Imaging Neuroscience
an open access journal

Research Article

Multilayer meta-matching: Translating phenotypic prediction models from multiple datasets to small data

Pansheng Chen[a,b,c,d], Lijun An[a,b,c,d], Naren Wulan[a,b,c,d], Chen Zhang[a,b,c,d], Shaoshi Zhang[a,b,c,d,e], Leon Qi Rong Ooi[a,b,c,d,e], Ru Kong[a,b,c,d], Jianzhong Chen[a,b,c,d], Jianxiao Wu[f,g], Sidhant Chopra[h], Danilo Bzdok[i,j], Simon B. Eickhoff[f,g], Avram J. Holmes[k], B.T. Thomas Yeo[a,b,c,d,e,l]

[a]Centre for Sleep & Cognition & Centre for Translational Magnetic Resonance Research, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

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

[c]Department of Medicine, Human Potential Translational Research Programme & Institute for Digital Medicine (WisDM), Yong Loo Lin School of Medicine, National University of Singapore, Singapore

[d]N.1 Institute for Health, National University of Singapore, Singapore

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

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

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

[h]Department of Psychology, Yale University, New Haven, CT, United States

[i]Department of Biomedical Engineering, McConnell Brain Imaging Centre (BIC), Montreal Neurological Institute (MNI), Faculty of Medicine, School of Computer Science, McGill University, Montreal QC, Canada

[j]Mila – Quebec Artificial Intelligence Institute, Montreal, QC, Canada

[k]Department of Psychiatry, Brain Health Institute, Rutgers University, Piscataway, NJ, United States

[l]Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, United States

Corresponding Author: B.T. Thomas Yeo (yeoyeo02+INau@gmail.com)

ABSTRACT

Resting-state functional connectivity (RSFC) is widely used to predict phenotypic traits in individuals. Large sample sizes can significantly improve prediction accuracies. However, for studies of certain clinical populations or focused neuroscience inquiries, small-scale datasets often remain a necessity. We have previously proposed a "meta-matching" approach to translate prediction models from large datasets to predict new phenotypes in small datasets. We demonstrated a large improvement over classical kernel ridge regression (KRR) when translating models from a single source dataset (UK Biobank) to the Human Connectome Project Young Adults (HCP-YA) dataset. In the current study, we propose two meta-matching variants ("meta-matching with dataset stacking" and "multilayer meta-matching") to translate models from multiple source datasets across disparate sample sizes to predict new phenotypes in small target datasets. We evaluate both approaches by translating models trained from five source datasets (with sample sizes ranging from 862 participants to 36,834 participants) to predict phenotypes in the HCP-YA and HCP-Aging datasets. We find that multilayer meta-matching modestly outperforms meta-matching with dataset stacking. Both meta-matching variants perform better than the original "meta-matching with stacking" approach trained only on the UK Biobank. All meta-matching variants outperform classical KRR and transfer learning by a large margin. In fact, KRR is better than classical transfer learning when less than 50 participants are available for finetuning, suggesting the difficulty of classical transfer learning in the very small sample regime. The multilayer meta-matching model is publicly available at https://github.com/ThomasYeoLab/Meta_matching_models/tree/main/rs-fMRI/v2.0.

Keywords: phenotypic prediction, meta-learning, transfer learning, neuroimaging, functional connectivity

Received: 17 November 2023 Revision: 17 May 2024 Accepted: 10 June 2024 Available Online: 08 July 2024

The MIT Press

Imaging Neuroscience, Volume 2, 2024
https://doi.org/10.1162/imag_a_00233

1. INTRODUCTION

There is growing interest in harnessing neuroimaging data to predict non-neuroimaging-related phenotypes, such as fluid intelligence or clinical outcomes, of individual participants (Eickhoff & Langner, 2019; Gabrieli et al., 2015; Varoquaux & Poldrack, 2019; Woo et al., 2017). However, most brain-behavior prediction studies suffer from underpowered samples, typically involving less than a few hundred participants, leading to low reproducibility and inflated performance (Arbabshirani et al., 2017; Bzdok & Meyer-Lindenberg, 2018; Marek et al., 2022; Masouleh et al., 2019; Poldrack et al., 2020). Adequately powered sample sizes can significantly improve prediction accuracy (Chu et al., 2012; Cui & Gong, 2018; He et al., 2020; Schulz et al., 2020), so large-scale datasets, such as the UK Biobank (Miller et al., 2016; Sudlow et al., 2015), are vital for enhancing prediction performance. However, for investigations of certain clinical populations or focused neuroscience inquiries, small-scale datasets often remain the norm.

We have previously proposed a "meta-matching" approach to translate prediction models from large datasets to improve the prediction of new phenotypes in small datasets (He et al., 2022). Meta-matching is grounded in the observation that many phenotypes exhibit inter-correlations, as demonstrated by previous studies identifying a small number of factors linking brain imaging data to various non-brain-imaging traits like cognition, mental health, demographics, and other health attributes (Kebets et al., 2019; Miller et al., 2016; Smith et al., 2015; Xia et al., 2018). As a result, a phenotype X in a smaller-scale study is likely correlated with a phenotype Y present in a larger population dataset. This means that a machine-learning model trained on phenotype Y from the larger dataset might be more effectively translated to predict phenotype X in the smaller study. Meta-matching exploited these inter-phenotype correlations and was thus referred to as "meta-matching." See Section 4 for further discussion.

In our previous study (He et al., 2022), we trained a deep neural network (DNN) to predict 67 non-brain-imaging phenotypes from resting-state functional connectivity (RSFC) in the UK Biobank. The DNN was then translated using meta-matching to predict non-brain-imaging phenotypes in the Human Connectome Project Young Adult (HCP-YA) dataset, yielding large improvements over classical KRR without meta-matching. Among the different meta-matching variants, complementing "advanced meta-matching (stacking)" (which we will refer to as "meta-matching with stacking") performed the best (He et al., 2022). Stacking is a well-known ensemble learning approach (Breiman, 1996; Wolpert, 1992) and has also enjoyed utility in neuroimaging (Liem et al., 2017; Ooi et al., 2022; Rahim et al., 2017).

The original study (He et al., 2022) experimented with only one source dataset (UK Biobank). Using multiple source datasets might lead to better generalization for multiple reasons. First, prediction performance tends to increase with larger sample sizes (Chu et al., 2012; Cui & Gong, 2018; He et al., 2020; Schulz et al., 2020). Second, given acquisition, preprocessing, and demographic differences across datasets, training on multiple source datasets might yield representations that are more generalizable to a new target population (Abraham et al., 2017). Third, different datasets collect overlapping and distinct non-brain-imaging phenotypes. Since meta-matching exploits inter-phenotype correlation, training on more diverse phenotypes might lead to better performance. Here, we investigated the performance of meta-matching models trained from five source datasets—UK Biobank (Miller et al., 2016; Sudlow et al., 2015), Adolescent Brain Cognitive Development (ABCD) study (Volkow et al., 2018), Genomics Superstruct Project (GSP; Holmes et al., 2015), Healthy Brain Network (HBN; Alexander et al., 2017), and the enhanced Nathan Kline Institute-Rockland sample (eNKI-RS; Nooner et al., 2012).

One major challenge is the extreme sample size imbalances across source datasets, for example, the UK Biobank is almost 40 times larger than the HBN dataset. Therefore, there might be diminishing returns from adding smaller source datasets despite an increase in population and phenotypic diversity. A second challenge is that the available phenotypes are different across datasets, so training a single DNN to predict all phenotypes is not straightforward. Here, we considered a naive extension of the original meta-matching with stacking approach by training independent prediction model(s) in each source dataset, and then performed stacking on the outputs of the prediction models in the target dataset. We refer to this extension as "meta-matching with dataset stacking." Because meta-matching can improve the prediction of smaller datasets, we also proposed an alternative "multilayer meta-matching" approach, which gradually applied meta-matching from large source datasets (e.g., UK Biobank) to smaller source datasets (e.g., GSP, HBN, etc), to generate additional features for a final round of stacking in the target dataset.

We evaluated the proposed approaches in two target datasets—HCP-YA (Van Essen et al., 2013) and HCP-Aging (Harms et al., 2018). We found that both approaches performed better than the original "meta-matching with stacking" approach trained only on the UK Biobank. Given the close relationship between meta-matching and transfer learning, instead of performing stacking on the DNN trained on the UK Biobank (i.e., meta-matching with stacking), we also considered a standard transfer learning baseline (Weiss et al., 2016), in which the DNN was finetuned on the target dataset. Of note, meta-matching with stacking significantly outperformed the transfer learning baseline. In fact, the transfer learning baseline was worse than classical kernel ridge regression when less than 50 participants were available for finetuning, suggesting the difficulty of transfer learning in the very small sample regime. Finally, we found that multilayer meta-matching modestly outperformed meta-matching with dataset stacking.

2. METHODS

2.1. Datasets

As illustrated in Figure 1, we used five source datasets for meta-training: the UK Biobank (Miller et al., 2016; Sudlow et al., 2015), the Adolescent Brain Cognitive Development (ABCD) study (Volkow et al., 2018), the Genomics Superstruct Project (GSP; Holmes et al., 2015), the Healthy Brain Network (HBN; Alexander et al., 2017) project, and the enhanced Nathan Kline Institute-Rockland sample (eNKI-RS; Nooner et al., 2012). The models from the five datasets were then adapted for phenotypic prediction in two meta-test datasets: Human Connectome Project Young Adults (HCP-YA; Van Essen et al., 2013) and HCP-Aging (Harms et al., 2018). All data collection and analysis procedures were approved by the respective Institutional Review Boards (IRBs), including the National University of Singapore IRB for the analysis presented in this paper.

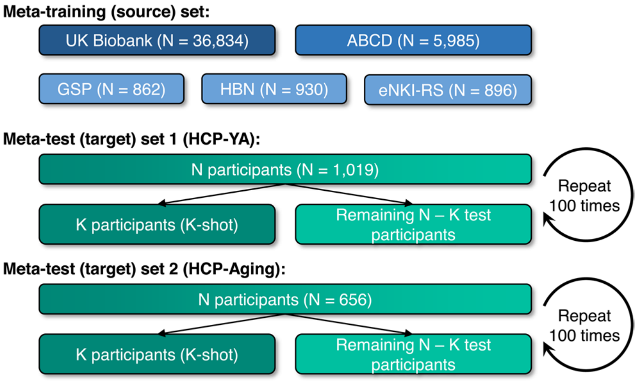
Meta-training (source) set:
UK Biobank (N = 36,834)
ABCD (N = 5,985)
GSP (N = 862)
HBN (N = 930)
eNKI-RS (N = 896)
Meta-test (target) set 1 (HCP-YA):
N participants (N = 1,019)
K participants (K-shot)
Remaining N – K test participants
Repeat 100 times
Meta-test (target) set 2 (HCP-Aging):
N participants (N = 656)
K participants (K-shot)
Remaining N – K test participants
Repeat 100 times

Fig. 1. Schematic of meta-training and meta-test sets. Datasets were assigned to meta-training set and meta-test set. Prediction models from the meta-training set were adapted to K participants from each meta-test dataset to predict target phenotypes. The adapted models were evaluated in the remaining N – K participants from the meta-test dataset. This procedure was repeated 100 times for stability. The meta-training set was differentiated into extra-large-scale (UK Biobank; dark blue), large-scale (ABCD; blue), and medium-scale (GSP, HBN, and eNKI-RS; light blue) source datasets.

The summary information of the datasets is listed in Table 1. Detailed information about the non-brain-imaging phenotypes (henceforth referred to as phenotypes) used can be found in Tables S2 to S8. The phenotypes covered a broad range of behavioral domains, ranging from cognitive performance, personality measures, lifestyle, and mental health scores. The following subsections describe each dataset and corresponding preprocessing procedures in greater detail.

We note that these datasets were opportunistically collated (e.g., by contacting potential collaborators or by downloading preprocessed data provided by the study), so the preprocessing steps varied considerably across datasets. However, we consider the heterogeneous preprocessing as a strength because the heterogeneity might help to improve (and demonstrate) generalization across preprocessing pipelines.

The phenotypes were predicted using 419 × 419 RSFC matrices, consistent with previous studies from our group (Chen et al., 2022; Kong et al., 2021; Li et al., 2022). The 419 × 419 RSFC matrices were computed using 400 cortical (Schaefer et al., 2018) and 19 subcortical parcels (Fischl et al., 2002). We note that the Schaefer parcellation is a group-level parcellation available in fsaverage, MNI, and fsLR space. For each participant, RSFC was computed as the Pearson's correlations between the average time series of each pair of brain parcels.

2.1.1. UK Biobank

The UK Biobank (UKBB) dataset is a population epidemiology study with 500,000 adults (age 40–69 years) recruited between 2006 and 2010 (Miller et al., 2016; Sudlow et al., 2015). We utilized fMRI data from 36,834 participants and 67 phenotypes (selected from a total of 3,937 phenotypes) from the UK Biobank dataset. The detailed phenotypic selection procedures followed our previous study (He et al., 2022). The sample size is slightly smaller than our previous study (He et al., 2022) because of participants voluntarily withdrawing from the UK Biobank study. More specifically, ICA-FIX pre-processed volumetric rs-fMRI time series in native participant space were downloaded from the UK Biobank (Alfaro-Almagro et al., 2018). The time series were then projected to MNI152 2-mm template space, and averaged within each cortical and each subcortical parcel. Here, the cortical parcels were based on the Schaefer parcellation in MNI152 space, while the subcortical parcels were obtained by FreeSurfer recon-all of the MNI152 template. Pearson's correlations were used to generate the 419 × 419 RSFC matrices.

Table 1. Summary information of datasets used in the current study.

Datasets		#Participants	Age range	Preprocessing notes	#Phenotypes
Meta-training datasets	UK Biobank	36,834	45-82	ICA-FIX & MNI152	67
	ABCD	5,985	9-10	GSR & fsaverage6	36
	GSP	862	18-35	GSR & fsaverage6	23
	HBN	930	5-21	GSR & fsaverage6	42
	eNKI-RS	896	6-85	ICA-AROMA & MNI152	61
Meta-test datasets	HCP-YA	1,019	22-35	ICA-FIX & fs_LR32k	35
	HCP-Aging	656	36-100+	ICA-FIX & MNI152	45

2.1.2. ABCD

The adolescent brain cognitive development (ABCD) is a dataset of children (age 9–10 years) and a diverse set of behavioral measures (Volkow et al., 2018). We considered data from 11,875 children from the ABCD 2.0.1 release. We used 36 phenotypes in total, including 16 cognitive measures, 9 personality measures, and 11 mental health measures, consistent with our previous studies (Chen et al., 2022; Ooi et al., 2022).

Details of the fMRI preprocessing can be found in previous studies (Chen et al., 2022; Ooi et al., 2022) but briefly, minimally preprocessed fMRI data (Hagler Jr et al., 2019) were further processed with the following steps: (1) removal of initial frames (number of frames removed depended on the type of scanner; Hagler Jr et al., 2019); (2) alignment with the T1 images using boundary-based registration (BBR; Greve & Fischl, 2009) with FsFast (http://surfer.nmr.mgh.harvard.edu/fswiki/FsFast); (3) respiratory pseudomotion motion filtering was performed by applying a bandstop filter of 0.31–0.43 Hz (Fair et al., 2020); (4) functional runs with BBR costs greater than 0.6 were excluded; and (5) motion correction and outlier detection: framewise displacement (FD; Jenkinson et al., 2002) and voxel-wise differentiated signal variance (DVARS; Power et al., 2012) were computed using fsl_motion_outliers. Volumes with FD > 0.3 mm or DVARS > 50, along with one volume before and two volumes after, were marked as outliers (i.e., censored frames). Uncensored segments of data containing fewer than five contiguous volumes were also censored (Gordon et al., 2016; Kong et al., 2019). BOLD runs with over half of frames censored and runs with max FD > 5 mm were removed; (6) the following nuisance covariates were regressed out of the fMRI time series: a vector of ones and linear trend, global signal, six motion correction parameters, averaged ventricular signal, averaged white matter signal, and their temporal derivatives. Regression coefficients were estimated from the non-censored volumes; (7) interpolation of censored frames with Lomb-Scargle periodogram (Power et al., 2014); (8) band-pass filtering (0.009 Hz $\leq$ f $\leq$ 0.08 Hz); (9) projection onto FreeSurfer (Fischl, 2012) fsaverage6 surface space; and (10) smoothing by a 6 mm full-width half-maximum (FWHM) kernel.

We also excluded participants who did not have at least 4 minutes for rs-fMRI and excluded participants without all 36 phenotypes, resulting in 5,985 participants. For each participant, the fMRI time series were averaged within each cortical and subcortical parcel. Here, the cortical parcels were based on the Schaefer parcellation in fsaverage space, while the subcortical parcels (from FreeSurfer recon-all) were projected from the participant's T1 native volumetric space to the participant's fMRI native volumetric space. Pearson's correlations were used to generate the 419 × 419 RSFC matrices.

2.1.3. GSP

The Brain Genomics Superstruct Project (GSP) contains fMRI and multiple behavioral measures from healthy young adults aged 18 to 35 years old (Holmes et al., 2015). We used 23 behavioral phenotypes, including cognitive and personality measures, consistent with our previous study (J. Li et al., 2019).

Details of the fMRI preprocessing can be found in previous studies (J. Li et al., 2019), but briefly, the pipeline comprised the following steps: (1) removal of the first four frames; (2) slice time correction with FSL (Jenkinson et al., 2012; Smith et al., 2004) package; and (3) motion correction and outlier detection: FD and DVARS were estimated using fsl_motion_outliers. Volumes with FD > 0.2 mm or DVARS > 50 were marked as outliers (censored frames). One frame before and two frames after these volumes were flagged as censored frames. Uncensored segments of data lasting fewer than five contiguous volumes were also labeled as censored frames (Gordon et al., 2016). BOLD runs with more than half of the volumes labeled as censored frames were removed; (4) alignment with structural image using boundary-based registration with FsFast (Greve & Fischl, 2009); (5) regress the following nuisance regressors: a vector of ones and linear trend, six motion correction parameters, averaged white matter signal, averaged ventricular signal, mean whole-brain signal, and their temporal derivatives. Regression coefficients were estimated from the non-censored volumes; (6) interpolation of censored frames with Lomb-Scargle periodogram; (7) band-pass filtering (0.009 Hz ≤ f ≤ 0.08 Hz); (8) projection onto the FreeSurfer fsaverage6 surface space; and (9) smoothing with 6 mm FWHM and down-sampling to fsaverage5 surface space.

We also removed participants without full 23 phenotypes, yielding 862 participants. For each participant, the fMRI time series were averaged within each cortical and subcortical parcel. Here, the cortical parcels were based on the Schaefer parcellation in fsaverage space, while the subcortical parcels (from FreeSurfer recon-all) were projected from the participant's T1 native volumetric space to the participant's fMRI native volumetric space. Pearson's correlations were used to generate the 419 × 419 RSFC matrices.

2.1.4. HBN

The Healthy Brain Network (HBN) contains New York area participants (age 5–21 years) with brain imaging, psychiatric, behavioral, cognitive, and lifestyle information (Alexander et al., 2017). We downloaded data from 2,196 participants (HBN release 1–7). We manually selected commonly used cognitive performance scores and behavioral scores with less than 10% of missing values, resulting in 42 phenotypes.

Resting-state fMRI data were pre-processed with the following steps: (1) removal of the first 8 frames; (2) slice time correction; (3) motion correction and outlier detection: frames with FD > 0.3 mm or DVARS > 60 were flagged as censored frames. 1 frame before and 2 frames after these volumes were flagged as censored frames. Uncensored segments of data lasting fewer than five contiguous frames were also labeled as censored frames. BOLD runs with over half of the frames censored and runs with max FD > 5 mm were removed; (4) correcting for spatial distortion caused by susceptibility-induced off-resonance field; (5) alignment with structural image using boundary-based registration; (6) nuisance regression: regressed out a vector of ones and linear trend, global signal, six motion correction parameters, averaged ventricular signal, averaged white matter signal, and their temporal derivatives. Regression coefficients were estimated from the non-censored volumes; (7) band-pass filtering (0.009 Hz ≤ f ≤ 0.08 Hz); (8) interpolation of censored frames with Lomb-Scargle periodogram; (9) projection onto the FreeSurfer fsaverage6 surface space; and (10) smoothing with 2 mm FWHM and down-sampling to fsaverage5 surface space.

We excluded individuals who did not have at least 4 minutes of uncensored rs-fMRI data and removed participants with no relevant phenotypes, resulting in 930 participants. For each participant, the fMRI time series were averaged within each cortical and subcortical parcel. Here, the cortical parcels were based on the Schaefer parcellation in fsaverage space, while the subcortical parcels (from FreeSurfer recon-all) were projected from the participant's T1 native volumetric space to the participant's fMRI native volumetric space. Pearson's correlations were used to generate the 419 × 419 RSFC matrices.

2.1.5. eNKI-RS

The enhanced Nathan Kline Institute-Rockland Sample (eNKI-RS) is a community sample of over 1,000 participants (age 6–85 years), with measures including various physiological and psychological assessments, genetic information, and neuroimaging data (Nooner et al., 2012). We manually selected commonly used cognitive performance measures and behavioral scores with less than 10% of missing value, yielding 61 phenotypes and 896 participants with at least one phenotype.

Details of the fMRI preprocessing can be found in our previous study (Wu et al., 2022), but briefly, eNKI-RS data were pre-processed with fMRIprep (Esteban et al., 2019) with default configuration and additional ICA-AROMA denoising (Pruim, Mennes, Buitelaar, et al., 2015; Pruim, Mennes, van Rooij, et al., 2015). Additional nuisance regression was then performed with regressors corresponding to 24 motion parameters, white matter signal, CSF signal, and their temporal derivatives (Wu et al., 2022). The pre-processed fMRI data in MNI152 space were used to compute 419 × 419 RSFC matrices. Here, the cortical parcels were based on the Schaefer parcellation in MNI152 space, while the subcortical parcels were obtained by FreeSurfer recon-all of the MNI152 template.

2.1.6. HCP-YA

The Human Connectome Project (HCP Young Adult, HCP-YA) contains brain imaging data and phenotypes from healthy young adults (age 22–35 years) (Van Essen et al., 2013). We used 35 phenotypes across cognition, personality, and emotion, consistent with our previous study (He et al., 2022). There are 1,019 participants with all 35 phenotypes in the end.

For the RSFC data, we used ICA-FIX MSMALL time series in the grayordinate (combined surface and subcortical volumetric) fsLR_32k space (Glasser et al., 2013). The time series were averaged within each cortical and subcortical parcel to calculate 419 × 419 RSFC matrices. Here, the cortical parcels were based on the Schaefer parcellation in fsLR space, while the subcortical parcels were defined by the HCP preprocessing pipeline based on FreeSurfer (Glasser et al., 2013).

2.1.7. HCP-Aging

The Human Connectome Project Aging (HCP-Aging) study enrolls 1,500+ healthy adults (age 36–100+ years) (Harms et al., 2018). We manually selected commonly used behavioral measures, resulting in 45 phenotypes and 656 participants with at least one phenotype. The resting-fMRI data after ICA-FIX denoising in MNI152 space were used, following our previous study (Wu et al., 2022). Nuisance regression was then implemented, controlling for 24 motion parameters, white matter signal, CSF signal, and their temporal derivatives (Wu et al., 2022). The time series were averaged within each cortical and subcortical parcel to calculate 419 × 419 RSFC matrices. Here, the cortical parcels were based on the Schaefer parcellation in MNI152 space, while the subcortical parcels were obtained by FreeSurfer recon-all of the MNI152 template.

2.2. Data split overview

We split the datasets into a meta-training (source) set and a meta-test (target) set, as shown in Figure 1. For each meta-training dataset, we randomly divided the participants into training and validation sets comprising 80% and 20% of the participants respectively. The training and validation sets are used to train and tune the hyper-parameters of one or more “base-learners” to predict corresponding source phenotypes from the meta-training dataset. We note that the splits into training and validation sets were completely random, and no attempt was made to match the demographics (e.g., age and sex) between training and validation sets. Matching demographics between training and validation sets might potentially improve the prediction in the validation sets, but it is unclear whether this would be helpful for the meta-test set, whose demographics might differ from the meta-training sets. In fact, one might even speculate that demographic differences between training and validation sets could help the base-learners to be more robust to demographic differences between meta-training and meta-test datasets.

For each meta-test dataset, there are target phenotypes we want to predict from RSFC. For cross-dataset prediction, we trained a “meta-learner” using K participants in the meta-test dataset (i.e., where K = 10, 20, 50, 100, 200) with observed meta-test phenotypes, which is a setting known as “K-shot learning” (Kadam & Vaidya, 2020). The meta-learner exploits the relationship between source and target phenotypes via the previously trained base-learners from the meta-training datasets, thus transferring knowledge from the meta-training datasets to the meta-test dataset. Finally, we evaluated the prediction performance of meta-test phenotypes on the remaining N – K meta-test participants, using Pearson’s correlation and predictive coefficient of determinant (COD) as metrics.

2.3. Prediction approaches

Across all approaches, we vectorized the lower triangular entries of each 419 × 419 RSFC matrix into a feature vector (i.e., 87,571 × 1 vector) to predict phenotypic measures. We note that certain datasets were processed with global signal regression (GSR), while others were processed with ICA-FIX (Table 1). It is well known that GSR centers the distribution of RSFC values at zero (Murphy et al., 2009), which is not the case for ICA-FIX. Therefore, for all cross-dataset algorithms (i.e., all algorithms except kernel ridge regression), we normalized the RSFC vector for each participant independently, by subtracting the mean and then dividing by the L2-norm of the 87,571 × 1 FC vector. Although we did not perform this normalization for classical kernel ridge regression (Section 2.3.1), we note that this normalization has no effect on kernel ridge regression. The reason is that we used the correlation metric to compute the kernel similarity (Section 2.3.1), so Pearson’s correlation between two normalized RSFC matrices will be the same as Pearson’s correlation between two unnormalized RSFC matrices.

Following our previous study (He et al., 2022), statistical difference between algorithms was evaluated using a bootstrapping approach (more details in Supplementary Methods S3). Multiple comparisons were corrected using a false discovery rate (FDR) of $q < 0.05$. FDR was applied to all K-shots, across all pairs of algorithms and both evaluation metrics (Pearson’s correlation and COD).

2.3.1. Baseline 1: Classical KRR

We choose kernel ridge regression (KRR; Fig. 2A) as a baseline algorithm that does not utilize meta-training on the meta-training set. KRR has been shown to be a highly competitive algorithm for MRI prediction of phenotypic measures (He et al., 2020; Kong et al., 2023; Ooi et al., 2022). Consistent with our previous studies, the kernel similarity between participants was defined based on similarity (Pearson’s correlation) between the lower triangular portions of the RSFC matrices. More specifically, the procedure is as follows. Suppose the meta-test dataset has N participants in total. For each target phenotype in the meta-test dataset, we trained a KRR model and tuned the hyper-parameter λ (L2 regularization weight) with 5-fold cross-validation, using K random participants with observed target phenotypes (i.e., K-shot). The optimal λ was then used to train a final KRR model using all

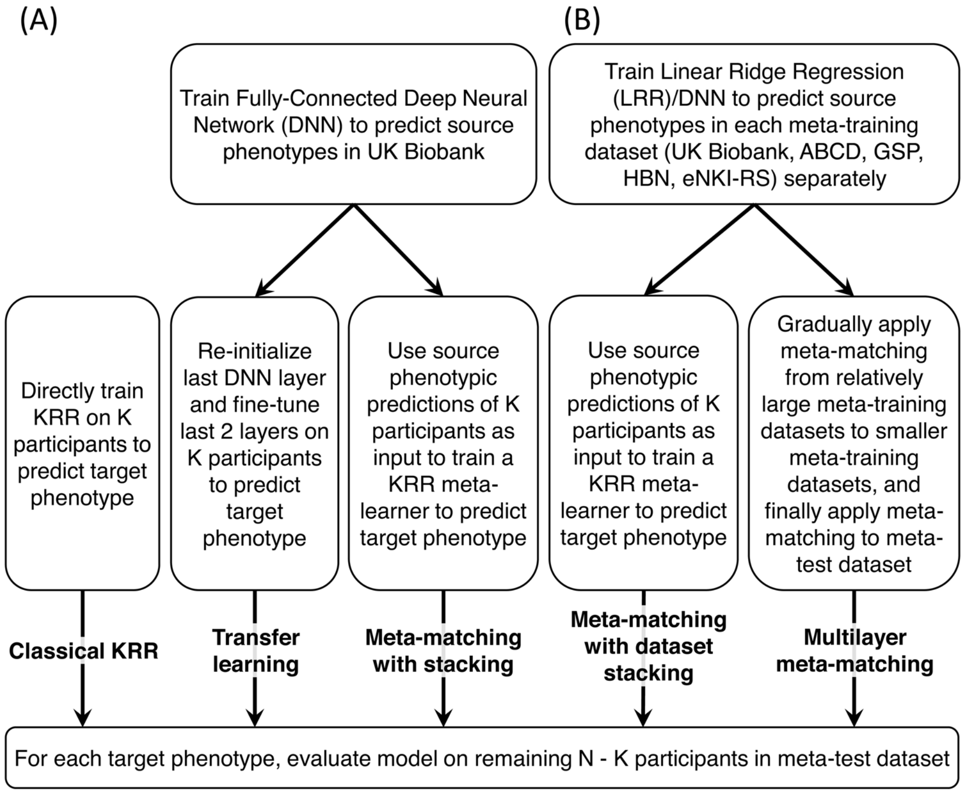
(A)
(B)
Train Fully-Connected Deep Neural Network (DNN) to predict source phenotypes in UK Biobank
Train Linear Ridge Regression (LRR)/DNN to predict source phenotypes in each meta-training dataset (UK Biobank, ABCD, GSP, HBN, eNKI-RS) separately
Directly train KRR on K participants to predict target phenotype
Re-initialize last DNN layer and fine-tune last 2 layers on K participants to predict target phenotype
Use source phenotypic predictions of K participants as input to train a KRR meta-learner to predict target phenotype
Use source phenotypic predictions of K participants as input to train a KRR meta-learner to predict target phenotype
Gradually apply meta-matching from relatively large meta-training datasets to smaller meta-training datasets, and finally apply meta-matching to meta-test dataset
Classical KRR
Transfer learning
Meta-matching with stacking
Meta-matching with dataset stacking
Multilayer meta-matching
For each target phenotype, evaluate model on remaining N - K participants in meta-test dataset

Fig. 2. Schematic of different approaches. (A) Schematic of three baselines: classical kernel ridge regression (KRR), transfer learning, and meta-matching with stacking from our previous study (He et al., 2022). (B) Schematic of two proposed approaches: meta-matching with dataset stacking and multilayer meta-matching. Observe the large sample imbalance in the meta-training set with the smallest source dataset comprising 862 participants and the largest source dataset comprising 36,834 participants.

K participants. We then evaluated the model performance on the remaining N – K participants using Pearson's correlation and COD. The procedure was repeated 100 times with a different random set of K participants. The evaluation metrics were averaged across the 100 repetitions to ensure the robustness of the results.

2.3.2. Baseline 2: Transfer learning

As a second baseline, we consider transfer learning (Weiss et al., 2016). As illustrated in Figure 2A, we pre-trained a deep neural network (DNN) in the UK Biobank to simultaneously predict 67 source phenotypes from RSFC (maximum training epochs = 100). The DNN is a simple fully-connected feedforward neural network (also known as a multi-layer perceptron) with 67 output nodes. Rectifying linear units (ReLU) were used as activation functions for all hidden layers. As mentioned in Section 2.2, 80% of the data was used for training and 20% was used for tuning DNN hyper-parameters. The hyper-parameters (e.g., number of layers, number of nodes, learning rate, dropout rate, etc.) were tuned using the Optuna package (Akiba et al., 2019). As a final step, we used 80% of the data for training with the optimal hyperparameters, and the remaining 20% of data for early stopping to reduce the possibility of overfitting. Detailed information about DNN hyper-parameters is found in Supplementary Methods S1.

The pre-trained DNN was then translated using K meta-test participants to predict a target phenotype. Because we are predicting different phenotypes in the meta-test dataset, for a given target phenotype, the last layer of the pre-trained DNN was re-initialized from scratch, and the last two layers of the DNN were then fine-tuned on K random participants with observed target phenotypes (i.e., K-shot). An optimal fixed learning rate was obtained by 5-fold cross-validation and grid search of the K participants. The optimal learning rate was then used to perform fine-tune a final model using all K participants. For both the 5-fold cross-validation and the final round of fine-tuning, the maximum fine-tuning epochs was set to be 10 with 80% of K participants used for training and 20% used to evaluate validation loss for early stopping, to reduce the possibility of overfitting. This final trained model was evaluated in the remaining N – K participants.

2.3.3. Baseline 3: Meta-matching with stacking

The third baseline is the "meta-matching with stacking" algorithm (Fig. 2A) from the original meta-matching study (He et al., 2022). The original study proposed several meta-matching algorithms. Here, we used the stacking approach because it exhibited the best prediction performance in the original study.

Similar to transfer learning, the meta-matching with stacking approach utilized the same pre-trained DNN from the UK Biobank (see Section 2.3.2). To adapt the DNN to the meta-test dataset, the DNN was applied to the RSFC of the K participants, yielding 67 predictions per participant. The 67 predictions were then used as features to train a KRR model for predicting the target phenotype using the K participants (i.e., stacking; Wolpert, 1992).

The KRR model utilized the correlation kernel, and the KRR hyperparameter λ was tuned using grid search and 5-fold cross-validation on the K participants. The optimal λ was then used to train a final KRR model using all K participants. The prediction performances were evaluated on the remaining N – K participants using Pearson's correlation and COD as metrics. This procedure was repeated 100 times with a different random sample of K participants.

It is worthwhile highlighting a deviation from the original meta-matching with stacking implementation (He et al., 2022). The original implementation utilized K features for stacking when $K < 67$. Here, we decided to simply use all 67 features because experimentation after the publication of our previous study (not shown) suggested the constraint was unnecessary.

2.3.4. Meta-matching with dataset stacking

A naive approach to extending meta-matching with stacking to multiple datasets is to train independent prediction model(s) in each meta-training (source) dataset and then "stack" the prediction models based on K participants in the meta-test dataset. We refer to this approach as meta-matching with dataset stacking (Fig. 2B).

For the UK Biobank, we trained a DNN model to predict 67 phenotypes, as well as 67 Linear Ridge Regression (LRR) models to predict 67 phenotypes, to improve prediction performance via ensemble learning (Dietterich, 2000), yielding $67 \times 2 = 138$ predictions. We note that the original version of our manuscript utilized KRR instead of LRR. However, KRR requires computing the similarity between a test individual's FC with the training individuals' FC. The implication is that a researcher applying meta-matching to their own small dataset would require access to the original FC data from the meta-training set, which is undesirable.

We note that the DNN model is identical to the pre-trained DNN from the transfer learning baseline. The remaining four datasets (ABCD, GSP, HBN, eNKI-RS) were a lot smaller than the UK Biobank, so instead of training a DNN, we simply trained an LRR model for each source phenotype and each meta-training dataset. The regularization hyperparameter λ was tuned using grid search and 5-fold cross-validation on the full dataset, and the optimal λ was then used to train a final LRR model using the full dataset. The LRR and DNN models were applied to the RSFC of the K participants (of the meta-test dataset), yielding a total of $67 \times 2 + 36 + 23 + 42 + 61 = 296$ phenotypic predictions for each participant.

Similar to the meta-matching with stacking approach (Section 2.3.3), the predictions were then used as features to train a KRR model for predicting the target phenotype using the K participants (i.e., stacking). The KRR model utilized the correlation kernel, and the KRR hyperparameter λ was tuned using grid search and 5-fold cross-validation on the K participants. The optimal λ was then used to train a final KRR model using all K participants.

The prediction performances were evaluated on the remaining N – K participants using Pearson's correlation and COD as metrics. This procedure was repeated 100 times with a different random sample of K participants.

2.3.5. Multilayer meta-matching

As an alternative to "meta-matching with dataset stacking," we made use of the fact "meta-matching with stacking" can improve the prediction of smaller datasets. Therefore, "multilayer meta-matching" (Fig. 2B) gradually applied meta-matching with stacking from relatively large source datasets (e.g., UK Biobank) to smaller datasets (e.g., GSP, HBN, etc), to generate additional features for a final round of stacking using the K participants from the meta-test dataset.

In the current study, we instantiated multilayer meta-matching by dividing the meta-training datasets into three groups: extra-large source dataset (comprising only UK Biobank in the current study), large source datasets (comprising only ABCD in the current study), and medium source datasets (comprising GSP, HBN, and eNKI-RS in the current study). Multilayer meta-matching proceeds as follows (Fig. 3).

In the case of the extra-large dataset (UK Biobank), we have previously trained DNN and LRR models to predict 67 phenotypes (Section 2.3.4). The same two models were applied to the K meta-test dataset participants, yielding $67 \times 2 = 134$ phenotypic predictions, which will be concatenated with the predictions from the other models (below) for stacking.

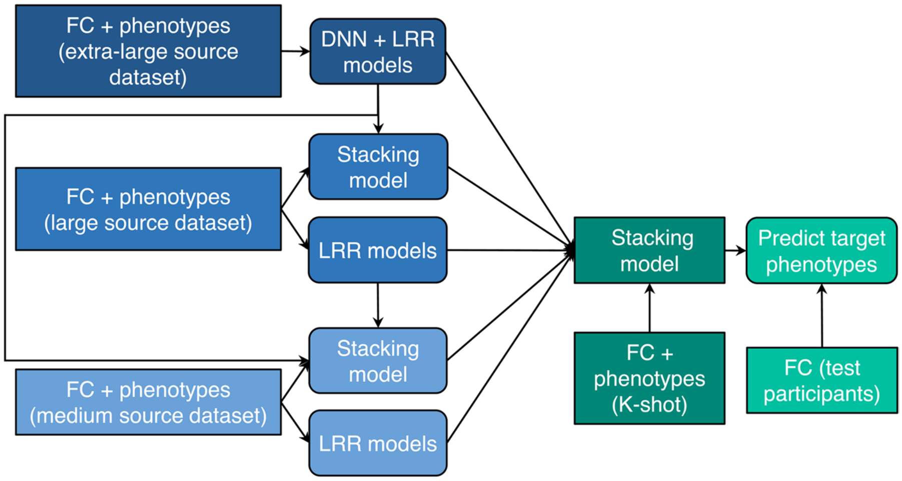
FC + phenotypes (extra-large source dataset)
DNN + LRR models
Stacking model
FC + phenotypes (large source dataset)
LRR models
Stacking model
FC + phenotypes (medium source dataset)
LRR models
Stacking model
Predict target phenotypes
FC + phenotypes (K-shot)
FC (test participants)

Fig. 3. Multilayer meta-matching. We divided source datasets into extra-large (UK Biobank), large (ABCD), and medium (GSP/HBN/eNKI-RS) source datasets. Multi-layer meta-matching gradually applied meta-matching with stacking from relatively large source datasets (e.g., UK Biobank) to smaller datasets (e.g., HCP), to generate additional features for a final round of stacking using the K participants from the meta-test dataset.

In the case of the large dataset (ABCD), we have previously trained an LRR model to predict 36 phenotypes in the ABCD dataset (Section 2.3.4). The same model was applied to the K meta-test dataset participants, yielding 36 predictions. Furthermore, the DNN and LRR models from the extra-large dataset (UK Biobank) were also combined to predict the 36 ABCD phenotypes via the meta-matching with stacking procedure (He et al., 2022). The resulting stacking model was applied to the K meta-test dataset participants, yielding 36 predictions. Therefore, models from the ABCD dataset yielded a total of 36 × 2 = 72 phenotypic predictions for each of the K meta-test dataset participants, which will be concatenated with the 134 predictions from the UK Biobank (above) and predictions from the other models (below) for stacking.

Finally, in the case of the medium source dataset (GSP, HBN, or eNKI-RS), let us use the GSP dataset, which had 23 phenotypes, as an example. First, we have previously trained an LRR model to predict 23 phenotypes in the GSP dataset (Section 2.3.4). The same model was applied to the K meta-test dataset participants, yielding 23 predictions. Second, the DNN and LRR models from the extra-large dataset (UK Biobank), as well as the LRR models from the large dataset (ABCD) were also combined to predict the 23 GSP phenotypes via the meta-matching with stacking procedure (He et al., 2022). The resulting stacking model was applied to the K meta-test dataset participants, yielding 23 predictions. Therefore, in total, the GSP dataset contributed 23 × 2 = 46 phenotypic predictions in each of the K meta-test dataset participants. Similarly, the HBN and eNKI-RS datasets contributed 42 × 2 = 84 and 61 × 2 = 122 phenotypic predictions.

Finally, all the phenotypic predictions (134 + 72 + 46 + 84 + 122 = 458) were concatenated and used to train a KRR model on the K meta-test dataset participants (i.e., stacking). Once again, the KRR model utilized the correlation kernel and the KRR hyperparameter λ was tuned using grid search and 5-fold cross-validation on the K participants. The optimal λ was then used to train a final KRR model using all K participants.

The prediction performances were evaluated on the remaining N – K participants using Pearson's correlation and COD as metrics. This procedure was repeated 100 times with a different random sample of K participants.

It is worth noting that the number of features used by the final stacking procedure was 458 in multilayer meta-matching, compared with 296 features in meta-matching with dataset stacking. More specifically, the number of features directly generated the UK Biobank models is 134 for both approaches. In the case of the large and medium-sized datasets, the number of features are doubled from 36 (ABCD), 23 (GSP), 42 (HBN), and 61 (eNKI-RS) to 72, 46, 84, and 122 respectively. We note that 458 features are still of much lower dimensionality than the raw FC matrices.

2.4. Feature importance based on the Haufe transform

Here, we are adapting models pre-trained with different phenotypes to predict new phenotypes in a meta-test dataset with potentially different demographics from the source datasets. A potential concern is that the interpretation of these adapted models (meta-matching or transfer learning models) might be "tainted" by this pre-training. To quantify this bias that might arise from pre-training, we needed to define a ground truth. Here, we assumed that the full HCP-YA and HCP-Aging datasets are sufficiently large, so that a model trained with the full dataset to predict a particular meta-test phenotype will not be biased by the pre-training (since there is no pre-training).

However, what predictive model should be used in this analysis? Since KRR has been shown to be a highly competitive algorithm for MRI prediction of phenotypic measures (He et al., 2020; Kong et al., 2023; Ooi et al., 2022), we decided to train a KRR model on the full HCP-YA (or HCP-Aging) dataset and then applied the Haufe transform to the KRR model to generate pseudo ground truth feature importance weights. The Haufe transform involved computing the covariance between each FC edge and the phenotypic prediction across all participants in the meta-test set (Chen et al., 2022; Haufe et al., 2014). The result is a feature importance value for each RSFC edge. A positive (or negative) feature importance value indicates that higher RSFC for the edge was associated with the prediction model predicting greater (or lower) value for the phenotype.

We chose the Haufe transform because it has been shown to be optimal for linear models (Haufe et al., 2014) and KRR can be reformulated as a linear model of our use of the linear kernel. Furthermore, previous studies have shown that the Haufe transform led to highly-reliable feature importance weights, which are similar across different predictive models (Chen et al., 2023; Tian & Zalesky, 2021), suggesting that our pseudo ground truth will not be sensitive to our choice of KRR as the pseudo ground truth predictive model.

We compared the Haufe transform of the pseudo ground truth with the Haufe transform for each approach (classical KRR, meta-matching, and transfer-learning) for the K = 100 scenario, which involved computing the covariance between each FC edge and the phenotypic prediction across the K participants (Chen et al., 2022; Haufe et al., 2014). We then correlated the resulting feature importance values of each approach with the pseudo ground truth. We repeated this procedure 100 times, and averaged the correlations with the pseudo ground truth across 100 repetitions. Given the relatively small sample (K = 100), we did not expect that meta-matching will yield very similar feature importance values as the pseudo ground truth. However, we hoped that the deviation between our meta-matching models and the pseudo ground truth is not worse than classical KRR (trained on 100 participants).

3. RESULTS

3.1. Meta-matching with stacking outperformed classical KRR and transfer learning

Figure 4A and B show the prediction accuracy (Pearson's correlation coefficient) of various approaches in the HCP-YA and HCP-Aging meta-test datasets respectively. Results were averaged across 35 HCP-YA (or 45 HCP-Aging) phenotypes. The horizontal axis is the number of few-shot participants (K, where K = 10, 20, 50, 100, 200). The vertical axis is Pearson's correlation of phenotypic prediction. Boxplots represent variability across the 100 repetitions of sampling K participants (i.e., K-shot). Figure 5 shows results for COD. Bootstrapping results are shown in Figures S1 and S2, while p values are reported in Tables S9 and S10. All bolded p values (Tables S9 and S10) survived an FDR of $q < 0.05$.

Consistent with our previous study (He et al., 2022), meta-matching with stacking outperformed classical KRR in the HCP-YA dataset (Figs. 4A and 5A; Tables S9). Here, we extended the previous results by showing consistent improvements over KRR in the HCP-Aging dataset. More specifically, in the case of the HCP-YA dataset and $K > 10$ (Table S9), meta-matching with stacking was statistically better than classical KRR with the largest $p < 0.01$ across both evaluation metrics (Pearson's correlation and COD). In the case of HCP-Aging and $K > 10$ (Table S10), meta-matching with stacking was statistically better than classical KRR with the largest $p < 0.002$ across both evaluation metrics.

Furthermore, meta-matching with stacking also outperformed transfer learning across both datasets (Figs. 4A and 5A). In the case of the HCP-YA dataset and $K \geq 10$ (Table S9), meta-matching with stacking was statistically better than transfer learning with p values <0.02 across both evaluation metrics (Pearson's correlation and COD). In the case of HCP-Aging and $K \geq 10$ (Table S10), meta-matching with stacking was statistically better than transfer learning with the largest $p < 0.001$ across both evaluation metrics.

Interestingly, transfer learning performed consistently worse than classical KRR for $K < 50$, especially for the COD metric (Figs. 4A and 5A).

3.2. Improvement from additional meta-training source datasets

By including additional meta-training datasets, meta-matching with dataset stacking and multilayer meta-matching were numerically better than meta-matching with stacking (which only utilized the UK Biobank) for almost all values of K (Figs. 4 and 5).

In the case of the HCP-YA dataset and $K > 100$ (Table S9), meta-matching with dataset stacking was statistically better than meta-matching with stacking with the largest $p < 0.001$ across both evaluation metrics (Pearson's correlation and COD). In the case of the HCP-Aging and $K > 20$ (Table S10), meta-matching with dataset stacking was statistically better than meta-matching

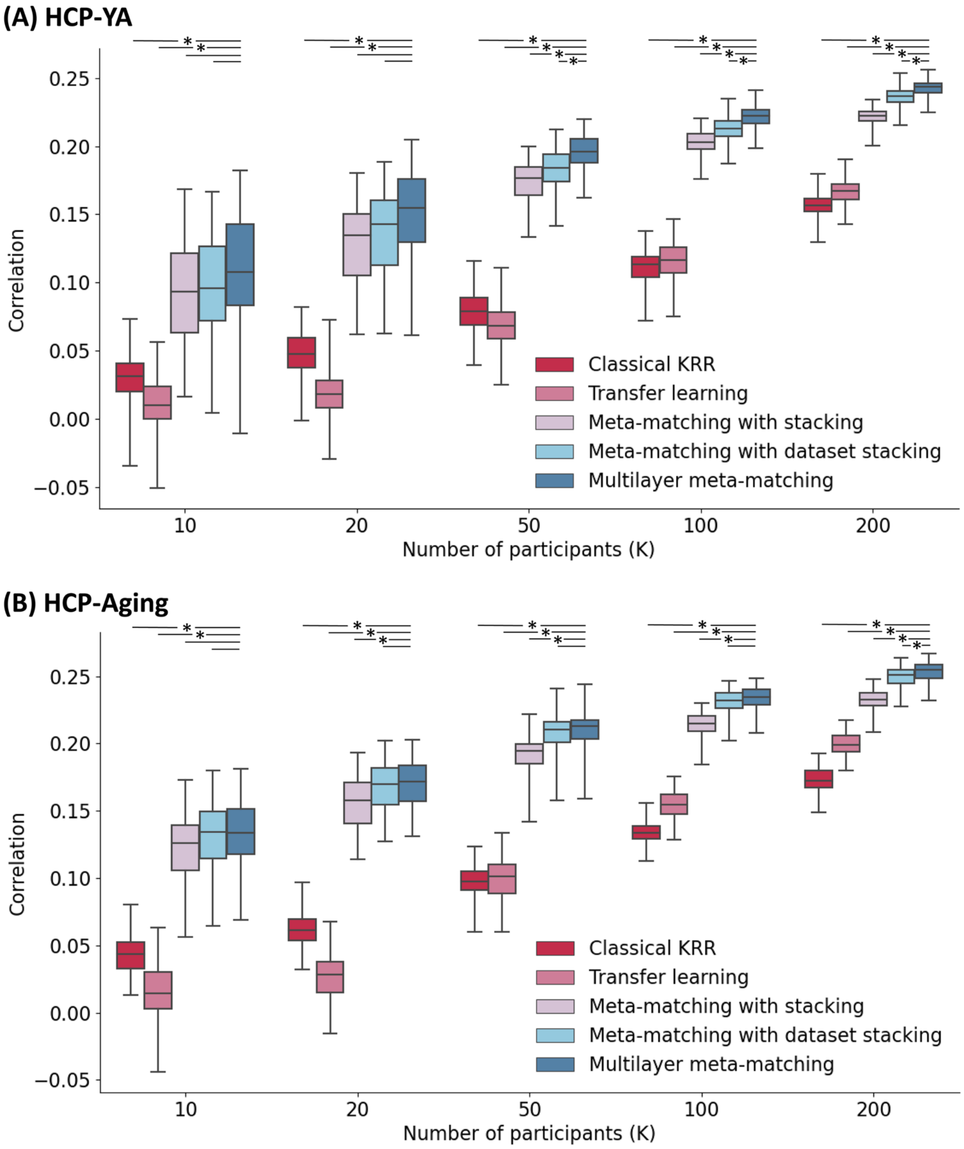

Fig. 4. Prediction performance (Pearson's correlation) in the HCP-YA and HCP-Aging datasets. (A) Phenotypic prediction performance in terms of Pearson's correlation (averaged across 35 meta-test phenotypes) in the HCP-YA dataset. Horizontal axis is the number of participants in the HCP-YA dataset used to adapt the models trained from the meta-training source datasets. Boxplots represent variability across 100 repetitions of sampling K participants. The bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Whiskers correspond to 1.5 times the interquartile range. "*" indicates statistical significance between multilayer meta-matching and other approaches (after correction for multiple comparisons with FDR $q < 0.05$). Dash line without "*" indicates a lack of significance. (B) Same plot as panel A except that the analyses were performed in the HCP-Aging dataset. The full set of p values are reported in Tables S9 and S10.

with stacking with the largest $p < 0.001$ across both evaluation metrics.

On the other hand, in the case of the HCP-YA dataset and $K > 20$ (Table S9), multilayer meta-matching was statistically better than meta-matching with stacking with the largest $p < 0.03$ across both evaluation metrics. In the case of the HCP-Aging and $K > 20$ (Table S10), multilayer meta-matching was statistically better than meta-matching with stacking with the largest $p < 0.001$ across both evaluation metrics.

We observe that the p values for multilayer meta-matching were generally stronger (i.e., smaller) than meta-matching with dataset stacking and will directly compare the two meta-matching variants in the next section.

3.3. Multilayer meta-matching modestly outperformed meta-matching with dataset stacking

Multi-layer meta-matching was numerically better than meta-matching with dataset stacking for almost all values

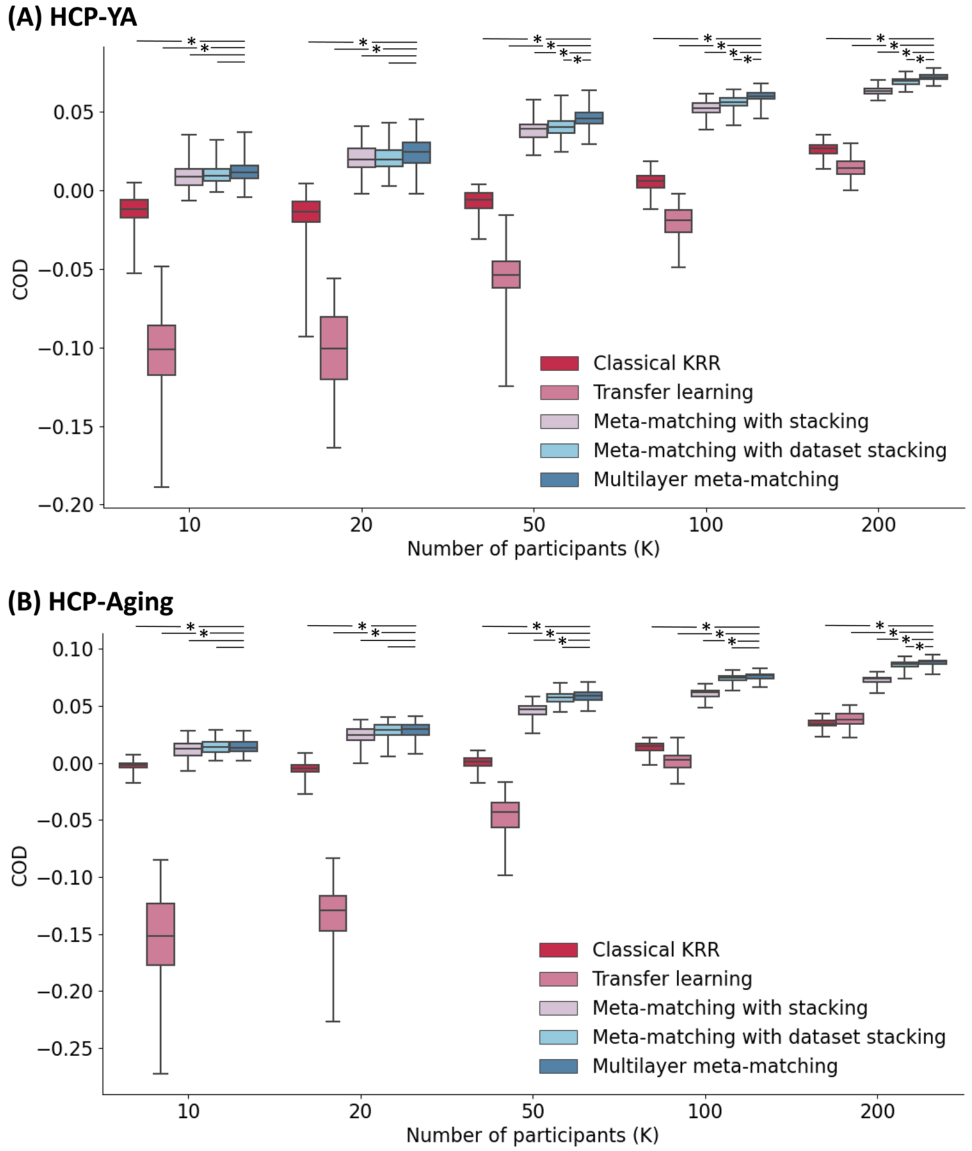

Fig. 5. Prediction performance (COD) in the HCP-YA and HCP-Aging datasets. (A) Phenotypic prediction performance in terms of COD (averaged across 35 meta-test phenotypes) in the HCP-YA meta-test set. Horizontal axis is the number of participants in the HCP-YA dataset used to adapt the models trained from the meta-training source datasets. Boxplots represent variability across 100 repetitions of sampling K participants. The bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Whiskers correspond to 1.5 times the interquartile range. "*" indicates statistical significance between multilayer meta-matching and other approaches (after correction for multiple comparisons with FDR $q < 0.05$). Dash line without "*" indicates a lack of significance. (B) Same plot as panel A, except that the analyses were performed in the HCP-Aging dataset. The full set of p values are reported in Tables S9 and S10.

of K. This improvement was significant for larger values of K. In the case of the HCP-YA dataset and $K > 20$ (Table S9), multi-layer meta-matching was statistically better than meta-matching with dataset stacking with the largest $p < 0.01$ for both evaluation metrics (correlation and COD). For HCP-Aging, multilayer meta-matching was statistically better than meta-matching with dataset stacking for $K = 200$ for both evaluation metrics ($p < 0.03$; Table S10). Overall, the results suggest that multilayer meta-matching was modestly more effective than meta-matching with dataset stacking.

In the introduction, we suggested that since meta-matching with stacking (He et al., 2022) improved prediction significantly in small datasets, by applying the original meta-matching with stacking to the smaller datasets, the resulting features might be more helpful for the final stacking procedure, compared with just training KRR models in the smaller datasets directly. To test this hypothesis, we performed 5-fold cross-validation on three medium datasets (i.e., GSP, HBN, and eNKI-RS), to predict phenotypes using classical KRR. We note that the KRR models are used by the meta-matching with dataset stacking approach in the meta-test set. We also performed 5-fold cross-validation on the medium datasets using meta-matching with stacking based on the DNN and LRR models from the UK Biobank and the LRR models from ABCD. We find that meta-matching with stacking yielded better prediction performance than the KRR models in two of the three datasets (Table 2), thus providing some support for our hypothesis.

In the same analysis, we found that phenotypic predictions from classical KRR and meta-matching with stacking are not strongly correlated (r = 0.4 to 0.6; Table 2). We remind the reader that the predictions of these models are used as features for stacking in the meta-test set. Since ensembles of diverse machine-learning models lead to better prediction performance (Kuncheva & Whitaker, 2003), we speculate that the more diverse predictions utilized by multilayer meta-matching might lead to better prediction performance than meta-matching with dataset stacking.

3.4. Different improvements on different phenotypes by multilayer meta-matching

Figure 6 shows the numerical improvement in prediction performance (Pearson's correlation) of multilayer meta-matching over the other approaches across different phenotypes. The corresponding plot for COD is shown in Figure S3. Table 3 shows the percentage of phenotypes in which multilayer meta-matching exhibits numerical improvement in prediction performance (Pearson's correlation) over other approaches. COD results are shown in Table S11. Compared with classical KRR, transfer learning, and meta-matching with stacking, we found that multilayer meta-matching exhibited numerical improvement for a vast majority of the phenotypes (Table 3; Table S11).

Figure 7 illustrates the 100-shot prediction performance (Pearson's correlation coefficient) of three example meta-test phenotypes across all approaches in the HCP-YA (Fig. 7A) and HCP-Aging (Fig. 7B) datasets. For three illustrated HCP-YA phenotypes ("Delay Discounting," "Manual Dexterity," "Arithmetic"), multilayer meta-matching exhibited numerically the best results. On the other hand, among the three illustrated HCP-Aging phenotypes, multilayer meta-matching was numerically worse than meta-matching with stacking and meta-matching with dataset stacking in the case of "Walking Endurance," but was numerically the best for "MOCA score" and "Perceived Hostility."

Tables S12 to S15 report the numerical improvement of multilayer meta-matching over other baselines for all HCP-YA and HCP-Aging phenotypes (in the 100-shot scenario). In the HCP-YA dataset, multilayer meta-matching was numerically better than classical KRR for many cognitive measures, but also some non-cognitive measures, such as strength and endurance (Tables S12 and S13). This was also the case for the HCP-Aging dataset, and interestingly the phenotype enjoying the greatest improvement was strength (Tables S14 and S15).

3.5. Feature importance using the Haufe transform

As shown in Figure 8, across both HCP-YA and HCP-Aging datasets, feature importance values of multilayer meta-matching and classical KRR were equally similar to the pseudo ground truth feature importance values. On the other hand, feature importance values from transfer learning were the most different from the pseudo ground truth. If we only focused on the transfer learning and meta-matching models, we observed a trend in increasing agreement with pseudo ground truth, which parallels the prediction accuracy increase from transfer learning to meta-matching with stacking to meta-matching with dataset stacking and then to multilayer meta-matching.

4. DISCUSSION

In this study, we proposed two meta-matching algorithms to translate phenotypic prediction models from source datasets with disparate sizes to predict new phenotypes in small datasets. Both approaches outperformed meta-matching using a single source dataset (UK Biobank). Both approaches also outperformed classical KRR and classical transfer learning by a big margin. Furthermore, multilayer meta-matching compared favorably with meta-matching with dataset stacking across both HCP-YA and HCP-Aging datasets. In terms of feature importance based on the Haufe transform, we found that feature importance values of multilayer meta-matching

Table 2. Prediction using classical KRR versus meta-matching with stacking on medium source datasets.

Datasets	Prediction performance (Pearson's correlation) of classical KRR	Prediction performance of (Pearson's correlation) of meta-matching w/ stacking (from UKBB + ABCD)	Correlation between phenotypic prediction by above two methods
GSP	0.0953	0.106	0.400
HBN	0.167	0.144	0.433
eNKI-RS	0.154	0.196	0.600

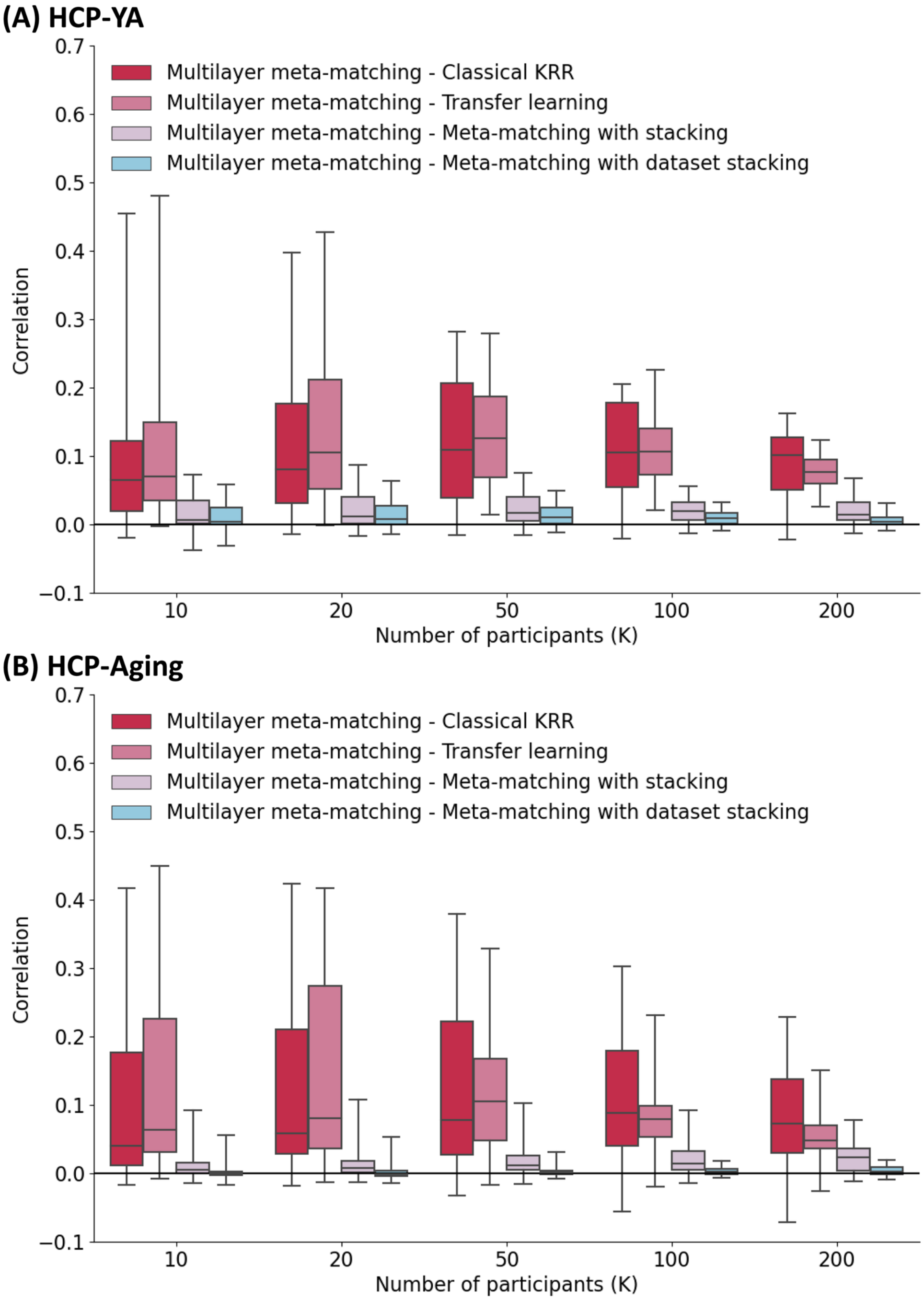

Fig. 6. Numerical improvement in prediction performance (Pearson's correlation) across different phenotypes in the HCP-YA and HCP-Aging datasets. (A) Phenotypic prediction performance (averaged across 100 repetitions of sampling K participants) in the HCP-YA dataset. Horizontal axis is the number of participants in the HCP-YA dataset used to adapt the models trained from the meta-training datasets. Boxplots represent variability across the 35 HCP-YA phenotypes. The bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Whiskers correspond to 1.5 times the interquartile range. (B) Same plot as panel A except that the analyses were performed in the HCP-Aging dataset with 45 phenotypes.

and classical KRR seemed to be equally similar to the pseudo ground truth, while feature importance values of transfer learning were the furthest away from the pseudo ground truth. Therefore, there was no trade-off between prediction accuracy and feature importance accuracy (with respect to the pseudo ground truth), which is consistent with our previous study (Chen et al., 2023).

The relatively poor performance of classical transfer learning was somewhat surprising but probably indicated the difficulty of finetuning so many parameters in the very small sample regime. We note that the transfer learning baseline is similar to a meta-matching variant "meta-matching finetune" from our previous study (He et al., 2022), except for one key difference. Both meta-matching

Table 3. Percentages of phenotypes with numerical improvement in prediction performance (Pearson's correlation).

Datasets	K	Multilayer MM vs. classical KRR	Multilayer MM vs. transfer learning	Multilayer MM vs. MM w/ stacking	Multilayer MM vs. MM w/ dataset stacking
HCP-YA	10	85.7%	94.3%	80.0%	74.3%
	20	85.7%	100%	77.1%	74.3%
	50	88.6%	100%	85.7%	80.0%
	100	91.4%	100%	85.7%	77.1%
	200	97.1%	100%	85.7%	74.3%
HCP-Aging	10	86.7%	95.6%	75.6%	44.4%
	20	86.7%	95.6%	77.8%	42.2%
	50	88.9%	93.3%	82.2%	57.7%
	100	88.9%	95.6%	77.8%	57.7%
	200	88.9%	95.6%	82.2%	64.4%

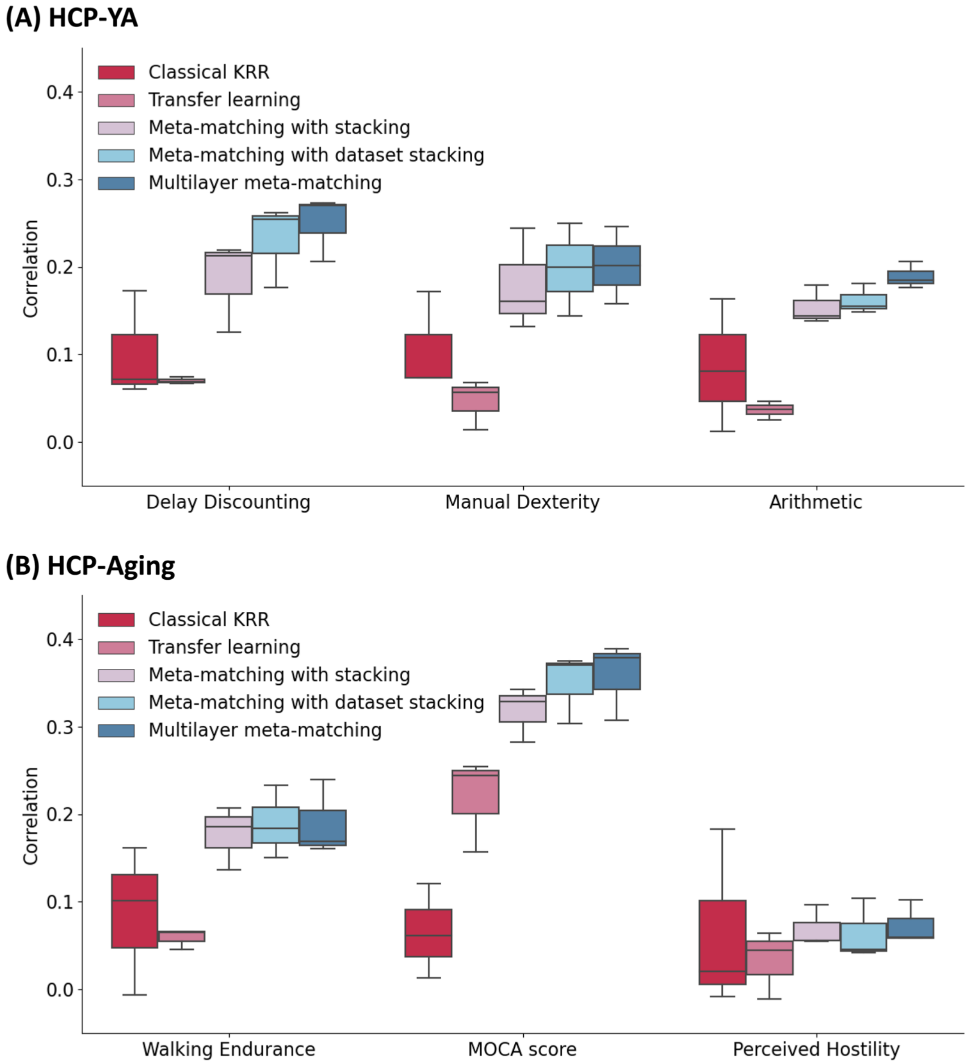

Fig. 7. Examples of phenotypic prediction performance in the (A) HCP-YA and (B) HCP-Aging datasets in the case of 100-shot learning (K = 100). Here, prediction performance was measured using Pearson's correlation. For each box plot, the horizontal line indicates the median. The bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Whiskers correspond to 1.5 times the interquartile range.

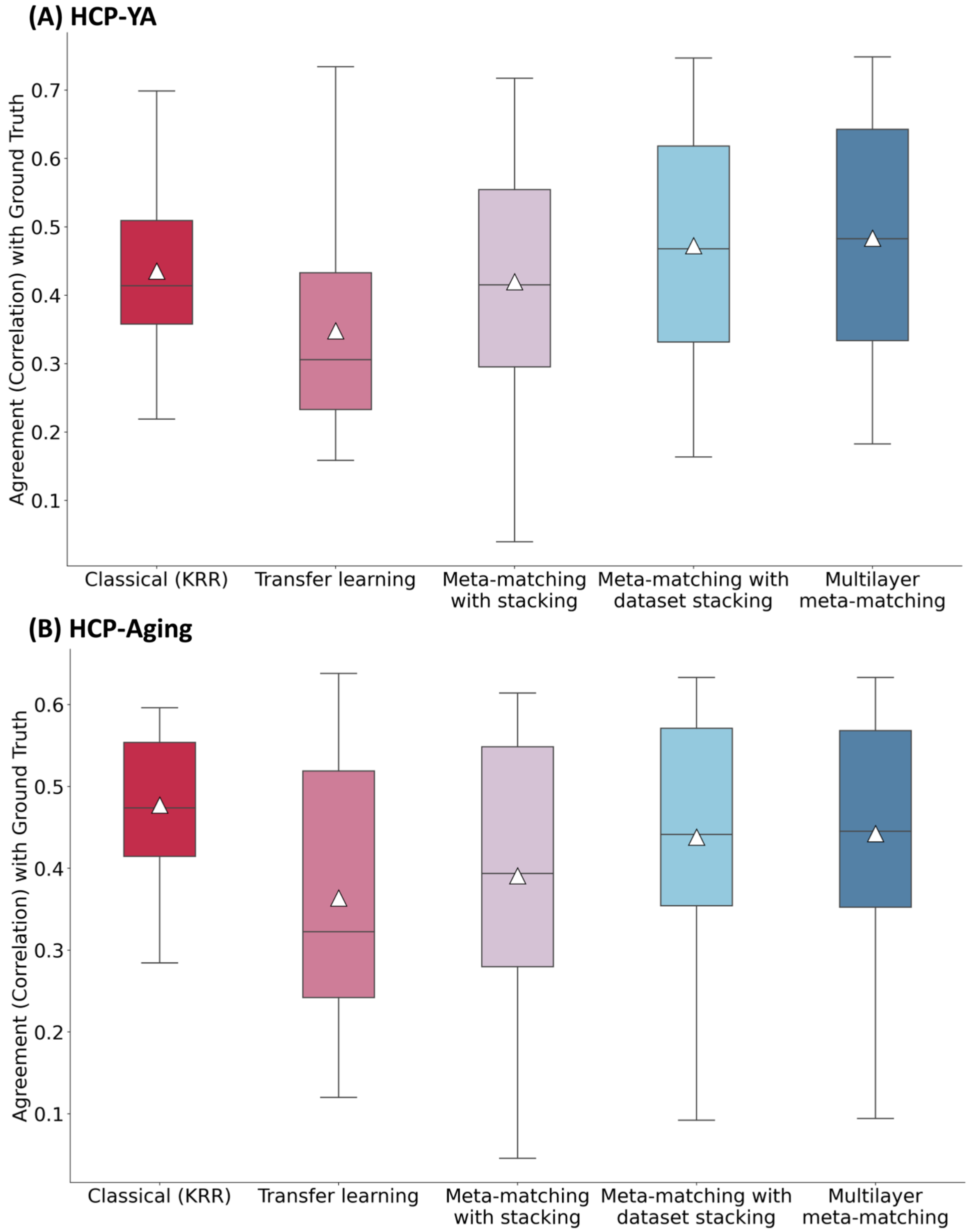

Fig. 8. Agreement (correlation) of feature importance values with pseudo ground truth in the (A) HCP-YA and (B) HCP-Aging datasets. For each approach, the Haufe transform was used to estimate feature importance in the 100-shot scenario (K = 100), which was then compared with the pseudo ground truth. Pseudo ground truth feature importance was generated by applying the Haufe transform to a KRR model trained from the full target dataset. For each box plot, the horizontal line indicates the median, and the triangle indicates the mean. The bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. Whiskers correspond to 1.5 times the interquartile range.

finetune and classical transfer learning finetuned the last two layers of the DNN. However, transfer learning initialized the last layer of the DNN from scratch (Section 2.3.2), and then finetuned the last two layers. On the other hand, meta-matching finetune first selected the output node that predicted the K meta-test participants the best (for a particular meta-test phenotype), and retained the weights leading to the output node. The last two layers of the DNN were then finetuned, given that meta-matching finetune was much better than classical KRR (He et al., 2022), but classical transfer learning was worse than KRR in the current study. This further supported the importance of the meta-matching idea.

4.1. Meta-learning, transfer learning, and related problems

We mentioned in the introduction that the name "meta-matching" was motivated by the "matching" of meta-training and meta-test phenotypes. The name "meta-matching" was also motivated by its close links with meta-learning (Andrychowicz et al., 2016; Fei-Fei et al., 2006; Finn et al., 2017; Ravi & Larochelle, 2016; Vanschoren, 2019). Meta-learning is often referred to as "learning to learn" and is closely related to "transfer learning" (Hospedales et al., 2021). Both meta-learning and transfer learning seek to improve prediction in a new domain with limited training data using knowledge gained from previous domains (Pan & Yang, 2009). The goal of learning from limited training data (e.g., K training examples) is often referred to as few-shot (or K-shot) learning (Hospedales et al., 2021).

Meta-learning typically involves two learning levels (Huisman et al., 2021). At one level, the algorithm seeks to rapidly learn a new task with limited quantity of data. This rapid learning of a new task is made possible by knowledge learned from earlier tasks at another level. Therefore, meta-matching is similar in spirit to meta-learning. At one level, meta-matching involves training models to predict meta-training phenotypes. These trained models are then rapidly adapted to predict a new meta-test set at another level. However, our meta-matching approach differs from modern meta-learning algorithms that typically involve a meta-objective that is used to optimize an inner-loop learner (Hospedales et al., 2021).

Transfer learning can be broadly defined as using past experience from one or more source tasks to improve learning on a target task (Hospedales et al., 2021). Therefore, meta-learning is one approach that can be used to improve transfer learning (Hospedales et al., 2021). Consequently, we can also consider meta-matching as a type of transfer learning algorithm. One distinction between meta-learning and transfer learning is that meta-learning always involves training a machine-learning model from a wide range of meta-training tasks and then adapting to perform a new prediction problem in the target dataset. On the other hand, in transfer learning, the prediction problem in the target dataset can be the same (Chen et al., 2020; Vakli et al., 2018; Zhang & Bellec, 2020) or different (Hon & Khan, 2017; Lu et al., 2021; Schirmer et al., 2021) in the source dataset. While the prediction problem is the same in the target and source domains, the input feature distribution might be different between the two domains, which is a problem known as domain shift (Hospedales et al., 2021). Domain adaptation is, therefore, a type of transfer learning which seeks to address the problem of domain shift.

Finally, we note that the stacking procedure employed by multilayer meta-matching (and other meta-matching variants) utilized the predictions of meta-training phenotypes as input features to predict new meta-test phenotypes. This contrasts with many RSFC-based prediction approaches that utilized the RSFC data directly (Finn et al., 2015; He et al., 2020), and is reminiscent of studies predicting a phenotype from previously predicted measures (Gal, Tik, et al., 2022; Yoo et al., 2022).

4.2. Limitations and future work

One important limitation of meta-matching is that the magnitude of prediction improvement heavily depends on the correlations between meta-training and meta-test phenotypes (He et al., 2022). Consequently, we do not expect all meta-test phenotypes to benefit from meta-matching (Fig. 6). However, it is important to note that this limitation exists for all meta-learning and transfer learning algorithms (Jose & Simeone, 2021; Zhang et al., 2017). Model transfer is easier if the source and target domains are more similar. Performance will degrade if the source and target domains are very different. This observation motivates the addition of more source datasets.

Based on the current trends (Figs. 4 and 5), we might expect multilayer meta-matching to remain better than classical KRR beyond 200 participants. However, we would expect classical KRR to catch up for larger K, and might ultimately be better than multilayer meta-matching for relatively large K. A hint of this crossover can be found in Table 2, where KRR was numerically better than meta-matching with stacking for 5-fold cross-validation of HBN (N = 930), but numerically worse than meta-matching with stacking for 5-fold cross-validation of GSP (N = 862) and eNKI-RS (N = 896).

Finally, we note that there are multiple possible extensions to the current work. Within the context of resting-state functional connectivity, we could explore the use of individual-specific parcellations, which have been shown to improve phenotypic prediction performance compared with group-level parcellations (Kong et al., 2021; M. Li et al., 2019). Furthermore, previous studies have suggested that other FC measures (e.g., partial correlations) can lead to better prediction performance than Pearson's correlation (Dadi et al., 2019; Farahibozorg et al., 2021; Pervaiz et al., 2020). Some studies have suggested that fine-grained FC might capture additional behavioral information (Feilong et al., 2021). Therefore, meta-matching models based on other FC measures (e.g., fine-grained FC and partial correlations) might also be explored.

Beyond resting-state functional connectivity, meta-matching can be applied to other imaging modalities, such as task-FC (Chen et al., 2022; Greene et al., 2018) and fMRI during naturalistic stimulus (Finn, 2021; Finn & Bandettini, 2021; Gal, Coldham, et al., 2022), which have shown improvements over RSFC for phenotypic prediction. However, developing meta-matching models for task-fMRI and naturalistic-FC is more challenging because large datasets with consistent task or movie paradigm are not common. Other alternative modalities include anatomical T1 images and diffusion MRI. In the case of anatomical T1 images, we could simply replace the fully connected feedforward DNN used in the current study with 3D convolutional neural networks (Wulan et al., 2024). Finally, the datasets in the current study comprised relatively healthy participants. Meta-matching might be potentially useful for psychiatric populations (Chopra et al., 2022). Including psychiatric datasets to the base model training might further improve generalization to new datasets by increasing the diversity of the source datasets.

DATA AND CODE AVAILABILITY

This study utilized publicly available data from the UK Biobank (https://www.ukbiobank.ac.uk/), ABCD (https://abcdstudy.org), GSP (http://neuroinformatics.harvard.edu/gsp/), HBN (https://fcon_1000.projects.nitrc.org/indi/cmi_healthy_brain_network), eNKI-RS (http://fcon_1000.projects.nitrc.org/indi/enhanced/), and HCP (https://www.humanconnectome.org/). Data can be accessed via data use agreements.

Code for the classical KRR baseline and meta-matching algorithms can be found here (https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/predict_phenotypes/Chen2024_MMM). The trained models for multilayer meta-matching are also publicly available (https://github.com/ThomasYeoLab/Meta_matching_models/tree/main/rs-fMRI/v2.0). The code was reviewed by two co-authors (LA and CZ) before merging into the GitHub repository to reduce the chance of coding errors.

AUTHOR CONTRIBUTIONS

P.C., L.A., N.W., C.Z., S.Z., L.Q.R.O., R.K., J.C., J.W., S.C., D.B., S.B.E., A.J.H., and B.T.T.Y. designed the research. P.C. conducted the research. P.C., L.A., N.W., C.Z., S.Z., L.Q.R.O., R.K., J.C., J.W., S.C., D.B., S.B.E., A.J.H., and B.T.T.Y. interpreted the results. P.C. and B.T.T.Y. wrote the manuscript and made the figures. P.C., L.A., and C.Z. reviewed and published the code. All authors contributed to project direction via discussion. All authors edited the manuscript.

DECLARATION OF COMPETING INTEREST

The authors declare no competing interests.

ACKNOWLEDGMENTS

This work was supported by the NUS Yong Loo Lin School of Medicine (NUHSRO/2020/124/TMR/LOA), the Singapore National Medical Research Council (NMRC) LCG (OFLCG19May-0035), NMRC CTG-IIT (CTGIIT-23jan-0001), NMRC STaR (STaR20nov-0003), Singapore Ministry of Health (MOH) Centre Grant (CG21APR1009), the Temasek Foundation (TF2223-IMH-01), and the United States National Institutes of Health (R01MH120080 & R01MH133334). Our computational work was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Singapore NRF, NMRC, MOH, or the Temasek Foundation. Data used in this research were provided by: (1) the UK Biobank resource under application 25163; (2) the Adolescent Brain Cognitive Development^{SM} (ABCD) Study (https://abcdstudy.org), held in the NIMH Data Archive (NDA). This is a multisite, longitudinal study designed to recruit more than 10,000 children age 9–10 and follow them over 10 years into early adulthood. The ABCD Study® is supported by the National Institutes of Health and additional federal partners under award numbers U01DA041048, U01DA050989, U01DA051016, U01DA041022, U01DA051018, U01DA051037, U01DA050987, U01DA041174, U01DA041106, U01DA041117, U01DA041028, U01DA041134, U01DA050988, U01DA051039, U01DA041156, U01DA041025, U01DA041120, U01DA051038, U01DA041148, U01DA041093, U01DA041089, U24DA041123, and U24DA041147. A full list of supporters is available at https://abcdstudy.org/federal-partners.html. A listing of participating sites and a complete listing of the study investigators can be found at https://abcdstudy.org/consortium_members/. ABCD consortium investigators designed and implemented the study and/or provided data but did not necessarily participate in the analysis or writing of this report. This manuscript reflects the views of the authors and may not reflect the opinions or views of the NIH or ABCD consortium investigators. The ABCD data repository grows and changes over time. The ABCD data used in this report came from http://dx.doi.org/10.15154/1504041; (3) 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. Twenty individual investigators at Harvard and MGH generously contributed data to the overall project; (4) The HBN (http://www.healthybrainnetwork.org) and its initiatives are supported by philanthropic contributions from the following individuals, foundations, and organizations: Margaret Bilotti; Brooklyn Nets; Agapi and Bruce Burkard; James Chang; Phyllis Green and Randolph Cōwen; Grieve Family

Fund; Susan Miller and Byron Grote; Sarah and Geoff Gund; George Hall; Jonathan M. Harris Family Foundation; Joseph P. Healey; The Hearst Foundations; Eve and Ross Jaffe; Howard & Irene Levine Family Foundation; Rachael and Marshall Levine; George and Nitzia Logothetis; Christine and Richard Mack; Julie Minskoff; Valerie Mnuchin; Morgan Stanley Foundation; Amy and John Phelan; Roberts Family Foundation; Jim and Linda Robinson Foundation, Inc.; The Schaps Family; Zibby Schwarzman; Abigail Pogrebin and David Shapiro; Stavros Niarchos Foundation; Preethi Krishna and Ram Sundaram; Amy and John Weinberg; Donors to the 2013 Child Advocacy Award Dinner Auction; Donors to the 2012 Brant Art Auction; (5) the enhanced Nathan Kline InstituteRockland Sample (eNKI-RS) database (https://fcon_1000.projects.nitrc.org/indi/enhanced/access.html); (6) the Human Connectome Project, the 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; and (7) 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).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available with the online version here: https://doi.org/10.1162/imag_a_00233

REFERENCES

Abraham, A., Milham, M. P., Di Martino, A., Craddock, R. C., Samaras, D., Thirion, B., & Varoquaux, G. (2017). Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. *NeuroImage*, *147*, 736–745. https://doi.org/10.1016/j.neuroimage.2016.10.045

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2623–2631). Association for Computing Machinery. https://doi.org/10.1145/3292500.3330701

Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega-Potler, N., Langer, N., Alexander, A., & Kovacs, M. (2017). An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, *4*(1), 1–26. https://doi.org/10.1038/sdata.2017.181

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., & Vallee, E. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage*, *166*, 400–424. https://doi.org/10.1016/j.neuroimage.2017.10.034

Andrychowicz, M., Denil, M., Gomez, S., Hoffman, M. W., Pfau, D., Schaul, T., Shillingford, B., & De Freitas, N. (2016). Learning to learn by gradient descent by gradient descent. *Advances in Neural Information Processing Systems*, *29*. https://proceedings.neurips.cc/paper_files/paper/2016/hash/fb87582825f9d28a8d42c5e5e5e8b23d-Abstract.html

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, *145*, 137–165. https://doi.org/10.1016/j.neuroimage.2016.02.079

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*(1), 49–64. https://doi.org/10.1007/BF00117832

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007

Chen, C.-L., Hsu, Y.-C., Yang, L.-Y., Tung, Y.-H., Luo, W.-B., Liu, C.-M., Hwang, T.-J., Hwu, H.-G., & Tseng, W.-Y. I. (2020). Generalization of diffusion magnetic resonance imaging–based brain age prediction model through transfer learning. *NeuroImage*, *217*, 116831. https://doi.org/10.1016/j.neuroimage.2020.116831

Chen, J., Ooi, L. Q. R., Tan, T. W. K., Zhang, S., Li, J., Asplund, C. L., Eickhoff, S. B., Bzdok, D., Holmes, A. J., & Yeo, B. T. (2023). Relationship between prediction accuracy and feature importance reliability: An empirical and theoretical study. *NeuroImage*, *274*, 120115. https://doi.org/10.1016/j.neuroimage.2023.120115

Chen, J., Tam, A., Kebets, V., Orban, C., Ooi, L. Q. R., Asplund, C. L., Marek, S., Dosenbach, N. U., Eickhoff, S. B., & Bzdok, D. (2022). Shared and unique brain network features predict cognitive, personality, and mental health scores in the ABCD study. *Nature Communications*, *13*(1), 2217. https://doi.org/10.1038/s41467-022-29766-8

Chopra, S., Dhamala, E., Lawhead, C., Ricard, J. A., Orchard, E. R., An, L., Chen, P., Wulan, N., Kumar, P., Rubenstein, A., Moses, J., Chen, L., Levi, P., Holmes, A., Aquino, K., Fornito, A., Harpaz-Rotem, I., Germine, L. T., Baker, J. T., … Holmes, A. J. (2022). Reliable and generalizable brain-based predictions of cognitive functioning across common psychiatric illness. *medRxiv*. https://doi.org/10.1101/2022.12.08.22283232

Chu, C., Hsu, A.-L., Chou, K.-H., Bandettini, P., Lin, C., & Initiative, A. D. N. (2012). Does feature selection improve classification accuracy? Impact of sample size and feature selection on classification using anatomical magnetic resonance images. *NeuroImage*, *60*(1), 59–70. https://doi.org/10.1016/j.neuroimage.2011.11.066

Cui, Z., & Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *NeuroImage*, *178*, 622–637. https://doi.org/10.1016/j.neuroimage.2018.06.001

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., & Varoquaux, G. (2019). Benchmarking functional connectome-based predictive models for resting-state fMRI. *NeuroImage*, *192*, 115–134. https://doi.org/10.1016/j.neuroimage.2019.02.062

Dietterich, T. G. (2000). Ensemble methods in machine learning. In G. Goos, J. Hartmanis, & J. Van Leeuwen

(Eds.), *Multiple classifier systems* (Vol. *1857*, pp. 1–15). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

Eickhoff, S. B., & Langner, R. (2019). Neuroimaging-based prediction of mental traits: Road to utopia or Orwell? *PLoS Biology*, *17*(11), e3000497. https://doi.org/10.1371/journal.pbio.3000497

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. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. https://doi.org/10.1038/s41592-018-0235-4

Fair, D. A., Miranda-Dominguez, O., Snyder, A. Z., Perrone, A., Earl, E. A., Van, A. N., Koller, J. M., Feczko, E., Tisdall, M. D., & van der Kouwe, A. (2020). Correction of respiratory artifacts in MRI head motion estimates. *NeuroImage*, *208*, 116400. https://doi.org/10.1016/j.neuroimage.2019.116400

Farahibozorg, S.-R., Bijsterbosch, J. D., Gong, W., Jbabdi, S., Smith, S. M., Harrison, S. J., & Woolrich, M. W. (2021). Hierarchical modelling of functional brain networks in population and individuals from big fMRI data. *NeuroImage*, *243*, 118513. https://doi.org/10.1016/j.neuroimage.2021.118513

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 594–611. https://doi.org/10.1109/tpami.2006.79

Feilong, M., Guntupalli, J. S., & Haxby, J. V. (2021). The neural basis of intelligence in fine-grained cortical topographies. *Elife*, *10*, e64058. https://doi.org/10.7554/elife.64058

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning*, *70*, 1126–1135. https://proceedings.mlr.press/v70/finn17a.html

Finn, E. S. (2021). Is it time to put rest to rest? *Trends in Cognitive Sciences*, *25*(12), 1021–1032. https://doi.org/10.1016/j.tics.2021.09.005

Finn, E. S., & Bandettini, P. A. (2021). Movie-watching outperforms rest for functional connectivity-based prediction of behavior. *NeuroImage*, *235*, 117963. https://doi.org/10.1016/j.neuroimage.2021.117963

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*(11), 1664–1671. https://doi.org/10.1038/nn.4135

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

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., & Klaveness, S. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341–355. https://doi.org/10.1016/s0896-6273(02)00569-x

Gabrieli, J. D., Ghosh, S. S., & Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron*, *85*(1), 11–26. https://doi.org/10.1016/j.neuron.2014.10.047

Gal, S., Coldham, Y., Tik, N., Bernstein-Eliav, M., & Tavor, I. (2022). Act natural: Functional connectivity from naturalistic stimuli fMRI outperforms resting-state in predicting brain activity. *NeuroImage*, *258*, 119359. https://doi.org/10.1016/j.neuroimage.2022.119359

Gal, S., Tik, N., Bernstein-Eliav, M., & Tavor, I. (2022). Predicting individual traits from unperformed tasks. *NeuroImage*, *249*, 118920. https://doi.org/10.1016/j.neuroimage.2022.118920

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

Gordon, E. M., Laumann, T. O., Adeyemo, B., Huckins, J. F., Kelley, W. M., & Petersen, S. E. (2016). Generation and evaluation of a cortical area parcellation from resting-state correlations. *Cerebral Cortex*, *26*(1), 288–303. https://doi.org/10.1093/cercor/bhu239

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

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *NeuroImage*, *48*(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Hagler Jr, D. J., Hatton, S., Cornejo, M. D., Makowski, C., Fair, D. A., Dick, A. S., Sutherland, M. T., Casey, B. J., Barch, D. M., & Harms, M. P. (2019). Image processing and analysis methods for the Adolescent Brain Cognitive Development Study. *NeuroImage*, *202*, 116091. https://doi.org/10.1016/j.neuroimage.2019.116091

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

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., 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

He, T., An, L., Chen, P., Chen, J., Feng, J., Bzdok, D., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. (2022). Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nature Neuroscience*, *25*(6), 795–804. https://doi.org/10.1038/s41593-022-01059-9

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., & Yeo, B. 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. (2015). Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. *Scientific Data*, *2*(1), 1–16. https://doi.org/10.1038/sdata.2015.31

Hon, M., & Khan, N. M. (2017). Towards Alzheimer's disease classification through transfer learning. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1166–1169). IEEE. https://ieeexplore.ieee.org/abstract/document/8217822/

Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2021). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5149–5169. https://doi.org/10.1109/tpami.2021.3079209

Huisman, M., Van Rijn, J. N., & Plaat, A. (2021). A survey of deep meta-learning. *Artificial Intelligence Review*, *54*(6), 4483–4541. https://doi.org/10.1007/s10462-021-10004-4

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, *17*(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *NeuroImage*, *62*(2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

Jose, S. T., & Simeone, O. (2021). An information-theoretic analysis of the impact of task similarity on meta-learning. In *2021 IEEE International Symposium on Information Theory (ISIT)* (pp. 1534–1539). IEEE. https://doi.org/10.1109/ISIT45174.2021.9517767

Kadam, S., & Vaidya, V. (2020). Review and analysis of zero, one and few shot learning approaches. In A. Abraham, A. K. Cherukuri, P. Melin, & N. Gandhi (Eds.), *Intelligent systems design and applications* (Vol. *940*, pp. 100–112). Springer International Publishing. https://doi.org/10.1007/978-3-030-16657-1_10

Kebets, V., Holmes, A. J., Orban, C., Tang, S., Li, J., Sun, N., Kong, R., Poldrack, R. A., & Yeo, B. T. (2019). Somatosensory-motor dysconnectivity spans multiple transdiagnostic dimensions of psychopathology. *Biological Psychiatry*, *86*(10), 779–791. https://doi.org/10.1016/j.biopsych.2019.06.013

Kong, R., Li, J., Orban, C., Sabuncu, M. R., Liu, H., Schaefer, A., Sun, N., Zuo, X.-N., Holmes, A. J., & Eickhoff, S. B. (2019). Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. *Cerebral Cortex*, *29*(6), 2533–2551. https://doi.org/10.1093/cercor/bhy123

Kong, R., Tan, Y. R., Wulan, N., Ooi, L. Q. R., Farahibozorg, S.-R., Harrison, S., Bijsterbosch, J. D., Bernhardt, B. C., Eickhoff, S., & Yeo, B. T. (2023). Comparison between gradients and parcellations for functional connectivity prediction of behavior. *NeuroImage*, *273*, 120044. https://doi.org/10.1016/j.neuroimage.2023.120044

Kong, R., Yang, Q., Gordon, E., Xue, A., Yan, X., Orban, C., Zuo, X.-N., Spreng, N., Ge, T., & Holmes, A. (2021). Individual-specific areal-level parcellations improve functional connectivity prediction of behavior. *Cerebral Cortex*, *31*(10), 4477–4500. https://doi.org/10.1093/cercor/bhab101

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, *51*, 181–207. https://doi.org/10.1023/a:1022859003006

Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L. Q. R., Holmes, A. J., Ge, T., Patil, K. R., Jabbi, M., Eickhoff, S. B., Yeo, B. T. T., & Genon, S. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Science Advances*, *8*(11), eabj1812. https://doi.org/10.1126/sciadv.abj1812

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

Li, M., Wang, D., Ren, J., Langs, G., Stoecklein, S., Brennan, B. P., Lu, J., Chen, H., & Liu, H. (2019). Performing group-level functional image analyses based on homologous functional regions mapped in individuals. *PLoS Biology*, *17*(3), e2007032. https://doi.org/10.1371/journal.pbio.2007032

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Masouleh, S. K., Huntenburg, J. M., Lampe, L., Rahim, M., Abraham, A., & Craddock, R. C. (2017). Predicting brain-age from multimodal imaging data captures cognitive impairment. *NeuroImage*, *148*, 179–188. https://doi.org/10.1016/j.neuroimage.2016.11.005

Lu, B., Li, H.-X., Chang, Z.-K., Li, L., Chen, N.-X., Zhu, Z.-C., Zhou, H.-X., Fan, Z., Yang, H., Chen, X., & Yan, C.-G. (2021). Classification of Sex and Alzheimer's disease via brain imaging-based deep learning on 85,721 samples. *bioRxiv*. https://doi.org/10.1101/2020.08.18.256594

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., & Hendrickson, T. J. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), 654–660. https://doi.org/10.1038/s41586-022-04492-9

Masouleh, S. K., Eickhoff, S. B., Hoffstaedter, F., Genon, S., & Initiative, A. D. N. (2019). Empirical examination of the replicability of associations between brain structure and psychological variables. *Elife*, *8*, e43464. https://doi.org/10.7554/elife.43464

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., & Andersson, J. L. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*, *19*(11), 1523–1536. https://doi.org/10.1038/nn.4393

Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., & Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? *NeuroImage*, *44*(3), 893–905. https://doi.org/10.1016/j.neuroimage.2008.09.036

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., Sikka, S., Gutman, D., Bangaru, S., Schlachter, R.T., Kamiel, S.M., Anwar, A.R., Hinz, C.M., Kaplan, M.S., Rachlin, A.B., … Milham, M.P. (2012). The NKI-Rockland sample: A model for accelerating the pace of discovery science in psychiatry. *Frontiers in Neuroscience*, *6*, 152. https://doi.org/10.3389/fnins.2012.00152

Ooi, L. Q. R., Chen, J., Zhang, S., Kong, R., Tam, A., Li, J., Dhamala, E., Zhou, J. H., Holmes, A. J., & Yeo, B. T. (2022). Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI. *NeuroImage*, *263*, 119636. https://doi.org/10.1016/j.neuroimage.2022.119636

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/tkde.2009.191

Pervaiz, U., Vidaurre, D., Woolrich, M. W., & Smith, S. M. (2020). Optimising network modelling methods for fMRI. *NeuroImage*, *211*, 116604. https://doi.org/10.1016/j.neuroimage.2020.116604

Poldrack, R. A., Huckins, G., & Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry*, *77*(5), 534–540. https://doi.org/10.1001/jamapsychiatry.2019.3671

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

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*, 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048

Pruim, R. H., 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. https://doi.org/10.1016/j.neuroimage.2015.02.063

Pruim, R. H., Mennes, M., van Rooij, D., 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. https://doi.org/10.1016/j.neuroimage.2015.02.064

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

Ravi, S., & Larochelle, H. (2016). Optimization as a model for few-shot learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rJY0-Kcll

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., & Yeo, B. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, *28*(9), 3095–3114. https://doi.org/10.1093/cercor/bhx179

Schirmer, M. D., Venkataraman, A., Rekik, I., Kim, M., Mostofsky, S. H., Nebel, M. B., Rosch, K., Seymour, K., Crocetti, D., & Irzan, H. (2021). Neuropsychiatric disease classification using functional connectomics-results of the connectomics in neuroimaging transfer learning challenge. *Medical Image Analysis*, *70*, 101972. https://doi.org/10.1016/j.media.2021.101972

Schulz, M.-A., Yeo, B. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., Richards, B., & Bzdok, D. (2020). Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications*, *11*(1), 4238. https://doi.org/10.1038/s41467-020-18037-z

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., & Flitney, D. E. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23*, S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., Behrens, T. E., Glasser, M. F., Ugurbil, K., Barch, D. M., Van Essen, D. C., & Miller, K. L. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, *18*(11), 1565–1567. https://doi.org/10.1038/nn.4125

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., & Landray, M. (2015). UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Medicine*, *12*(3), e1001779. https://doi.org/10.1371/journal.pmed.1001779

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

Vakli, P., Deák-Meszlényi, R. J., Hermann, P., & Vidnyánszky, Z. (2018). Transfer learning improves resting-state functional connectivity pattern analysis using convolutional neural networks. *Gigascience*, *7*(12), giy130. https://doi.org/10.1093/gigascience/giy130

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., & Consortium, W.-M. H. (2013). The WU-Minn human connectome project: An overview. *NeuroImage*, *80*, 62–79. https://doi.org/10.1016/j.neuroimage.2013.05.041

Vanschoren, J. (2019). Meta-learning. In F. Hutter, L. Kotthoff, & J. Vanschoren (Eds.), *Automated Machine Learning: Methods, Systems, Challenges* (pp. 35–61). Springer. https://doi.org/10.1007/978-3-030-05318-5_2

Varoquaux, G., & Poldrack, R. A. (2019). Predictive models avoid excessive reductionism in cognitive neuroimaging. *Current Opinion in Neurobiology*, *55*, 1–6. https://doi.org/10.1016/j.conb.2018.11.002

Volkow, N. D., Koob, G. F., Croyle, R. T., Bianchi, D. W., Gordon, J. A., Koroshetz, W. J., Pérez-Stable, E. J., Riley, W. T., Bloch, M. H., & Conway, K. (2018). The conception of the ABCD study: From substance use to a broad NIH collaboration. *Developmental Cognitive Neuroscience*, *32*, 4–7. https://doi.org/10.1016/j.dcn.2017.10.002

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data*, *3*(1), 1–40. https://doi.org/10.1186/s40537-016-0043-6

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259. https://doi.org/10.1016/s0893-6080(05)80023-1

Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience*, *20*(3), 365–377. https://doi.org/10.1038/nn.4478

Wu, J., Li, J., Eickhoff, S. B., Hoffstaedter, F., Hanke, M., Yeo, B. T., & Genon, S. (2022). Cross-cohort replicability and generalizability of connectivity-based psychometric prediction patterns. *NeuroImage*, *262*, 119569. https://doi.org/10.1016/j.neuroimage.2022.119569

Wulan, N., An, L., Zhang, C., Kong, R., Chen, P., Bzdok, D., Eickhoff, S. B., Holmes, A. J., & Yeo, B. T. (2024). Translating phenotypic prediction models from big to small anatomical MRI data using meta-matching. *bioRxiv*. https://doi.org/10.1101/2023.12.31.573801

Xia, C. H., Ma, Z., Ciric, R., Gu, S., Betzel, R. F., Kaczkurkin, A. N., Calkins, M. E., Cook, P. A., García de la Garza, A., & Vandekar, S. N. (2018). Linked dimensions of psychopathology and connectivity in functional brain networks. *Nature Communications*, *9*(1), 3003. https://doi.org/10.1038/s41467-018-05317-y

Yoo, K., Rosenberg, M. D., Kwon, Y. H., Scheinost, D., Constable, R. T., & Chun, M. M. (2022). A cognitive state transformation model for task-general and task-specific subsystems of the brain connectome. *NeuroImage*, *257*, 119279. https://doi.org/10.1016/j.neuroimage.2022.119279

Zhang, W., Fang, Y., & Ma, Z. (2017). The effect of task similarity on deep transfer learning. In D. Liu, S. Xie, Y. Li, D. Zhao, & E.-S. M. El-Alfy (Eds.), *Neural information processing* (Vol. *10635*, pp. 256–265). Springer International Publishing. https://doi.org/10.1007/978-3-319-70096-0_27

Zhang, Y., & Bellec, P. (2020). Transferability of brain decoding using graph convolutional networks. *bioRxiv*. https://doi.org/10.1101/2020.06.21.163964