be required to identify the actual causes. Potentially, several factors influencing the prediction accuracies or patterns could be suggested for further investigation, including the differences in psychometric test implementation, sample characteristics, imaging protocols and data quality across the different cohorts.

For the GSP cohort, the shorter time series may have undermined both the prediction accuracies and interpretation based on prediction patterns. Scan duration and number of scans have been reported to influence the reliability of RSFC (Mueller et al. 2015; Shah et al. 2016; Noble et al. 2017; Noble et al. 2019), which could in turn affect the predictive power of the derived FC features. A previous study has also shown that prediction accuracies in GSP were lower than those in HCP-YA across multiple behavior phenotypes, not only for fluid cognition and openness (Li et al. 2019). Thus, altogether these findings may suggest that the length of the time series may play an important role in the reliability and validity of the connectivity-based prediction of behavioral phenotype in healthy populations.

One potential factor not investigated in this work is the sample size, which is less concerning when large open datasets and the easy-to-acquire resting-state data are used. However, in studies using recruited subjects or task-based fMRI (Beaty et al. 2018; Christov-Moore et al. 2020; Kwon et al. 2021; Speer et al. 2022), sample sizes tend to be much smaller. Most self-recruited sample includes fewer than 100 subjects and may suffer from shorter scan duration (Yeung et al. 2022). While some task-based fMRI samples include more than 100 subjects, the reproducibility of the analysis may still suffer if the amount of data for each subject is insufficient (Turner et al. 2018; Nee 2019).

4.5. Region-wise Models for Brain Prediction Pattern Analysis and Identification of Key Predictive Regions

The prediction accuracy distribution maps based on region-wise models (Wu et al. 2021) were particularly useful in our analysis, as it allows comparisons to be made with direct reference to brain regions' contributions to the prediction of each psychometric variable. In contrast, the

Haufe transformed patterns were harder to interpret. As there is little similarity between the Haufe transformed patterns from different atlases or cohorts, we were not able to identify a representative pattern for any psychometric variable based on the Haufe transformed weights. It should be noted that how well the Haufe transformed patterns captured the true brain-behavior relationships is mostly dependent on the accuracy of the backward regression model (Haufe et al. 2014). As the part of variance in psychometric variables in healthy adult population that can be explained by interindividual variability in RSFC is limited, the field of CBPP inherently suffers from low prediction performance. Therefore, the validity of the regression weights themselves, based on models with low predictive power, may be questionable, further limiting analyses relying on interpreting these weights. While it has been shown that the transformation improves the robustness of the weight patterns in comparison to using regression weights directly (Chen et al. 2022), it has been also shown that the Haufe transformed weights themselves generally have poor reliability (Tian and Zalesky 2021). Our results hence converged with previous findings by revealing the poor reliability of the Haufe transformed patterns.

Several other studies have also investigated the replicability of feature importance under the framework of Connectome Predictive Modelling (CPM). As the CPM process selects features most correlated with the prediction target based on the training set, the group of selected features can be considered a representation of prediction pattern too. A pattern consisting of features selected in all cross-validation splits can be used to infer brain-behavior association pattern (Jiang et al. 2018; Jiang et al. 2020). Nevertheless, this feature selection has been shown to be unstable even in the same dataset across cross-validation splits (O’Connor et al. 2021). More research would be required to assess the extent to which these selected features can be related to the underlying brain-behavior association.

Despite the overall low prediction accuracies, the region-wise accuracy distribution maps allow researchers to easily identify and hence focus on the important regions for the psychometric variable in question. From a general and technical perspective, the optimal number of features for the highly correlated RSFC features in a dataset of N subjects would be $\sqrt{N}$ (Jain and Waller 1978). As a result, studies with relatively small sample sizes, for instance, studies using clinical population or locally acquired datasets have to select a low number of features. This is usually done by focusing on canonical networks (such as the default mode network or the cognitive control network) derived from task-fMRI meta-analysis or RSFC network atlases (Nostro et al. 2018; Chen et al. 2020; Plaeschke et al. 2020). Nevertheless, such approaches are prone to neglecting potentially relevant regions because those where not highlighted in activation studies or not included in the selected RSFC network. These regions may, nonetheless, be involved in information processing and cognitive processes in general, or have high interindividual variability in terms of the features used for prediction, such as the three regions identified in our fluid cognition predictions. In this context, the region-wise CBPP approach can help to select a set of key regions in a data-driven fashion.

Nonetheless, the utility of the region-wise CBPP approach for feature selection is dependent on, and limited by, the generalizability of the region-wise prediction patterns. Key regions for predicting a psychometric variable may be identified by selecting the better performing brain regions in the prediction patterns. However, in order to use these key regions in the smaller sample of interest, the predictive model based on these regions need to be generalizable to the new sample. Overall, our results suggested limited generalizability of region-wise prediction patterns, although a small set of key regions did show decent generalizability. Therefore, the prediction patterns derived from the region-wise approach should be interpreted with caution. The key regions identified may serve as a general guidance to select the regions of interest, but not always applicable directly to new cohorts.

It should be noted that the comparison between the Haufe transformation and the region-wise approach was only discussed in relation to brain prediction pattern analysis. As our focus was on assessing the replicability and generalizability of prediction patterns, we used the same linear model for all approaches to make them more comparable. Therefore, our results were not indicative of the predictive power of the region-wise approach or the whole-brain approach. In cases where the focus is on maximizing prediction performance, the whole-brain approach should be preferred. As demonstrated in our previous work (Wu et al. 2021), implementing region-wise models in addition to the whole-brain model can help to bring additional insights into the relevant brain-behavior association. As the whole-brain and region-wise approaches address different objectives and applications of connectivity-based prediction, they appear as two complementary rather than competitive approaches for the field.

4.6. Conclusion

To conclude, we examined the convergence and divergence of connectivity-based psychometric prediction patterns across four distinct population neuroscience cohorts. While similar prediction accuracies could be achieved for several fluid intelligence and NEO-FFI openness measures across cohorts, the prediction patterns could not be replicated across cohorts in many cases. In the case where the prediction patterns were partly replicated for the prediction of fluid cognition across HCP-YA, HCP-A and eNKI-RS cohorts, we further demonstrated that some extent of cross-dataset generalizability could be achieved. Accordingly, making use of a region-wise CBPP approach, we revealed a set of common brain regions potentially involved in fluid cognitive ability, hence demonstrating that region-wise CBPP could provide regions of interest in a data-driven way for future studies in smaller cohorts to focus on. In view of our results, we caution researchers to not be overoptimistic in replicating brain-behavior relationships discoveries in distinct cohorts. While many large population neuroimaging datasets are now available, predictive models and the corresponding brain-behavior association patterns identified based on such datasets could still only represent a small portion of the general population. Generalizing these models and patterns to the general population may remain a challenge for a long time.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, GE 2835/2–1, EI 816/ 4–1), the Helmholtz Portfolio Theme ‘Supercomputing and Modelling for the Human Brain’ and the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No. 720270 (HBP SGA1) and Grant Agreement No. 785907 (HBP SGA2).

SBE acknowledges funding by the European Union’s Horizon 2020 Research and Innovation Program (grant agreements 945539 (HBP SGA3) and 826421 (VBC)), the Deutsche Forschungsgemeinschaft (DFG, SFB 1451 & IRTG 2150) and the National Institute of Health (R01 MH074457).

BTTY is supported by the Singapore National Research Foundation (NRF) Fellowship (Class of 2017), the NUS Yong Loo Lin School of Medicine (NUHSRO/2020/124/TMR/LOA), the Singapore National Medical Research Council (NMRC) LCG (OFLCG19May-0035), NMRC STaR (STaR20nov-0003) and the USA NIH (R01MH120080). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Singapore NRF or the Singapore NMRC.

For the HCP-YA cohort, 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.

For the HCP-A cohort, research was supported by the National Institute On Aging of the National Institutes of Health under Award Number U01AG052564. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Ntional Institutes of Health. The associated study ID is 1376 (http://dx.doi.org/10.15154/1524254).

For the GSP cohort, data were provided by the Brain Genomics Superstruct Project of Harvard University and the Massachusetts General Hospital (Principal Investigators: Randy Buckner, Joshua Roffman, and Jordan Smoller), with support from the Center for Brain Science Neuroinformatics Research Group, the Athinoula A. Martinos Center for Biomedical Imaging, and the Center for Human Genetic Research. 20 individual investigators at Harvard and MGH generously contributed data to the overall project.

The authors also thank Laura Waite for curation of the DataLad datasets used in the work.

References

Arslan, S., Ktena, S.I., Makropoulos, A., Robinson, E.C., Rueckert, D., Parisot, S., 2018. Human brain mapping: A systematic comparison of parcellation methods for the human cerebral cortex. NeuroImage. 170, 5-30. https://doi.org/10.1016/j.neuroimage.2017.04.014.

Avery, E.W., Yoo, K., Rosenberg, M.D., Greene, A.S., Gao, S., Na, D.L., Scheinost, D., Constable, T.R., Chun, M.M., 2020. Distributed patterns of functional connectivity predict working memory performance in novel healthy and memory-impaired individuals. J. Cogn. Neuroscie. 32, 241-255. https://doi.org/10.1162/jocn_a_01487.

Barbey, A.K., Colom, R., Paul, E.J., Grafman, J., 2013. Architecture of fluid intelligence and working memory revealed by lesion mapping. Brain Struct. Funct. 219, 485-494. https://doi.org/10.1007/s00429-013-0512-z.

Beaty, R.E., Kenett, Y.N., Christensen, A.P., Rosenberg, M.D., Benedek, M., Chen, Q., Fink, A., Qui, J., Kwapil, T.R., Kane, M.J., Silvia, P.J., 2018. Robust prediction of individual creative ability from brain functional connectivity. Proc. Natl. Acad. Sci. U.S.A. 115, 1087-1092. https://doi.org/10.1073/pnas.1713532115.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. 59, 289-300. http://doi.wiley.com/10.1111/j.2517-6161.1995.tb02031.x.

Bilker, W.B., Hansen, J.A., Brensinger, C.M., Richard, J., Gur, R.E., Gur R.C., 2012. Development of abbreviated nine-item forms of the Raven's standard progressive matrices test. Assess. 19, 354-369. https://doi.org/10.1177/1073191112446655.

Bookheimer, S.Y., Salat, D.H., Terpstra, M., Ances, B.M., Barch, D.M., Buckner, R.L., Burgess, G.C., Curtiss, S.W., Diaz-Santos, M., Elam, J.S., et al., 2019. The lifespan Human Connectome Project in aging: An overview. NeuroImage. 185, 335-348. https://doi.org/10.1016/j.neuroimage.2018.10.009.

Boser, E.B., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. Proc. Comput. Learn. Theory. 144-152. https://doi.org/10.1145/130385.130401.

Caruso, J.C., 2000. Reliability generalization of the NEO personality scales. Educ. Psychol. Meas. 60, 236-254. https://doi.org/10.1177/00131640021970484.

Caspers, S., Moebus, S., Lux, S., Pundt, N., Schütz, H., Mühleisen, T.W., Gras, V., Eickhoff, S.B., Romanzetti, S., Stöcker, T., et al., 2014. Studying variability in human brain aging in a population-based German cohort – Rationale and design of 1000BRAINS. Front. Aging Neurosci. 6, 149. https://doi.org/10.3389/fnagi.2014.00149.

Chen, J., Mueller, V.I., Dukart, J., Hoffstaedter, F., Baker, J.T., Holmes, A.J., Vatansever, D., Nickl-Jockschat, T., Liu, X., Derntl, B., et al., 2020. Intrinsic connectivity patterns of task-defined brain networks allow individual prediction of cognitive symptom dimension of schizophrenia and are linked to molecular architecture. Biol. Psychiatry. 89, 308-319. https://doi.org/10.1016/j.biopsych.2020.09.024.

Chen, J., Tam, A., Kebets, V., Orban, C., Ooi, L.Q.R., Asplund, C.L., Marek, S., Dosenbach, N.U.F., Eickhoff, S.B., Bzdok, D., Holmes, A.J., Yeo, B.T.T., 2022. Shared and unique brain network features predict cognitive, personality, and mental health scores in the ABCD study. Nat. Commun. 13, 2217. https://doi.org/10.1038/s41467-022-29766-8.

Christov-Moore, L., Reggente, N., Douglas, P.K., Feusner, J.D., Iacoboni, M., 2020. Predicting empathy from resting state brain connectivity: A multivariate approach. Front. Integr. Neurosci. 14, 3. https://doi.org/10.3389/fnint.2020.00003.

Cortes, C., Vapnik, V.N., 1995. Support-vector networks. Mach. Learn. 20, 273-297. https://doi.org/10.1007/BF00994018.

Costa, P.T., McCrea, R.R., 1992. NEO PI-R Professional Manual. Psychol. Assess. 4, 5-13.

Deary, I.J., Weiss, A., Batty, D., 2011. Intelligence and personality as predictors of illness and death: How researchers in differential psychology and chronic disease epidemiology are collaborating to understand and address health inequalities. Psychol. Sci. Public Interest. 11, 53-79. https://doi.org/10.1177/1529100610387081.

Dubois, J., Galdi, P., Han, Y., Paul, L.K., Adolphs, R., 2018a. Resting-state functional brain connectivity best predicts personality dimension of openness to experience. Pers. Neurosci. 1, E6. https://doi.org/10.1017/pen.2018.8.

Dubois, J., Galdi, P., Paul, L.K., Adolphs, R., 2018b. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. Philos. Trans. R. Soc. B Biol. Sci. 373, 20170284. https://doi.org/10.1098/rstb.2017.0284.

Ebisch, S.J., Perrucci, M.G., Mercuri, P., Romanelli, R., Mantini, D., Romani, G.L., Colom, R., Saggino, A., 2012. Common and unique neuro-functional basis of induction, visualization, and spatial relationships as cognitive components of fluid intelligence. NeuroImage. 62, 331-342. https://doi.org/10.1016/j.neuroimage.2012.04.053.

Egan, V., Deary, I., Austin, E., 2000. The NEO-FFI: emerging British norms and an item-level analysis suggest N, A and C are more reliable than O and E. Pers. Individ. Differ. 29, 907-920. https://doi.org/10.1016/S0191-8869(99)00242-1.

Esteban, O., Markiewicz, C.J., Blair, R.W., Moodie, C.A., Isik, A.I., Erramuzpe, A., Kent, J.D., Goncalves, M., DuPre, E., Snyder, M., et al., 2019. fMRIPrep; A robust preprocessing pipeline for functional MRI. Nat. Methods. 16, 111-116. https://doi.org/10.1038/s41592-018-0235-4.

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetric, 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.

Geake, J.G., Hansen, P.C., 2005. Neural correlates of intelligence as revealed by fMRI of fluid analogies. NeuroImage. 26, 555-564. https://doi.org/10.1016/j.neuroimage.2005.01.035.

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., et al., 2013. The minimal preprocessing pipeline for the Human Connectome Project. NeuroImage. 80, 105-124. https://doi.org/10.1016/j.neuroimage.2013.04.127.

Gray, J.R., Chabris, C.F., Braver, T.S., 2003. Neural mechanisms of general fluid intelligence. Nat. Neurosci. 6, 316-322. https://doi.org/10.1038/nn1014.

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., et al., 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.

Halchenko, Y.O., Meyer, K., Poldrack, B., Solanky, D.S., Wagner, A.S., Gors, J., MacFarlane, D., Pustina, D., Sochat, V., Ghosh, S.S., et al. (2021). DataLad: distributed system for joint management of code, data, and their relationship. Journal of Open Source Software, 6, 3262. https://joss.theoj.org/papers/10.21105/joss.03262.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J., Blankertz B., Bießmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage. 87, 96-110. https://doi.org/10.1016/j.neuroimage.2013.10.067.

Harms, M.P., Somerville, L.H., Ances, B.M., Andersson, J., Barch, D.M., Bastiani, M., Bookheimer, S.Y., Brown, T.B., Buckner, R.L., Burgess, G.C., 2018. Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. NeuroImage. 183, 972-984. https://doi.org/10.1016/j.neuroimage.2018.09.060.

He, T., Kong, R., Holmes, A.J., Nguyen, M., Sabuncu, M.R., Eickhoff, S.B., Bzdok, D., Feng, J., Yeo, B.T.T., 2020. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. NeuroImage. 206, 116276. https://doi.org/10.1016/j.neuroimage.2019.116276.

Holmes, A.J., Hollinshead, M.O., O'Keefe, T.M., Petrov, V.I., Fariello, G.R., Wald, L.L., Fischl, B., Rosen, B.R., Mair, R.W., Roffman, J.L., et al., 2015. Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. Sci. Data. 2, 150031. https://doi.org/10.1038/sdata.2015.31.

Horien, C., Greene, A.S., Constable, T., Scheinost, D., 2019. Regions and connections: Complementary approaches to characterize brain organization and function. Neuroscientist. 26, 117-133. https://doi.org/10.1177/1073858419860115.

Humphreys, M.S., Revelle, W., 1984. Personality, motivation, and performance: A theory for the relationship between individual differences and information processing. Psychol. Rev. 91, 153-184. https://doi.org/10.1037/0033-295X.91.2.153.

Jain A.K., Waller, W.G., 1978. On the optimal number of features in the classification of multivariate Gaussian data. Pattern Recognit. 10, 365-374. https://doi.org/10.1016/0031-3203(78)90008-0.

Jiang, R., Calhoun, V.D., Zuo, N., Lin, D., Li, J., Fan, L., Qi, S., sun, H., Fu, Z., Song, M., et al., 2018. Connectome-based individualized prediction of temperament trait scores. NeuroImage. 183, 366-374. https://doi.org/10.1016/j.neuroimage.2018.08.038.

Jiang, R., Calhoun, V., Fan, L., Zuo, N., Jung, R., Qi, S., Lin, D., Li, J., Zhuo, C., Song, M., 2020. Gender differences in connectome-based predictions of individualized intelligence quotient and sub-domain scores. Cereb. Cortex. 30, 888-900. https://doi.org/10.1093/cercor/bhz134.

Joliot, M., Jobard, G., Naveau, M., Delcroix, N., Petit, L., Zago, L., Crivello, F., Mellet, E., Mayzoyer, B., Tzourio-Mazoyer, N., 2015. AICHA: An atlas of intrinsic connectivity of homotopic areas. J. Neurosci. Methods. 254, 46-59. https://doi.org/10.1016/j.jneumeth.2015.07.013.

Kane, M.J., Engle, R.W., 2002. The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. Psychon. Bull. Rev. 9, 637-671. https://doi.org/10.3758/BF03196323.

Kong, R., Yang, Q., Gordon, E., Xue, A., Yan, X., Orban, C., Zuo, X., Spreng, N., Ge, T., Holmes, A., et al., 2021. Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. Cereb. Cortex. 31, 4477-4500. https://doi.org/10.1093/cercor/bhab101.

Kwon, Y.H., Yoo, K., Nguyen, H., Jeong, Y., Chung, M.M., 2021. Predicting multilingual effects on executive function and individual connectomes in children: An ABCD study. Proc. Natl. Acad. Sci. U.S.A. 118, e2110811118. https://doi.org/10.1073/pnas.2110811118.

Li, J., Kong, R., Liegeois, R., Orban, C., Tan, Y., Sun, N., Holmes, A., Sabuncu, M.R., Ge, T., Yeo, B.T.T., 2019. Global signal regression strengthens association between resting-state functional connectivity and behaviour. NeuroImage. 196, 126-141. https://doi.org/10.1016/j.neuroimage.2019.04.016.

Maglanoc, L.A., Kaufmann, T., van der Meer, D., Marquand, A.F., Wolfers, T., Jonassen, R., Hilland, E., Andreassen, O.A., Landrø, N.I., Westlye, L.T., 2019. Brain connectome mapping of complex human traits and their polygenic architecture using machine learning. Biol. Psychol. 87, 717-726. https://doi.org/10.1016/j.biopsych.2019.10.011.

McCrae, R.R., Costa, P.T., 2004. A contemplated revision of the NEO Five-Factor Inventory. Pers. Individ. Differ. 36, 587-596. https://doi.org/10.1016/S0191-8869(03)00118-1.

Mueller, S., Wang, D., Fox, M.D., Yeo, B.T.T., Sepulcre, J., Sabuncu, M.R., Shafee, R., Lu, J., Liu, H., 2012. Individual variability in functional connectivity architecture of the human brain. Neuron. 77, 586-595. https://doi.org/10.1016/j.neuron.2012.12.028.

Mueller, S., Wang, D., Fox, M.D., Pan, R., Lu, J., Li, K., Sun, W., Buckner, R.L., Liu, H., 2015. Reliability correction for functional connectivity: Theory and implementation. 36, 4664-4680. https://doi.org/10.1002/hbm.22947.

Murray, G., Rawlings, D., Allen, N.B., Trinder, J., 2003. NEO Five-Factor Inventory scores: Psychometric properties in a community sample. Meas. Eval. Couns. Dev. 36, 140-149. https://doi.org/10.1080/07481756.2003.11909738.

Nee, D.E., 2019. fMRI replicability depends upon sufficient individual-level data. Commun. Biol. 2, 130. https://doi.org/10.1038/s42003-019-0378-6.

Noble, S., Spann, M.N., Tokoglu, F., Shen, X., Constable, R.T., Scheinost, D., 2017. Influences on the test-retest reliability of functional connectivity MRI and its relationship with behavioural utility. Cereb. Cortex. 27, 5415-5429. https://doi.org/10.1093/cercor/bhx230.

Noble, S., Scheinost, D., Constable, R.T., 2019. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. NeuroImage. 203, 116157. https://doi.org/10.1016/j.neuroimage.2019.116157.

Nooner, K.B., Colcombe, S.J., Tobe, R.H., Mennes, M., Benedict, M.M., Moreno, A.L., Panek, L.J., Brown, S., Zavitz, S.T., Li, Q., et al., 2012. The NKI-Rockland Sample: A model for accelerating the pace of discovery science in psychiatry. Front. Neurosci. 6, 152. https://doi.org/10.3389/fnins.2012.00152.

Nostro, A.D., Mueller, V., Varikuti, D., Plaeschke, R., Hoffstaedter, F., Langner, R., Patil, K., Eickhoff, S.B., 2018. Predicting personality from network-based resting-state functional connectivity. Brain Struct. Funct. 223, 2699-2719. https://doi.org/10.1007/s00429-018-1651-z.

O'Connor, D., Lake, E.M.R., Scheinost, D., Constable, R.T., 2021. Resample aggregating improves the generalizability of Connectome Predictive Modelling. NeuroImage, 118044. https://doi.org/10.1016/j.neuroimage.2021.118044.

Pervaiz, U., Vidaurre, D., Woolrich, M.W., Smith, S.M., 2020. Optimising network modelling methods for fMRI. NeuroImage. 221, 116604. https://doi.org/10.1016/j.neuroimage.2020.116604.

Plaeschke, R.N., Patil, K.R., Cieslik, E.C., Nostro, A.D., Varikuti, D.P., Plachti, A., Losche, P., Hoffstaedter, F., Langner, R., Eickhoff, S.B., 2020. Age differences in predicting working memory performance from network-based functional connectivity. Cortex. 132, 441-459. https://doi.org/10.1016/j.cortex.2020.08.012.

Preusse, F., van der Meer, E., Deshpande, G., Krueger, F., Wartenburger, I., 2011. Fluid intelligence allows flexible recruitment of the parieto-frontal network in analogical reasoning. Front. Hum. Neurosci. 5, 22. https://doi.org/10.3389/fnhum.2011.00022.

Pruim, R.H.R., Mennes, M., van Rooij, Daan, Llera, A., Buitelaar, J.K., Beckmann, C.F., 2015. ICA-AROMA: A robust ICA-based strategy for removing motion artifacts from fMRI data. NeuroImage. 112, 267-277. http://dx.doi.org/10.1016/j.neuroimage.2015.02.064.

Pruim, R.H.R., Mennes, M., Buitelaar, J.K., Beckmann, C.F., 2015. Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fMRI. NeuroImage. 112, 278-287. http://dx.doi.org/10.1016/j.neuroimage.2015.02.063.

Qian, J., Hastie, T., Friedman, J., Tibshirani, R., Simon, N., 2013. Glmnet for Matlab. http://www.stanford.edu/~hastie/glmnet_matlab (last accessed 15 March 2019).

Rosa, M., Parr, A., Thompson, R., Woolgar, A., Torralva, T., Antoun, N., Manes, F., Duncan, J., 2010. Executive function and fluid intelligence after frontal lobe lesions. Brain. 133, 234-247. https://doi.org/10.1093/brain/awp269.

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.

Salimi-Khorshidi, G., Dounaud, 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.

Santarnecchi, E., Momi, D., Mencarelli, L., Plessow, F., Saxena, S., Rossi, S., Rossi, A., Mathan, S., Pascual-Leone, A., 2021. Overlapping and dissociable brain activations for fluid intelligence and executive functions. Cogn. Affect. Behav. Neurosci. 21, 327-346. https://doi.org/10.3758/s13415-021-00870-4.

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. 28, 3095-3114. https://doi.org/10.1093/cercor/bhx179.

Shah, L.M., Cramer, J.A., Ferguson, M.A., Birn, R.M., Anderson, J.S., 2016. Reliability and reproducibility of individual differences in functional connectivity acquired during task and resting state. Brain Behav. 6, e00456. https://doi.org/10.1002/brb3.456.

Smith, S.M., Beckmann, C.F., Andersson, J., Auerbach, E., Bijsterbosch, J., Dounaud, G., Duff, E., Feinberg, D.A., Griffanti, L., Harms, M.P., et al., 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., Vidaurre, D., Glasser, M., Winkler, A., McCarthy, P., Robinson, E., Chen, X., Horton, W., Jenkinson, M., Duff, E., et al., 2016. Second beta-release oft he HCP functional connectivity MegaTrawl. Available at: http://db.humanconnectome.org/megatraw (Accessed: 15 Mar 2019).

Speer, S.P.H., Smidts, A., Boksem, M.A.S., 2022. Individual differences in (dis)honesty are represented in the brain's functional connectivity at rest. NeuroImage. 246, 118761. https://doi.org/10.1016/j.neuroimage.2021.118761.

Sui, J., Jiang, R., Bustillo, J., Calhoun, V., 2020. Neuroimaging-based individualized prediction of cognition and behavior for mental disorders and health: Methods and premises. Biol. Psychol. 88, 818-828. https://doi.org/10.1016/j.biopsych.2020.02.016.

Tian, Y., Margulies, D.S., Breakspear, M., Zalesky, A., 2020. Topographic organization of the human subcortex unveiled with functional connectivity gradients. Nat. Neurosci. 23, 1421-1432. https://doi.org/10.1038/s41593-020-00711-6.

Tian, Y., Zalesky, A., 2021. Machine learning prediction of cognition from functional connectivity: Are feature weights reliable? NeuroImage. 245, 118648. https://doi.org/10.1016/j.neuroimage.2021.118648.

Turner, B.O., Paul, E.J., Miller, M.B., Barbey, A.K., 2018. Small sample sizes reduce the replicability of task-based fMRI studies. Commun. Biol. 1, 62. https://doi.org/10.1038/s42003-018-0073-z.

van den Heuvel, M.P., Sporns, O. 2013. Network hubs in the human brain. Trends Cogn. Sci. 17, 683-696. https://doi.org/10.1016/j.tics.2013.09.012.

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

Vanderwal, T., Kelly, C., Eilott, J., Mayes, L.C., Castellanos, F.X., 2015. Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. NeuroImage. 122, 222-232. https://doi.org/10.1016/j.neuroimage.2015.07.069.

Varikuti, D.P., Genon, S., Sotiras, A., Schwender, H., Hoffstaedter, F., Patil, K.R., Jockwitz, C., Caspers, S., Moebus, S., Amunts, K., 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.

Wu, J., Eickhoff, S.B., Hoffstaedter, F., Patil, K.R., Schwender, H., Yeo, B.T.T., Genon, S., 2021. A connectivity-based psychometric prediction framework for brain-behavior relationship studies. Cereb. Cortex. 31, 3732-3751. https://doi.org/10.1093/cercor/bhab044.

Yeung, A.W.K., More, S., Wu, J., Eickhoff, S.B., 2022. Reporting details of neuroimaging studies on individual traits prediction: A literature survey. NeuroImage. 119275. https://doi.org/10.1016/j.neuroimage.2022.119275.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. 67:301-320. http://doi.wiley.com/10.1111/j.1467-9868.2005.00503.x.