bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Translating phenotypic prediction models from big to small anatomical MRI data using meta-matching

Naren Wulan[1,2,3], Lijun An[1,2,3], Chen Zhang[1,2,3], Ru Kong[1,2,3], Pansheng Chen[1,2,3], Danilo Bzdok[5,6], Simon B Eickhoff[7,8], Avram J Holmes[9], B. T. Thomas Yeo[1,2,3,4,10]

[1]Centre for Sleep & Cognition & Centre for Translational Magnetic Resonance Research, Yong Loo Lin School of Medicine, National University of Singapore [2]Department of Electrical and Computer Engineering, National University of Singapore, Singapore [3]N.1 Institute for Health & Institute for Digital Medicine, National University of Singapore, Singapore [4]Integrative Sciences and Engineering Programme (ISEP), National University of Singapore, Singapore [5]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 [6]Mila – Quebec Artificial Intelligence Institute, Montreal, QC, Canada [7]Institute for Systems Neuroscience, Medical Faculty, Heinrich-Heine University Düsseldorf, Düsseldorf, Germany [8]Institute of Neuroscience and Medicine, Brain & Behavior (INM-7), Research Center Jülich, Jülich, Germany [9]Department of Psychiatry, Brain Health Institute, Rutgers University, Piscataway, NJ, USA [10]Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA, USA

Address correspondence to:

B.T. Thomas Yeo
ECE, CSC, TMR, N.1 & WisDM
National University of Singapore
Email: thomas.yeo@nus.edu.sg

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Individualized phenotypic prediction based on structural MRI is an important goal in neuroscience. Prediction performance increases with larger samples, but small-scale datasets with fewer than 200 participants are often unavoidable. We have previously proposed a "meta-matching" framework to translate models trained from large datasets to improve the prediction of new unseen phenotypes in small collection efforts. Meta-matching exploits correlations between phenotypes, yielding large improvement over classical machine learning when applied to prediction models using resting-state functional connectivity as input features. Here, we adapt the two best performing meta-matching variants ("meta-matching finetune" and "meta-matching stacking") from our previous study to work with T1-weighted MRI data by changing the base neural network architecture to a 3D convolution neural network. We compare the two meta-matching variants with elastic net and classical transfer learning using the UK Biobank (N = 36,461), Human Connectome Project Young Adults (HCP-YA) dataset (N = 1,017) and HCP-Aging dataset (N = 656). We find that meta-matching outperforms elastic net and classical transfer learning by a large margin, both when translating models within the same dataset, as well as translating models across datasets with different MRI scanners, acquisition protocols and demographics. For example, when translating a UK Biobank model to 100 HCP-YA participants, meta-matching finetune yielded a 136% improvement in variance explained over transfer learning, with an average absolute gain of 2.6% (minimum = -0.9%, maximum = 17.6%) across 35 phenotypes. Overall, our results highlight the versatility of the meta-matching framework.

Keywords:

Structural MRI, transfer learning, meta-matching, phenotypic prediction

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1. Introduction

An important goal in systems neuroscience is to understand how variation in brain structure relates to individual differences in behavior (Genon et al., 2022). Structural T1-weighted magnetic resonance imaging (MRI) is a non-invasive technique for examining the anatomy of the human brain, providing high contrast between gray and white matter (Gifford et al., 2020). Structural MRI is widely used to predict behavioral traits, clinical symptoms and diagnostic categories in both healthy individuals and individuals with neuropsychiatric disorders (Sabuncu et al., 2015; Arbabshirani et al., 2017; Bhagwat et al., 2019; Cohen et al., 2021; Ooi et al., 2022). However, most prediction studies use datasets with fewer than a few hundred participants, leading to low reproducibility and inflated performance (Arbabshirani et al., 2017; Bzdok et al., 2018; Masouleh et al., 2019; Poldrack et al., 2020; Marek et al., 2022). Studies have shown that prediction performance increases with larger sample sizes (Chu et al., 2012; Cui & Gong, 2018; He et al., 2020; Schulz et al., 2020), but for investigations of certain clinical populations or focused neuroscience inquiries, small-scale datasets remain unavoidable. Here, to address this fundamental issue, we seek to establish a framework to translate prediction models from large-scale datasets to predict new non-brain-imaging phenotypes in small-scale datasets based on anatomical T1-weighted images.

More specifically, given a large-scale anatomical MRI dataset ($N > 10{,}000$) with multiple phenotypes, we seek to translate models trained from the large dataset to new unseen phenotypes in a small independent dataset ($N \leqslant 200$). In machine learning, this problem is often referred to as meta-learning, lifelong learning, learning-to-learn or few-shot learning (Fei-Fei et al., 2006; Andrychowicz et al., 2016; Finn et al., 2017; Ravi & Larochelle, 2016; Vanschoren, 2019), and is closely related to transfer learning (Weiss et al., 2016; Hospedales et al., 2021). Broadly speaking, meta-learning and transfer learning methods usually train a model on abundant data on a related problem, called the source dataset, and seek to translate knowledge learned from the large-scale dataset to the small dataset, called the target dataset. During the translation, a subset of the target dataset is typically used to adapt the pre-trained model to the new sample. One distinction between meta-learning and transfer learning is that in transfer learning, the prediction problem in the target dataset can be same (Ghafoorian et al., 2017; Aderghal et al., 2018; Wee et al., 2019) or different (Dawud et al., 2019; Talo et al., 2019; Mehmood et al., 2021) from the source dataset. On the other hand, meta-learning always involves the translation of the prediction model to perform a *new* prediction problem

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

in the target dataset – providing the imaging neuroscience community with a versatile modeling framework that, once established, can be applied to a diversity of research goals.

In our previous study (He et al., 2022), we developed a simple "meta-matching" approach to translate prediction models from large datasets to improve the prediction of new phenotypes in small datasets. Meta-matching is grounded in the observation that many phenotypes are correlated, 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 (Smith et al., 2015; Miller et al., 2016; Xia et al., 2018; Kebets et al., 2019). As a result, a phenotype X in a smaller-scale study is likely correlated, sharing a latent relationship, with a phenotype Y present in a larger population dataset. Therefore, a model trained to predict phenotype Y from the larger dataset might be predisposed to features useful for predicting phenotype X. Consequently, the predictive model of Y can be more effectively translated to predict phenotype X in the smaller study. As a demonstration of meta-matching (He et al., 2022), we trained a simple fully-connected feedforward neural network to predict 67 non-brain-imaging phenotypes from resting-state functional connectivity (RSFC) in the UK Biobank. The neural network 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 in prediction accuracies over classical kernel ridge regression (without meta-learning or transfer learning).

In the current study, we investigated whether the two best performing meta-matching variants ("meta-matching finetune" and "meta-matching stacking") from our previous study (He et al., 2022) can be adapted to work with T1 MRI data. More specifically, given the different modalities (RSFC versus T1), the base neural network architecture was changed from a fully-connected feedforward neural network to the simple fully convolutional network (SFCN; Peng et al., 2021). The SFCN was chosen because of its simplicity and top performance in the Predictive Analysis Challenge 2019 of brain age prediction (Peng et al., 2021). We compared the two meta-matching variants with classical elastic net and classical transfer learning using the UK Biobank (Sudlow et al., 2015; Miller et al., 2016), Human Connectome Project Young Adults (HCP-YA) dataset (Van Essen et al., 2013) and HCP-Aging dataset (Harms et al., 2018; Bookheimer et al., 2019).

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

It is worth mentioning that it is not obvious that meta-matching will confer great benefits in anatomical MRI, compared with RSFC (He et al., 2022). The reason is that RSFC-based prediction typically utilizes high dimensional features derived from N × N RSFC matrices, where N is the number of brain parcels (or independent component analysis components). On the other hand, T1-based prediction can utilize low dimensional N × 1 volumetric and/or thickness measures. Therefore, classical machine learning techniques (e.g., elastic net) might work really well in the small sample regime (≤ 200 participants). Nevertheless, we found that meta-matching significantly outperformed classical elastic net and transfer learning, highlighting the versatility of the meta-matching framework.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2. Methods

2.1. Datasets and preprocessing

In this section, we describe the datasets and preprocessing used in the current study. Table 1 summarizes the demographics and acquisition parameters of the three datasets we considered. We will evaluate meta-matching based on prediction accuracy when translating prediction models within the same dataset (UK Biobank), as well as across datasets, i.e., from UK Biobank to HCP-YA and HCP-Aging datasets. The very different age ranges between HCP-YA and UK Biobank served as a strong test of the generalizability of meta-matching. 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.

	Age	Sex (M/F)	Scanner(s)	Resolution
UK Biobank	45-82	53%/47%	Siemens Skyra 3T scanner	1mm
HCP-YA	22-35	47%/53%	Customized Skyra 3T scanner	0.7mm
HCP-Aging	36-100	44%/56%	Siemens Prisma 3T scanner	0.8mm

Table 1. Summary of demographics and acquisition parameters of the three datasets used in the current study.

2.1.1. UK Biobank

The UK Biobank (UKBB) dataset is a large-scale epidemiology study of over 500,000 adults from the United Kingdom (Alfaro-Almagro et al., 2018). The volunteers were recruited between 2006 and 2010 from 22 centers across the UK. Participants were asked to answer a variety of questionnaires about different aspects of health and lifestyle. In addition, a range of physiological measurements was also collected. We considered the same set of 36,848 participants and 67 non-brain-imaging phenotypes (referred to as phenotypes henceforth; Table S1) from our previous study (He et al., 2022).

As part of the UK Biobank pipeline (Alfaro-Almagro et al., 2018), FreeSurfer recon-all was used to derive thickness and volume measures with the Desikan-Killiany-Tourville (DKT40)

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

cortical atlas (Klein et al., 2012) and subcortical segmentation (Fischl et al., 2002). We considered the subset of regions present in most participants, yielding 164 morphometric measures, comprising intracranial volume (ICV) and thickness measures of 62 cortical regions, as well as volumes of 62 cortical regions and 39 subcortical regions (Tables S4 and S5). After excluding participants who have dropped out since our previous study (He et al., 2022) and excluding participants without all 164 morphometric measures, we ended up with 36,461 participants. As a baseline, these 164 measures will be utilized by the elastic net algorithm for phenotypic prediction (see Section 2.3).

Furthermore, we used FMRIB's Linear Image Registration Tool (FLIRT) to transform the bias-field-corrected version of the brain-extracted T1 (from the UK Biobank provided preprocessing outputs) to MNI152 standard-space T1 template with 1 mm resolution (Jenkinson et al., 2001; Jenkinson et al., 2002). Each T1 image was cropped to dimensions 160 x 192 x 160, and then divided by the mean value within each image following Peng et al. (2021). The normalized T1 images will be used by a convolutional neural network for phenotypic prediction (Section 2.2).

2.1.2. HCP Young Adult (HCP-YA) dataset

We utilized the Human Connectome Project Young Adult (HCP-YA) dataset (Van Essen et al., 2013), which included healthy young adults. We considered 1,019 participants and 35 non-brain-imaging phenotypes, consistent with our previous study (He et al., 2022). The phenotypes are found in Table S2.

FreeSurfer recon-all from the HCP pipeline was used to derive thickness and volume measures with the DKT40 cortical atlas (Klein et al., 2012) and ASEG subcortical segmentation (Fischl et al., 2002). We considered the subset of regions present in most participants, yielding 166 morphometric measures, comprising intracranial volume (ICV) and thickness measures of 62 cortical regions, as well as volumes of 62 cortical regions and 41 subcortical regions (Tables S4 and S5). We note that the difference in the number of morphometric measures between UK Biobank and HCP-YA datasets (164 vs 166) arose because the 5th-Ventricle and non-WM-hypointensities were missing in most participants from the UK Biobank dataset. As a baseline, these 166 measures will be utilized by the elastic net algorithm for phenotypic prediction (see Section 2.2).

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Moreover, we considered T1 images of 0.7 mm resolution which had been transformed to MNI152 space by FLIRT from the HCP PreFreesurfer pipeline. We noted the files of two participants were missing in the HCP filesystem, so we ended up with 1,017 participants. Each T1 image was downsampled to 1mm, cropped to dimensions 160 x 192 x 160, and then divided by the mean value within each image following Peng et al. (2021). The processed T1 images will be used by a convolutional neural network for phenotypic prediction (Section 2.2).

2.1.3. HCP-Aging dataset

Besides the HCP-YA dataset, we also used the Human Connectome Project Aging (HCP-Aging) dataset (Harms et al., 2018; Bookheimer et al., 2019) consisting of healthy participants. We manually selected commonly used non-brain-imaging phenotypic measures across cognition, emotion, motor, sensor, and life experience, resulting in 45 phenotypes (Table S3). By only considering participants with at least 90% of the phenotypes, we ended up with 656 participants (out of 725 participants). Similar to the HCP-YA dataset, we used the same 166 morphometric measures generated by the FreeSurfer recon-all procedure from the HCP pipeline. Moreover, we considered T1 images of 0.8 mm resolution, which had been transformed to MNI152 space by FLIRT from the HCP PreFreesurfer pipeline. Each T1 image was downsampled to 1mm, cropped to dimensions 160 x 192 x 160, and then divided by the mean value within each image following Peng et al. (2021). The processed T1 images will be used by a convolutional neural network for phenotypic prediction (Section 2.2).

2.2. Data split for different analyses

We performed two sets of analyses. First, we benchmarked meta-matching within the UK Biobank. Second, we translated predictive models from the UK Biobank to the HCP-YA and HCP-Aging datasets.

2.2.1. Data split within UK Biobank

For the UK Biobank analysis, we considered 36,461 participants with T1 structural MRI and 67 phenotypes. As illustrated in Figure 1, we randomly split the data into a meta-training set comprising 26,573 participants with 33 phenotypes, as well as a meta-test set comprising

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

9,888 participants with 34 phenotypes. There was no overlap between the participants and phenotypes across the meta-training set and meta-test set.

We further randomly split the meta-training set into a training set with 21,258 participants (80% of 26,573 participants) and a validation set with 5,315 participants (20% of 26,573 participants). The validation set was used for tuning hyperparameters of the predictive models.

For the meta-test set, we randomly split 9,888 participants into K participants (K-shot) and 9,888− K participants, where K had a value of 10, 20, 50, 100, and 200. The group of K participants mimicked traditional small-N studies. Various trained models from the meta-training set were translated to the meta-test set using the K participants. The models were then evaluated using the remaining N – K participants. Each random K-shot split was repeated 100 times to ensure stability.

2.2.2. Data split scheme for cross-dataset analyses

To translate predictive models from the UK Biobank to other datasets, we considered HCP-YA and HCP-Aging datasets. As illustrated in Figure 2, the meta-training set comprised all 36,461 participants with all 67 phenotypes from the UK Biobank dataset. The first meta-test set consisted of 1,017 participants with 35 phenotypes from the HCP-YA dataset. The second meta-test set consisted of 656 participants with 45 phenotypes from the HCP-Aging dataset. There was no overlap between the participants and phenotypes across the meta-training and meta-test sets because they were from totally different datasets. For the meta-training set, we further randomly split it into a training set with 29,169 participants (80% of 36,461 participants) and a validation set with 7,292 participants (20% of 36,461 participants). The validation set was used for tuning hyperparameters of the predictive models.

For the HCP-YA dataset, we randomly split 1,017 participants into K participants (K-shot) and 1,017− K participants, where K had a value of 10, 20, 50, 100, and 200. Various trained models from the meta-training set were translated to the meta-test set using the K participants. The models were then evaluated using the remaining N – K participants. Each random K-shot split was repeated 100 times to ensure stability. The same procedure was applied to the HCP-Aging dataset.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

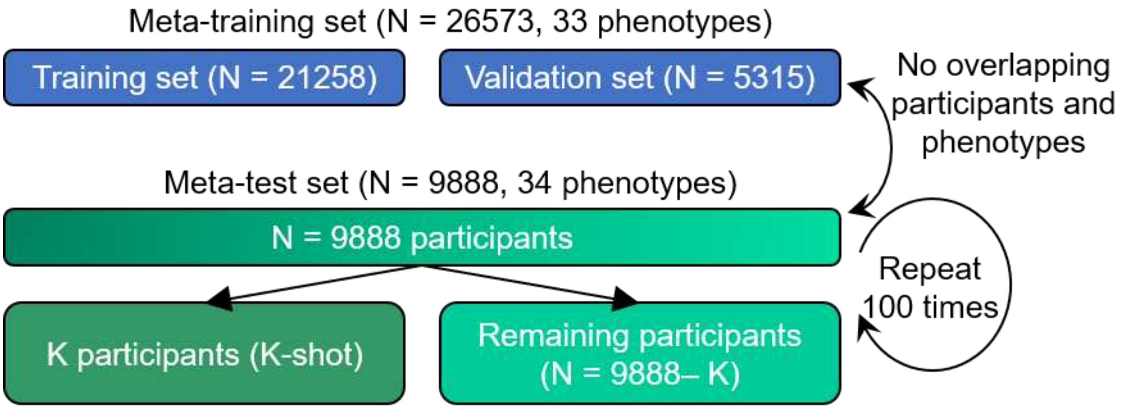

Meta-training set (N = 26573, 33 phenotypes)
Training set (N = 21258)
Validation set (N = 5315)
No overlapping participants and phenotypes
Meta-test set (N = 9888, 34 phenotypes)
N = 9888 participants
Repeat 100 times
K participants (K-shot)
Remaining participants (N = 9888– K)

Figure 1. Data split scheme for within-UK Biobank analysis. The UK Biobank dataset was divided into a meta-training set comprising 26,573 participants and 33 phenotypes, as well as a meta-test set comprising 9,888 participants and 34 other phenotypes. There was no participant or phenotype overlap between meta-training and meta-test sets. The meta-test set was, in turn, split into K participants (K = 10, 20, 50, 100 and 200) and remaining 9,888 − K participants. The group of K participants mimicked studies with traditionally common sample sizes. Various trained models from the meta-training set were translated to the meta-test set using the K participants. The models were then evaluated using the remaining N – K participants. This random split was repeated 100 times for robustness.

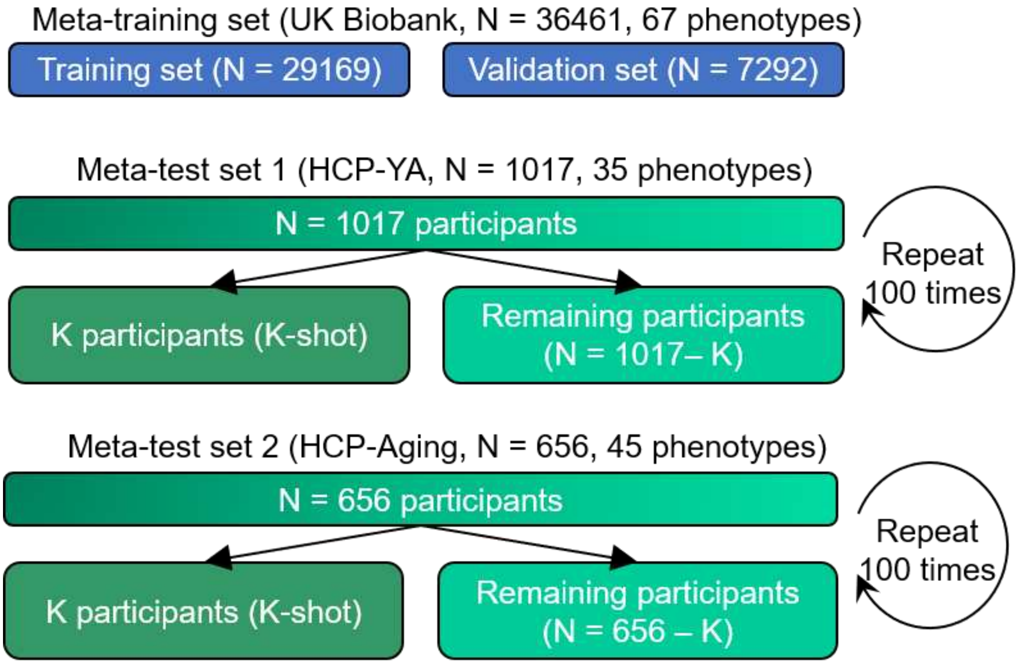

Meta-training set (UK Biobank, N = 36461, 67 phenotypes)
Training set (N = 29169)
Validation set (N = 7292)
Meta-test set 1 (HCP-YA, N = 1017, 35 phenotypes)
N = 1017 participants
Repeat 100 times
K participants (K-shot)
Remaining participants (N = 1017– K)
Meta-test set 2 (HCP-Aging, N = 656, 45 phenotypes)
N = 656 participants
Repeat 100 times
K participants (K-shot)
Remaining participants (N = 656 – K)

Figure 2. **Data split scheme for cross-dataset analysis.** The meta-training set comprised 36,461 UK Biobank participants and 67 phenotypes. The first meta-test set comprised 1,017 HCP-YA participants and 35 phenotypes. The second meta-test set comprised 656 HCP-Aging participants and 45 phenotypes. Each meta-test was, in turn, split into K participants (K = 10, 20, 50, 100 and 200) and remaining participants. The group of K participants mimicked studies with traditionally common sample sizes. Various trained models from the meta-training set were translated to the meta-test set using the K participants. The models

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

were then evaluated using the remaining N – K participants. This random split was repeated 100 times for robustness.

2.3. Predictive models

Figure 3 provides an overview of the different approaches we will compare. Across all approaches, we z-normalize each non-imaging phenotype to have zero mean and unit variance across participants. More specifically, in the case of the meta-training set, the mean and standard deviation were computed using all the participants to apply the z-normalization. In the case of the meta-test set, for each phenotype, the mean and standard deviation were computed from the K participants and subsequently carried over to the full meta-test set comprising the K participants and the remaining N – K test participants.

Following our previous study (He et al., 2022), statistical difference between algorithms was evaluated using a bootstrapping approach (more details in Supplementary Methods S2). More specifically, we will compare both meta-matching variants (Sections 2.3.3 and 2.3.4) with the two baselines (Sections 2.3.1 and 2.3.2). Multiple comparisons were corrected using a false discovery rate (FDR) of $q < 0.05$. FDR was applied to all K-shots, across all comparisons and both evaluation metrics (Pearson's correlation and COD).

2.3.1. Baseline 1: Elastic net

As a baseline, we used thickness and volumetric measures as input features to predict individuals' phenotypes using elastic net (Figure 3). Elastic net is a linear regression model with an L1 lasso and L2 ridge regularization terms (Zou et al., 2005). Here, we chose elastic net as a baseline because previous studies have suggested that elastic net yielded strong prediction performance in phenotypic prediction for brain MRI data (Pervaiz et al., 2020; Peng et al., 2021; Ooi et al., 2022).

Given K participants from the meta-test set, the morphometric (volumetric and thickness) measures were z-normalized based on the mean and standard deviation computed from the K participants. We note that the morphometric measures of the N – K participants were also z-normalized using the mean and standard deviation computed from the K participants. The z-normalized morphometric measures were used as input to train the elastic net model on the K participants.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

More specifically, we performed five-fold cross-validation on the K participants with different combinations of the hyperparameters λ_1 and λ_2 (which controlled the strength of the L1 and L2 regularizations). We used coefficient of determination (COD) to evaluate prediction performance to choose the best hyperparameters for λ_1 and λ_2 across the 5-fold cross-validation.

The best hyperparameters λ_1 and λ_2 were then used to train the elastic net model using all K participants. The trained elastic net model was then applied to the remaining N – K test participants in the meta-test set. Pearson's correlation and the COD were used to evaluate prediction performance. This procedure was repeated for each of the 100 random splits.

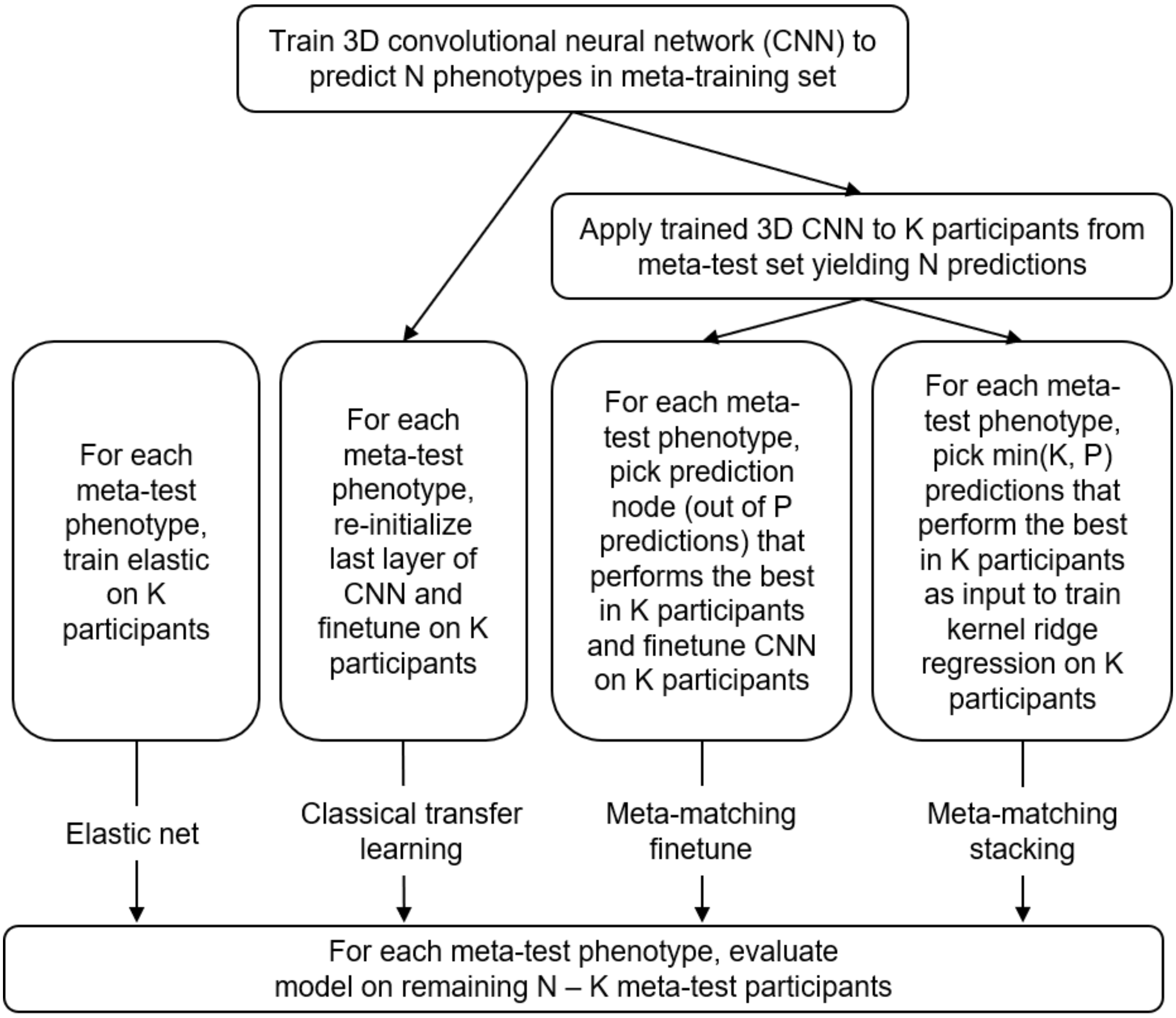

Train 3D convolutional neural network (CNN) to predict N phenotypes in meta-training set
Apply trained 3D CNN to K participants from meta-test set yielding N predictions
For each meta-test phenotype, train elastic on K participants
For each meta-test phenotype, re-initialize last layer of CNN and finetune on K participants
For each meta-test phenotype, pick prediction node (out of P predictions) that performs the best in K participants and finetune CNN on K participants
For each meta-test phenotype, pick min(K, P) predictions that perform the best in K participants as input to train kernel ridge regression on K participants
Elastic net
Classical transfer learning
Meta-matching finetune
Meta-matching stacking
For each meta-test phenotype, evaluate model on remaining N – K meta-test participants

Figure 3. **Overview of different approaches.** We considered two baselines: elastic net and classical transfer learning. We proposed two meta-matching variants: meta-matching finetune and meta-matching stacking.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.3.2. Baseline 2: Classical transfer learning

To perform classical transfer learning, we first trained a simple fully convolutional network (SFCN) introduced by Peng et al., (2021) in the meta-training set to jointly predict all the available meta-training phenotypes.

The input to the SFCN is the mean-normalized T1 image affine transformed to MNI152 standard space (Section 2.1). The SFCN's convolutional neural network (CNN) architecture was based on VGG (Simonyan et al., 2014) and used a fully convolutional structure (Long et al., 2015). We chose the SFCN given its simplicity and top performance in the Predictive Analysis Challenge 2019 of brain age prediction (Peng et al., 2021). In the original study (Peng et al., 2021), the last layer comprised 40 nodes that represented the predicted probability of the age interval that a participant's age falls into. Here, we modified the last layer to predict P phenotypes simultaneously. P is equal to 33 in the within-UK Biobank analysis (Figure 1) and P is equal to 67 in the cross-dataset analysis (Figure 2).

Figure 4 shows the overall network architecture. The 3D CNN consisted of several convolutional blocks for feature extraction. Each feature extraction block (except the last block) consisted of a 3D convolutional layer, a batch normalization layer, a max pooling layer, and a ReLU activation layer. The last block was similar to the previous blocks but without the max pooling layer. The feature maps from the last block were fed into an average pooling layer (green in Figure 4).

Since the elastic net utilized ICV as one of the features, while affine registration of T1 to MNI152 space removed this information, for the comparison to be comparable, we concatenated z-normalized ICV with the outputs of the average pooling layer. More specifically, for both meta-training and meta-test sets, ICV of each participant was z-normalized using the mean and standard deviation computed from the participants of the training set within the meta-training set. The concatenated features were then fed into a dropout layer and then went through a 3D convolution layer with 1x1x1 kernel size to produce the final outputs.

The hyperparameters of the CNN were empirically determined based on the meta-training set from the within UK Biobank analysis (Figure 1A). Both within-UK Biobank and cross-dataset analyses used the same set of hyperparameters. More details about model architecture

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and hyperparameters (for example, the number of blocks, number of channels per block, and kernel size per block) can be found in Supplementary Methods S1.

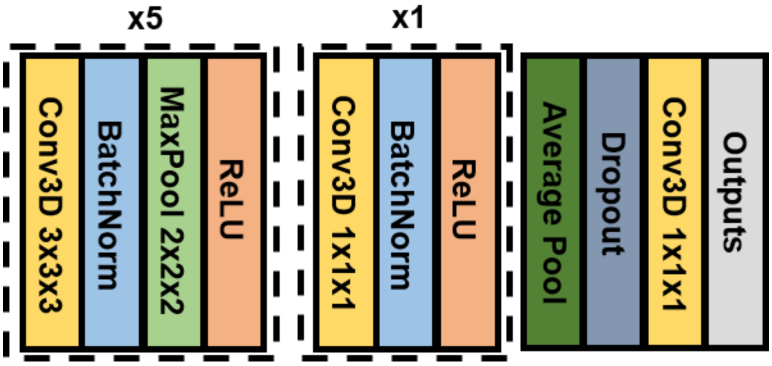


Figure 4. **Network architecture of the Simple Fully Convolutional Neural Network (SFCN) model (Peng et al., 2021) adapted to the current study**. In the original study (Peng et al., 2021), the last layer comprised 40 nodes that represented the predicted probability of the age interval that a participant's age falls into. Here, we modified the last layer to predict P phenotypes simultaneously.

The same transfer learning procedure was used for both within-UK Biobank analysis and cross-dataset analysis (Figure 3). The only difference is that the within-UK Biobank analysis used a CNN model trained on 26,573 participants and 33 phenotypes, while the cross-dataset analysis used a CNN model trained on 36,461 participants and 67 phenotypes.

To perform transfer learning, we first replaced the last layer of the 3D CNN model (trained on the meta-training set) with a new convolutional layer with 1x1x1 kernel size and one output node. The new convolutional layer was initialized with random weights. For each meta-test phenotype, the last two layers of the CNN model were then finetuned on K participants in the meta-test set, while the weights of the remaining layers were frozen.

The optimal learning rate was determined using grid search and five-fold cross-validation on the K participants. The optimal learning rate was then used to perform fine-tune a final model using all K participants. For both the five-fold cross validation and the final round of fine-tuning, the maximum fine-tuning epochs was set to be 50 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 in the meta-test set. Pearson’s correlation and COD were used to evaluate the prediction performance. This procedure was repeated for each of the 100 random splits.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.3.3. Meta-matching finetune

As an alternative to transfer learning, we considered the “meta-matching finetune” approach (Figure 3) introduced in our previous study (He et al., 2022). To explain the meta-matching finetune procedure in the current study, we will focus on the experimental setup for the within-UK Biobank analysis.

Recall from Section 2.3.2 that we have trained a 3D CNN model to predict 33 phenotypes in the meta-training set from the UK Biobank. Given K participants from the meta-test set, we applied the CNN yielding 33 predictions. For each meta-test phenotype (out of 34 phenotypes), we calculated the accuracy (COD) with each of the 33 predictions for the K participants. The output node of the CNN model with the best COD was chosen, while the remaining 32 nodes were removed. The last two layers of the CNN model were finetuned using the K participants, while the weights of the remaining layers were frozen.

Therefore, the difference between meta-matching finetune and classical transfer learning (Section 2.3.2) is the initialization of the last layer. Classical transfer learning randomly initialized the last layer, while meta-matching finetune initialized the last layer by choosing the “closest” phenotypic prediction model from the meta-training set.

The optimal learning rate for finetuning was determined using grid search and five-fold cross-validation on the K participants. The optimal learning rate was then used to perform fine-tune a final model using all K participants. For both the five-fold cross validation and the final round of fine-tuning, the maximum fine-tuning epochs was set to be 50 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 in the meta-test set. Pearson’s correlation and COD were used to evaluate the prediction performance. This procedure was repeated for each of the 100 random splits. The same procedure was used for both within-UK Biobank analysis and cross-dataset analysis (Figure 3). The only difference was that the within-UK Biobank analysis used a CNN model trained on 26,573 participants and 33 phenotypes, while the cross-dataset analysis used a CNN model trained on 36,461 participants and 67 phenotypes.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

2.3.4. Meta-matching stacking

We also considered the meta-matching stacking approach (Figure 3) introduced in our previous study (He et al., 2022). To explain the meta-matching stacking procedure in the current study, we will again focus on the experimental setup for the within-UK Biobank analysis.

Recall from Section 2.3.2 that we have trained a 3D CNN model to predict 33 phenotypes in the meta-training set from the UK Biobank. Given K participants from the meta-test set, we applied the CNN yielding 33 predictions. For each meta-test phenotype (out of 34 phenotypes), we calculated the accuracy (COD) with each of the 33 predictions for the K participants, and selected the top M predictions. The value of M was set to be the minimum of K and 33 to reduce overfitting. For example, when K = 20, then M was set to be 20. When K = 50, then M was set to be 33.

A stacking procedure was then performed (Wolpert, 1992; Breiman, 1996), in which a kernel ridge regression (KRR) model was trained on K participants using the M predictions as input to predict the meta-test phenotype. Similar to our previous study (He et al., 2022), we used the correlation kernel. The hyperparameter λ was tuned using grid search and five-fold cross-validation on the K participants. The optimal λ was then used to train a final KRR model using all K participants.

The trained KRR model was then applied to the remaining N – K participants in the meta-test set. Pearson's correlation and the COD were used to evaluate the prediction performance. This procedure was repeated for each of the 100 random splits.

2.4. Deep neural network implementation

The deep neural network was implemented using PyTorch (Paszke et al., 2017) and computed on NVIDIA RTX 3090 GPUs with CUDA 11.0. More details can be found in Supplementary Methods S1.

2.5. Model interpretation

To illustrate how meta-matching models can be interpreted, similar to our previous study (He et al., 2022), we utilized Haufe transform (Haufe et al., 2014) to interpret the meta-

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

matching stacking prediction of Rey Auditory Verbal Learning Test (RAVLT) score and Montreal Cognitive Assessment (MOCA) in the HCP-Aging dataset for K = 100 participants.

For a predictive model with T1 structural MRI as input and phenotype as output, Haufe transform produces a feature importance value for each voxel. A positive (or negative) predictive feature value implied that higher T1 intensity was related to predicting greater (or lower) phenotypic score.

More specifically, for each phenotype, Haufe transform was calculated as the covariance between the phenotype's prediction based on the meta-matching stacking model and the intensity value of each T1 voxel (across the 100 participants), yielding a 3D volume. The 3D volumes were averaged across the 100 random sampling of 100 participants, and were then visualized in MNI152 space.

2.6. Data and code availability

The code can be found here (https://github.com/ThomasYeoLab/CBIG/XXX). Two co-authors (Lijun An and Chen Zhang) reviewed the code before merging it into the GitHub repository to reduce the chance of coding errors. The trained models for meta-matching are also publicly available (https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/predict_phenotypes/NWULAN2024_MMS).

This study used publicly available data from the UK Biobank (https://www.ukbiobank.ac.uk/), as well as the HCP-YA and HCP-Aging datasets (https://www.humanconnectome.org/). Data can be accessed via data use agreements.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

3. Results

3.1. Meta-matching outperforms elastic net and transfer learning within the UK Biobank

Four approaches (elastic net, classical transfer learning, meta-matching finetune, and meta-matching stacking) were applied to the UK Biobank to predict 34 meta-test phenotypes. The models were trained or adapted based on K participants and then evaluated on the remaining 9,888 – K participants (Figure 1).

Figures 5A and 6A show the Pearson's correlation and COD respectively, averaged across all 34 meta-test phenotypes. Each boxplot represents 100 random samplings of K participants. Figures 5B and 6B show the outcomes of the statistical tests obtained by a bootstrapping procedure (Supplementary Methods S2). The actual p values are reported in Table S6. Colors indicate effect sizes of differences (Cohen's D) between approaches.

In the case of Pearson's correlation (Figure 5), both meta-matching finetune and meta-matching stacking greatly outperformed elastic net and classical transfer learning for all values of K. Meta-matching stacking was statistically better than meta-matching finetune for $K \geq 100$.

In the case of COD (Figure 6), both meta-matching finetune and meta-matching stacking greatly outperformed elastic net for all values of K. For $K \leq 20$, classical transfer learning was numerically better but not statistically better than meta-matching stacking. From $K \geq 50$, meta-matching stacking was numerically better than transfer learning with statistical significance from $K = 100$ onwards.

On the other hand, for $K = 10$, classical transfer learning was numerically better than meta-matching finetune, while meta-matching finetune was better than classical transfer learning for remaining other values of K with large effect sizes (light green in Figure 6B). We note that there was no statistical test between meta-matching finetune and classical transfer learning because of the huge computational cost of the two approaches, so no bootstrapping was performed for either approach.

Another relevant point is that COD for all approaches were negative for $K = 10$. COD was positive for meta-matching finetune for $K = 20$ onwards, and positive for meta-matching

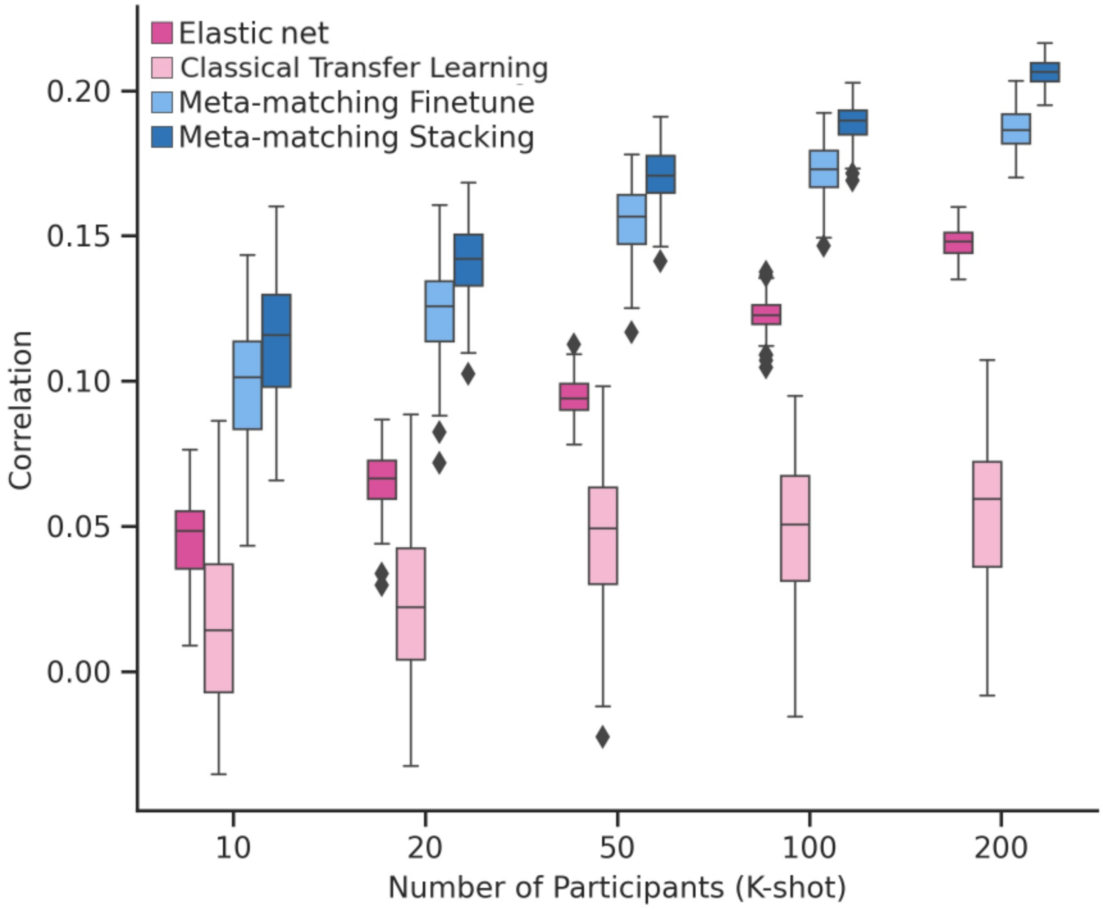
bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

stacking for K = 50 onwards. This suggests that absolute prediction accuracy (i.e., COD) is difficult even with meta-learning or transfer learning, when the sample size is very small.

Overall, meta-matching was better than elastic net for all values of K for both evaluation metrics (Pearson's correlation and COD). On the other hand, meta-matching compared favorably with respect to transfer learning for all values of K for Pearson's correlation and for larger values of K for COD.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(A) Prediction performance



(B) Statistical difference

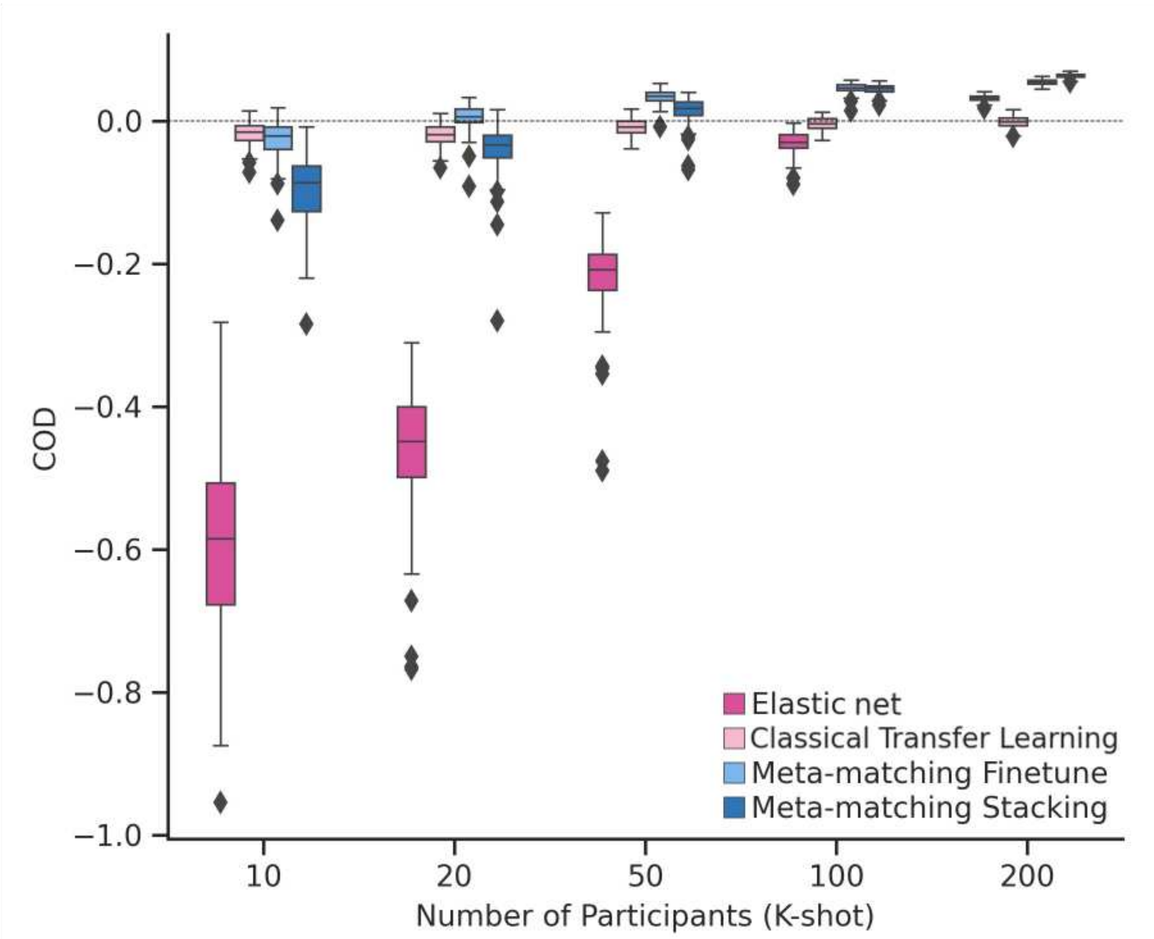
K-shot	10	20	50	100	200
MM stacking vs elastic net	*	*	*	*	*
MM finetune vs elastic net	*	*	*	*	*
MM stacking vs classical transfer learning	*	*	*	*	*
MM finetune vs classical transfer learning	🟩	🟩	🟩	🟩	🟩
MM stacking vs MM finetune	ns	ns	ns	*	*

Figure 5. Meta-matching compared favorably with elastic net and direct transfer learning in terms of Pearson's correlation within the UK Biobank. (A) Phenotypic prediction performance (Pearson's correlation) averaged across 34 meta-test phenotypes in the UK Biobank. X-axis is the number of participants in the meta-test set of the UK Biobank used to train an elastic net or adapt the pretrained model from the meta-training set of the UK Biobank. Each boxplot shows the distribution of performance over 100 repetitions of sampling K participants. (B) Statistical difference between the prediction performance (Pearson's correlation) of baseline methods and meta-matching algorithms. P values were calculated based on a two-sided bootstrapping test. '*' indicates statistical significance after multiple comparisons correction (FDR q < 0.05). 'ns' indicates statistical test did not survive FDR correction. We note that there was no statistical test between meta-matching finetune

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and classical transfer learning because the bootstrapping procedure was too expensive for the two methods. Colors indicate effect sizes of differences (Cohen's D) between approaches. Light green color indicates effect size ≥ 0.8. Dark green color indicates 0 ≤ effect size < 0.8. Dark pink indicates -0.8 < effect size < 0. Light pink color indicates effect size ≤ -0.8. There is no color for the comparison between meta-matching finetune and stacking since they are both our proposed methods.

(A) Prediction performance



(B) Statistical difference

K-shot	10	20	50	100	200
MM stacking vs elastic net	*	*	*	*	*
MM finetune vs elastic net	*	*	*	*	*
MM stacking vs classical transfer learning	ns	ns	ns	*	*
MM finetune vs classical transfer learning	🟪	🟩	🟩	🟩	🟩
MM stacking vs MM finetune	ns	ns	ns	ns	ns

Figure 6. Meta-matching compared favorably with elastic net and direct transfer learning in terms of coefficient of determination (COD) within the UK Biobank. (A)

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Phenotypic prediction performance (COD) averaged across 34 meta-test phenotypes in the UK Biobank. X-axis is the number of participants in the meta-test set of the UK Biobank used to train an elastic net or adapt the pretrained model from the meta-training set of the UK Biobank. Each boxplot shows the distribution of performance over 100 repetitions of sampling K participants. (B) Statistical difference between the prediction performance (COD) of baseline methods and meta-matching algorithms. P values were calculated based on a two-sided bootstrapping test. '*' indicates statistical significance after multiple comparisons correction (FDR $q < 0.05$). 'ns' indicates statistical test did not survive FDR correction. We note that there was no statistical test between meta-matching finetune and classical transfer learning because the bootstrapping procedure was too expensive for the two methods. Colors indicate effect sizes of differences (Cohen's D) between approaches. Light green color indicates effect size ≥ 0.8. Dark green color indicates $0 \leq$ effect size < 0.8. Dark pink indicates $-0.8 <$ effect size < 0. Light pink color indicates effect size ≤ -0.8. There is no color for the comparison between meta-matching finetune and stacking since they are both our proposed methods.

3.2. Meta-matching outperforms baselines in the HCP-YA dataset

The previous experiment results (Figures 5 and 6) suggest that meta-matching can perform well when transferring within the same dataset (UK Biobank). We now evaluate the generalizability of meta-matching across datasets, using the HCP-YA and HCP-Aging datasets (Figure 2) in the following section and next section respectively.

Figures 7A and 8A show the Pearson's correlation and COD respectively, averaged across all 35 meta-test phenotypes in the HCP-YA dataset. Each boxplot represents 100 random samplings of K participants. Figures 7B and 8B show the outcomes of the statistical tests obtained by a bootstrapping procedure (Supplementary Methods S2). The actual p values are reported in Table S7. Colors indicate effect sizes of differences (Cohen's D) between approaches.

In the case of Pearson's correlation (Figure 7), both meta-matching finetune and meta-matching stacking were better than elastic net and classical transfer learning for all values of K with large effect sizes (light green in Figure 7B). Meta-matching finetune was statistically better than elastic net for $K = 20$ and 50. Meta-matching stacking was statistically better than classical transfer learning for $K \geq 20$. For this cross-dataset analysis, meta-matching finetune was generally numerically better, but not statistically better than meta-matching stacking.

In the case of COD (Figure 8), both meta-matching finetune and meta-matching stacking greatly outperformed elastic net for $K \leq 100$. For $K \leq 50$, classical transfer learning was

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

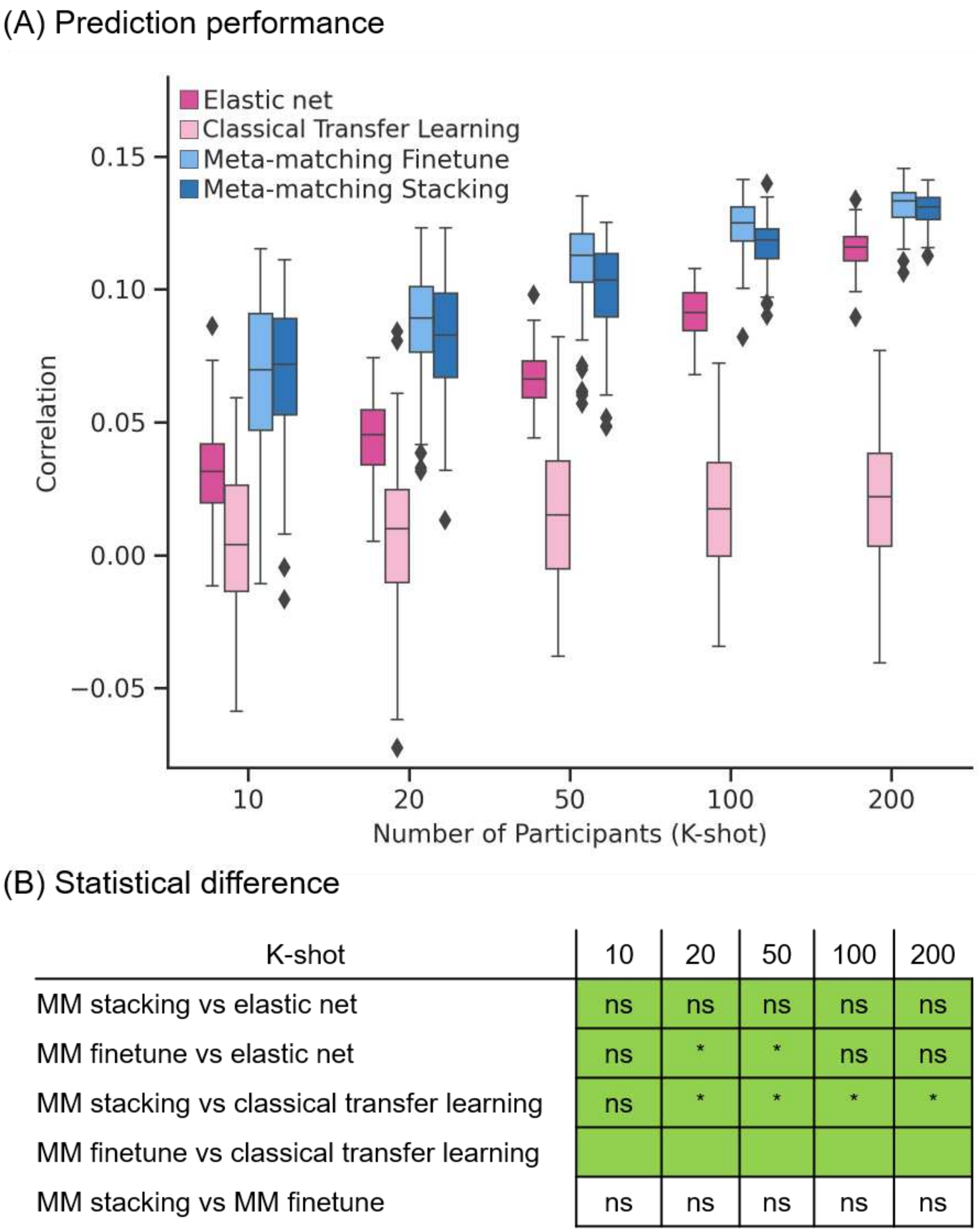
numerically better than meta-matching stacking with large effect sizes (light pink in Figure 8B), but the differences were not significant. From K ≥ 100, meta-matching stacking was numerically better, but not statistically better than transfer learning.

On the other hand, for all values of K, meta-matching finetune was numerically better than classical transfer learning with large effect sizes (light green in Figure 8B). We note that there was no statistical test between meta-matching finetune and classical transfer learning because of the huge computational cost of the two approaches, so no bootstrapping was performed for either approach.

Another relevant point is that COD for all approaches were negative (or almost zero) for K = 10, and only positive for meta-matching finetune for K ≥ 20, suggesting that absolute prediction accuracy (i.e., COD) is difficult even with meta-learning or transfer learning when the sample size is very small.

Overall, meta-matching compared favorably with respect to elastic net for all values of K for both evaluation metrics (Pearson's correlation and COD). On the other hand, meta-matching compared favorably with respect to transfer learning for all values of K for Pearson's correlation and for K ≥ 100 for COD.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

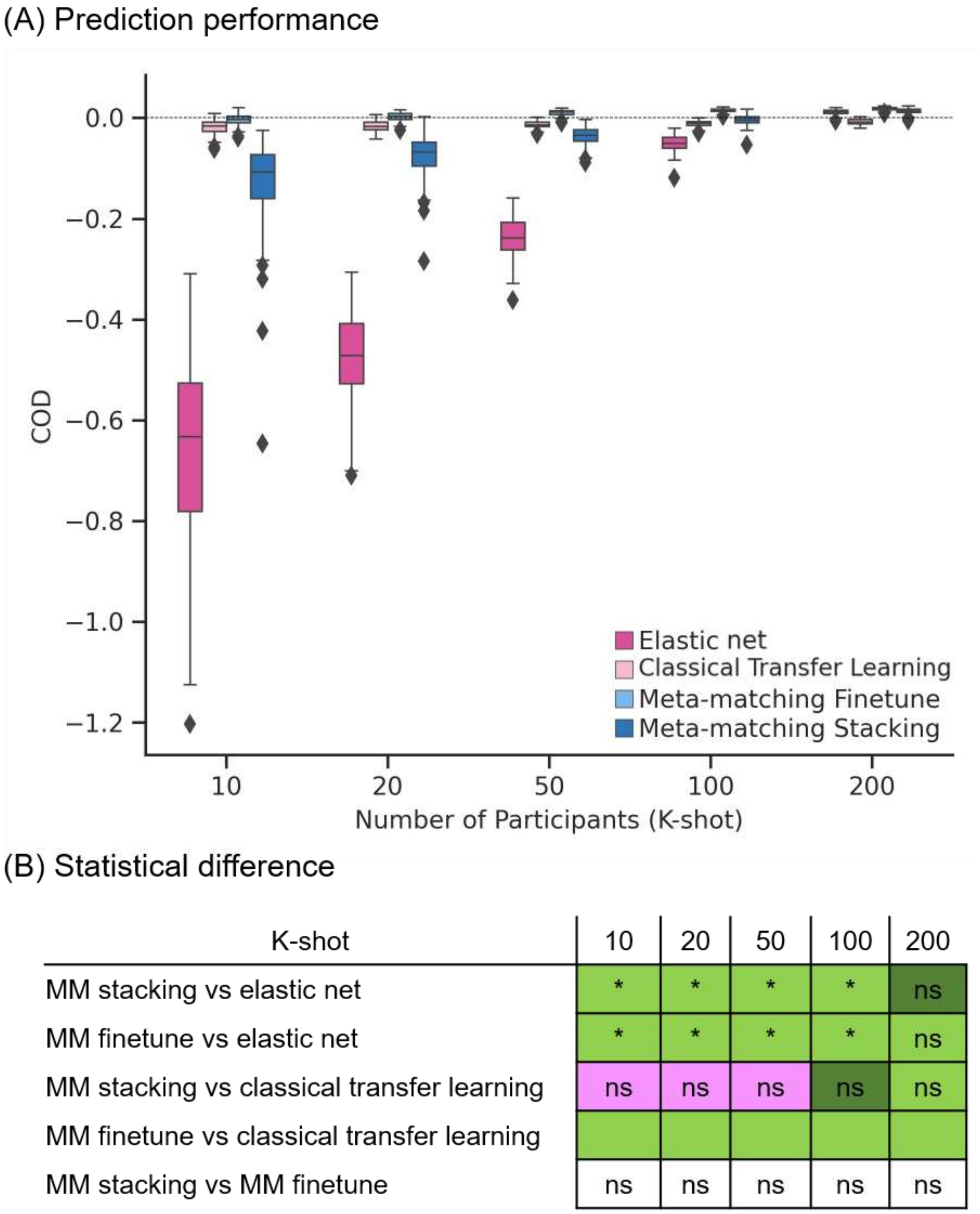


K-shot	10	20	50	100	200
MM stacking vs elastic net	ns	ns	ns	ns	ns
MM finetune vs elastic net	ns	*	*	ns	ns
MM stacking vs classical transfer learning	ns	*	*	*	*
MM finetune vs classical transfer learning					
MM stacking vs MM finetune	ns	ns	ns	ns	ns

Figure 7. **Meta-matching compared favorably with elastic net and direct transfer learning in terms of Pearson's correlation when translating models from UK Biobank to HCP-YA dataset.** (A) Phenotypic prediction performance (Pearson's correlation) averaged across 35 meta-test phenotypes in the HCP-YA dataset. X-axis is the number of participants from the HCP-YA dataset used to train an elastic net or adapt the pretrained model from the meta-training set. Each boxplot shows the distribution of performance over 100 repetitions of sampling K participants. (B) Statistical difference between the prediction performance (Pearson's correlation) of baseline methods and meta-matching algorithms. P values were calculated based on a two-sided bootstrapping test. '*' indicates statistical significance after

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

multiple comparisons correction (FDR q < 0.05). 'ns' indicates statistical test did not survive FDR correction. We note that there was no statistical test between meta-matching finetune and classical transfer learning because the bootstrapping procedure was too expensive for the two methods. Colors indicate effect sizes of differences (Cohen's D) between approaches. Light green color indicates effect size ≥ 0.8. Dark green color indicates 0 ≤ effect size < 0.8. Dark pink indicates -0.8 < effect size < 0. Light pink color indicates effect size ≤ -0.8. There is no color for the comparison between meta-matching finetune and stacking since they are both our proposed methods.



K-shot	10	20	50	100	200
MM stacking vs elastic net	*	*	*	*	ns
MM finetune vs elastic net	*	*	*	*	ns
MM stacking vs classical transfer learning	ns	ns	ns	ns	ns
MM finetune vs classical transfer learning					
MM stacking vs MM finetune	ns	ns	ns	ns	ns

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 8. **Meta-matching compared favorably with elastic net and direct transfer learning in terms of coefficient of determination (COD) when translating models from UK Biobank to HCP-YA dataset.** (A) Phenotypic prediction performance (COD) averaged across 35 meta-test phenotypes in the HCP-YA dataset. X-axis is the number of participants from the HCP-YA dataset used to train an elastic net or adapt the pretrained model from the meta-training set. Each boxplot shows the distribution of performance over 100 repetitions of sampling K participants. (B) Statistical difference between the prediction performance (COD) of baseline methods and meta-matching algorithms. P values were calculated based on a two-sided bootstrapping test. '*' indicates statistical significance after multiple comparisons correction (FDR $q < 0.05$). 'ns' indicates statistical test did not survive FDR correction. We note that there was no statistical test between meta-matching finetune and classical transfer learning because the bootstrapping procedure was too expensive for the two methods. Colors indicate effect sizes of differences (Cohen's D) between approaches. Light green color indicates effect size ≥ 0.8. Dark green color indicates 0 ≤ effect size < 0.8. Dark pink indicates -0.8 < effect size < 0. Light pink color indicates effect size ≤ -0.8. There is no color for the comparison between meta-matching finetune and stacking since they are both our proposed methods.

3.3. Meta-matching outperforms baselines in the HCP-Aging dataset

Figures 9A and 10A show the Pearson's correlation and COD respectively, averaged across all 45 meta-test phenotypes in the HCP-Aging dataset. Each boxplot represents 100 random samplings of K participants. Figures 9B and 10B show the outcomes of the statistical tests obtained by a bootstrapping procedure (Supplementary Methods S2). The actual p values are reported in Table S8. Colors indicate effect sizes of differences (Cohen's D) between approaches.

In the case of Pearson's correlation (Figure 9), both meta-matching finetune and meta-matching stacking greatly outperformed elastic net and classical transfer learning for most values of K. Meta-matching stacking was statistically better than elastic net for K ≥ 20. Meta-matching stacking was statistically better than classical transfer learning for all values of K. For this cross-dataset analysis, meta-matching stacking was numerically better, but not statistically better than meta-matching finetune.

In the case of COD (Figure 10), both meta-matching finetune and meta-matching stacking greatly outperformed elastic net for all values of K. For K ≤ 20, classical transfer learning was numerically better, but not statistically better than meta-matching stacking. From K ≥ 50, meta-matching stacking was numerically better than transfer learning with statistical significance achieved for K ≥ 100.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

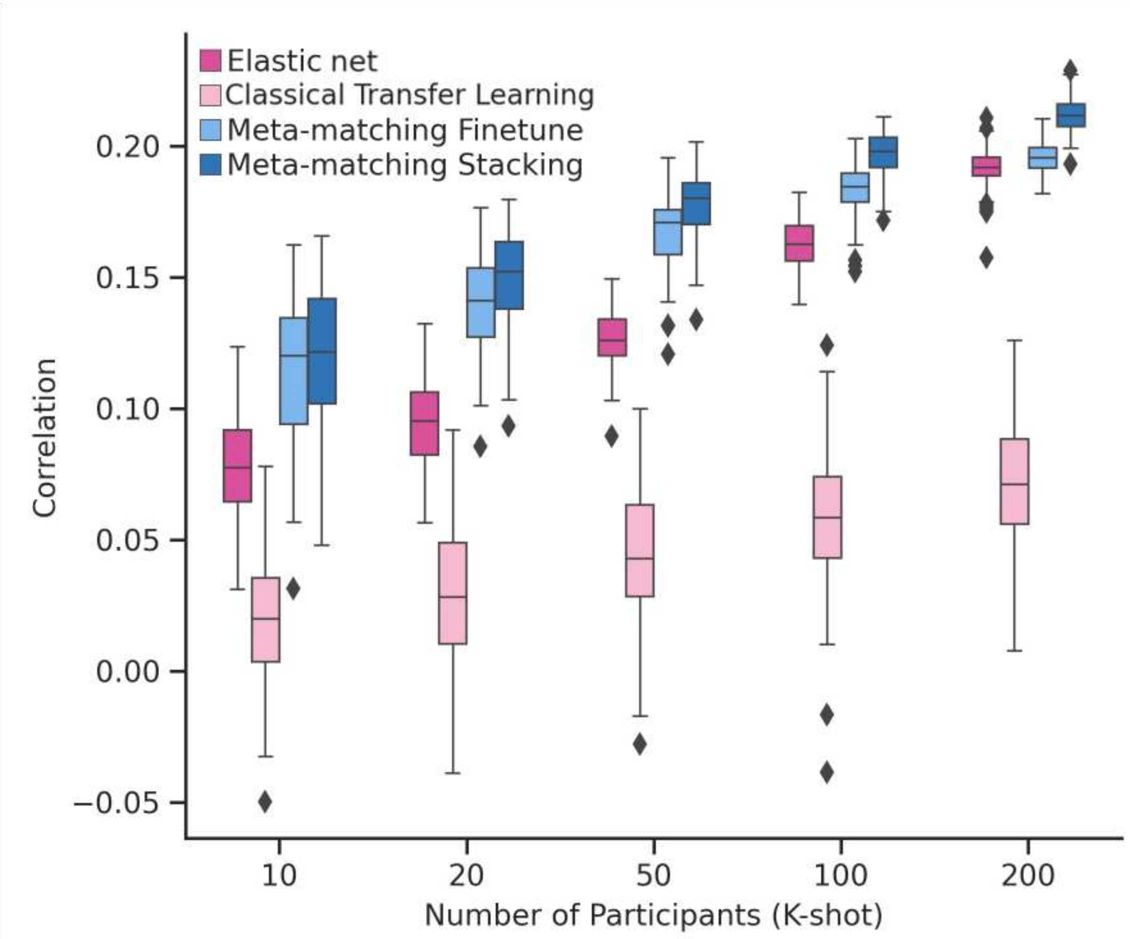
On the other hand, for all values of K, meta-matching finetune was numerically better than classical transfer learning with large effect sizes (light green in Figure 10B). We note that there was no statistical test between meta-matching finetune and classical transfer learning because of the huge computational cost of the two approaches, so no bootstrapping was performed for either approach.

Another relevant point is that COD for all approaches were negative (or almost zero) for $K = 10$, and only positive for meta-matching finetune for $K \geq 20$, suggesting that absolute prediction accuracy (i.e., COD) is difficult even with meta-learning or transfer learning when the sample size is very small.

Overall, meta-matching was better than elastic net for all values of K for both evaluation metrics (Pearson's correlation and COD). On the other hand, meta-matching compared favorably with respect to transfer learning for all values of K for Pearson's correlation and for $K \geq 50$ for COD.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

(A) Prediction performance



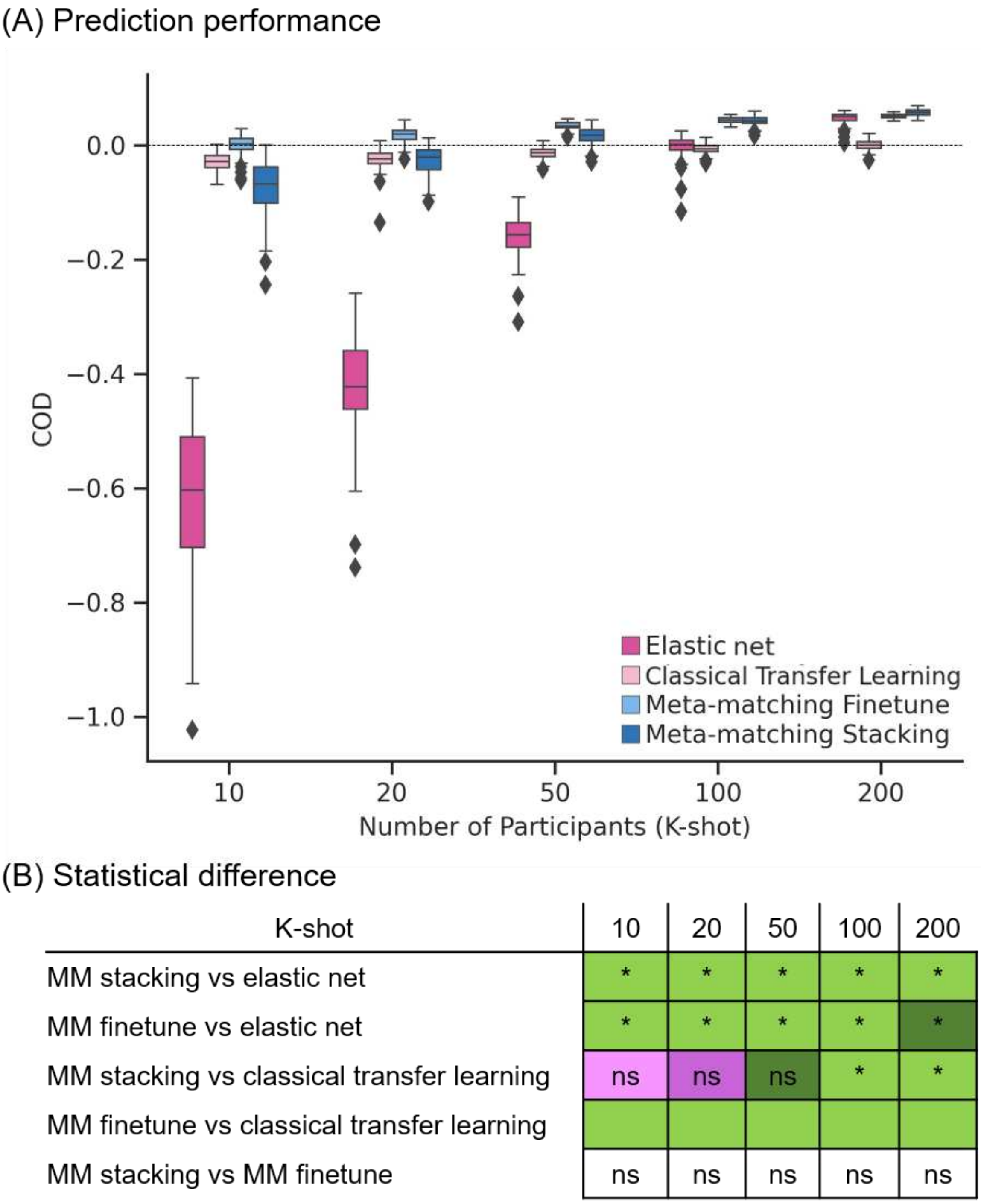
(B) Statistical difference

K-shot	10	20	50	100	200
MM stacking vs elastic net	ns	*	*	*	*
MM finetune vs elastic net	ns	*	*	*	*
MM stacking vs classical transfer learning	*	*	*	*	*
MM finetune vs classical transfer learning	🟩	🟩	🟩	🟩	🟩
MM stacking vs MM finetune	ns	ns	ns	ns	ns

Figure 9. **Meta-matching compared favorably with elastic net and direct transfer learning in terms of Pearson's correlation when translating models from UK Biobank to HCP-Aging dataset.** (A) Phenotypic prediction performance (Pearson's correlation) averaged across 45 meta-test phenotypes in the HCP-Aging dataset. X-axis is the number of participants from the HCP-Aging dataset used to train an elastic net or adapt the pretrained model from the meta-training set. Each boxplot shows the distribution of performance over 100 repetitions of sampling K participants. (B) Statistical difference between the prediction performance (Pearson's correlation) of baseline methods and meta-matching algorithms. P values were calculated based on a two-sided bootstrapping test. '*' indicates statistical significance after multiple comparisons correction (FDR q < 0.05). 'ns' indicates statistical test did not survive FDR correction. We note that there was no statistical test between meta-

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

matching finetune and classical transfer learning because the bootstrapping procedure was too expensive for the two methods. Colors indicate effect sizes of differences (Cohen's D) between approaches. Light green color indicates effect size ≥ 0.8. Dark green color indicates 0 ≤ effect size < 0.8. Dark pink indicates -0.8 < effect size < 0. Light pink color indicates effect size ≤ -0.8. There is no color for the comparison between meta-matching finetune and stacking since they are both our proposed methods.



K-shot	10	20	50	100	200
MM stacking vs elastic net	*	*	*	*	*
MM finetune vs elastic net	*	*	*	*	*
MM stacking vs classical transfer learning	ns	ns	ns	*	*
MM finetune vs classical transfer learning	🟩	🟩	🟩	🟩	🟩
MM stacking vs MM finetune	ns	ns	ns	ns	ns

Figure 10. **Meta-matching compared favorably with elastic net and direct transfer learning in terms of coefficient of determination (COD) when translating models from UK Biobank to HCP-Aging dataset.** (A) Phenotypic prediction performance (COD)

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

averaged across 45 meta-test phenotypes in the HCP-Aging dataset. X-axis is the number of participants from the HCP-Aging dataset used to train an elastic net or adapt the pretrained model from the meta-training set. Each boxplot shows the distribution of performance over 100 repetitions of sampling K participants. (B) Statistical difference between the prediction performance (COD) of baseline methods and meta-matching algorithms. P values were calculated based on a two-sided bootstrapping test. '*' indicates statistical significance after multiple comparisons correction (FDR $q < 0.05$). 'ns' indicates statistical test did not survive FDR correction. We note that there was no statistical test between meta-matching finetune and classical transfer learning because the bootstrapping procedure was too expensive for the two methods. Colors indicate effect sizes of differences (Cohen's D) between approaches. Light green color indicates effect size ≥ 0.8. Dark green color indicates 0 ≤ effect size < 0.8. Dark pink indicates -0.8 < effect size < 0. Light pink color indicates effect size ≤ -0.8. There is no color for the comparison between meta-matching finetune and stacking since they are both our proposed methods.

3.4. Different improvements on different phenotypes

Overall, meta-matching improved prediction on average across multiple phenotypes. However, we note that the improvement was not uniform across phenotypes. Figure 11 illustrates the prediction performance (Pearson's correlation) of three non-brain-imaging phenotypes for K = 100 participants. In the case of the HCP-YA dataset (Figure 11A), meta-matching finetune compared favorably with other approaches for predicting dexterity and language, but only achieved similar prediction accuracy on emotion. In the case of HCP-Aging dataset (Figure 11B), meta-matching stacking compared favorably with other approaches for predicting fear somatic and anger aggression, but only achieved similar prediction accuracy on perceived rejection.

Given that meta-matching exploits correlations among phenotypes, we hypothesized that variability in prediction improvements was driven by inter-phenotype correlations between the meta-training and meta-test sets. Figure 12 shows the performance improvement (Pearson's correlation) of meta-matching stacking as a function of the maximum correlation between each meta-test phenotype and meta-training phenotype in the within-UK Biobank analysis. As expected, meta-test phenotypes with stronger correlations with at least one meta-training phenotype led to greater prediction improvement with meta-matching. We note that this analysis required meta-training and meta-test phenotypes to be present in the same participants, so could only be performed for the within-UK Biobank analysis.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

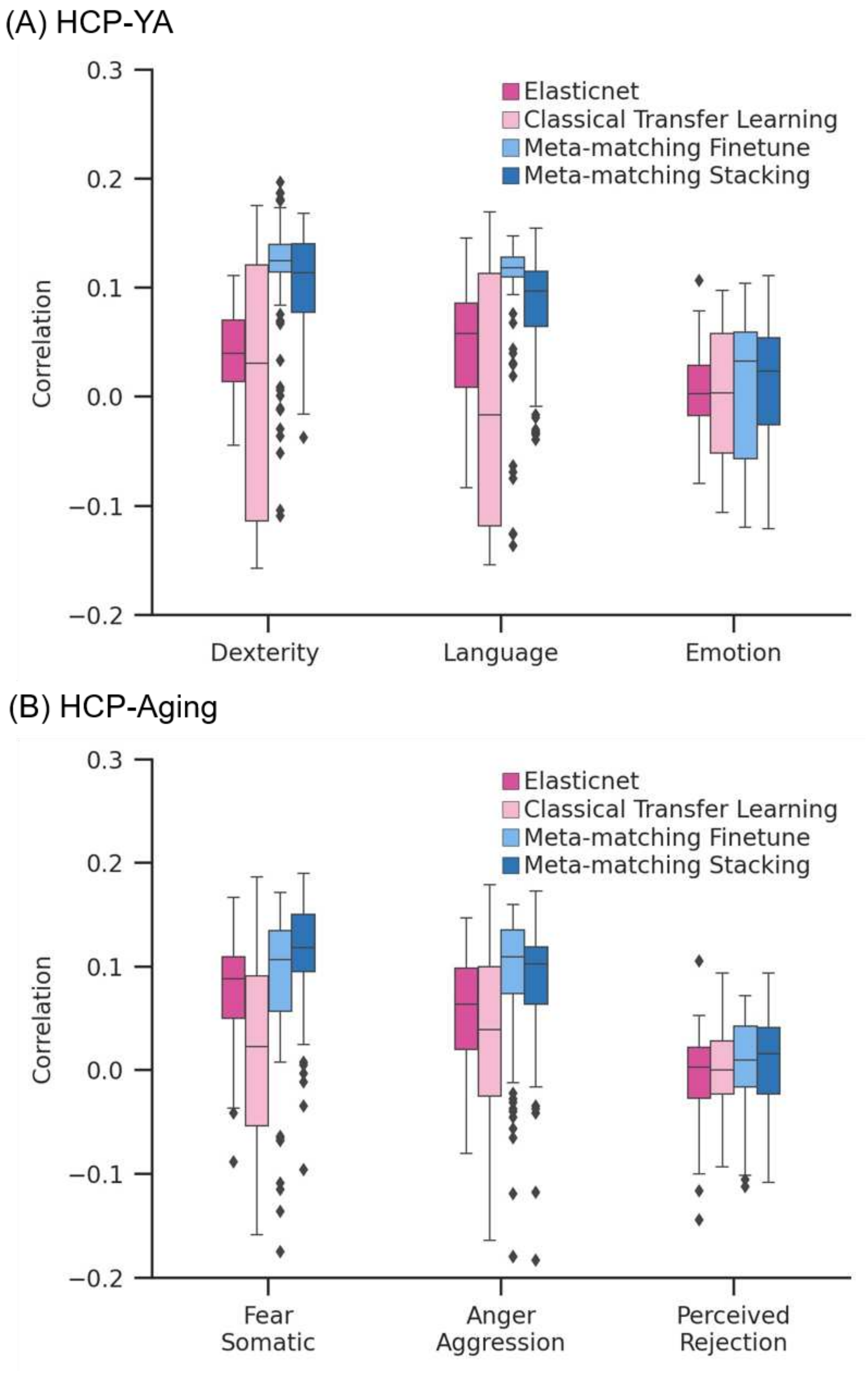


Figure 11. Examples of prediction performance (Pearson's correlation) for different non-brain-imaging phenotypes in the (A) HCP-YA and (B) HCP-Aging datasets in the case of K = 100 participants.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

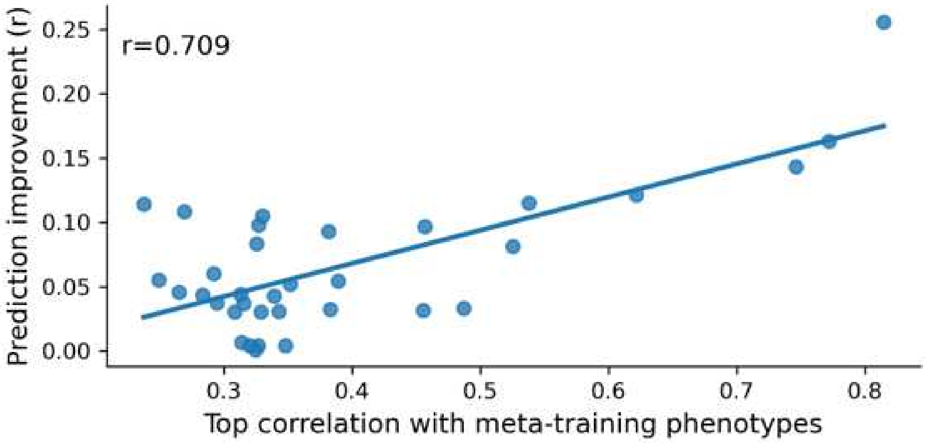


Figure 12. Prediction improvements were driven by correlations between meta-training and meta-test phenotypes. Vertical axis shows the prediction improvement of meta-matching stacking with respect to elastic net baseline under the 100-shot scenario. Prediction performance was measured using Pearson's correlation. Each dot represents a meta-test phenotype. Horizontal axis shows each test phenotype's top absolute Pearson's correlation with phenotypes in the meta-training set. Test phenotypes with stronger correlations with at least one training phenotype led to greater prediction improvement with meta-matching.

3.5. Interpreting meta-matching stacking with Haufe transform

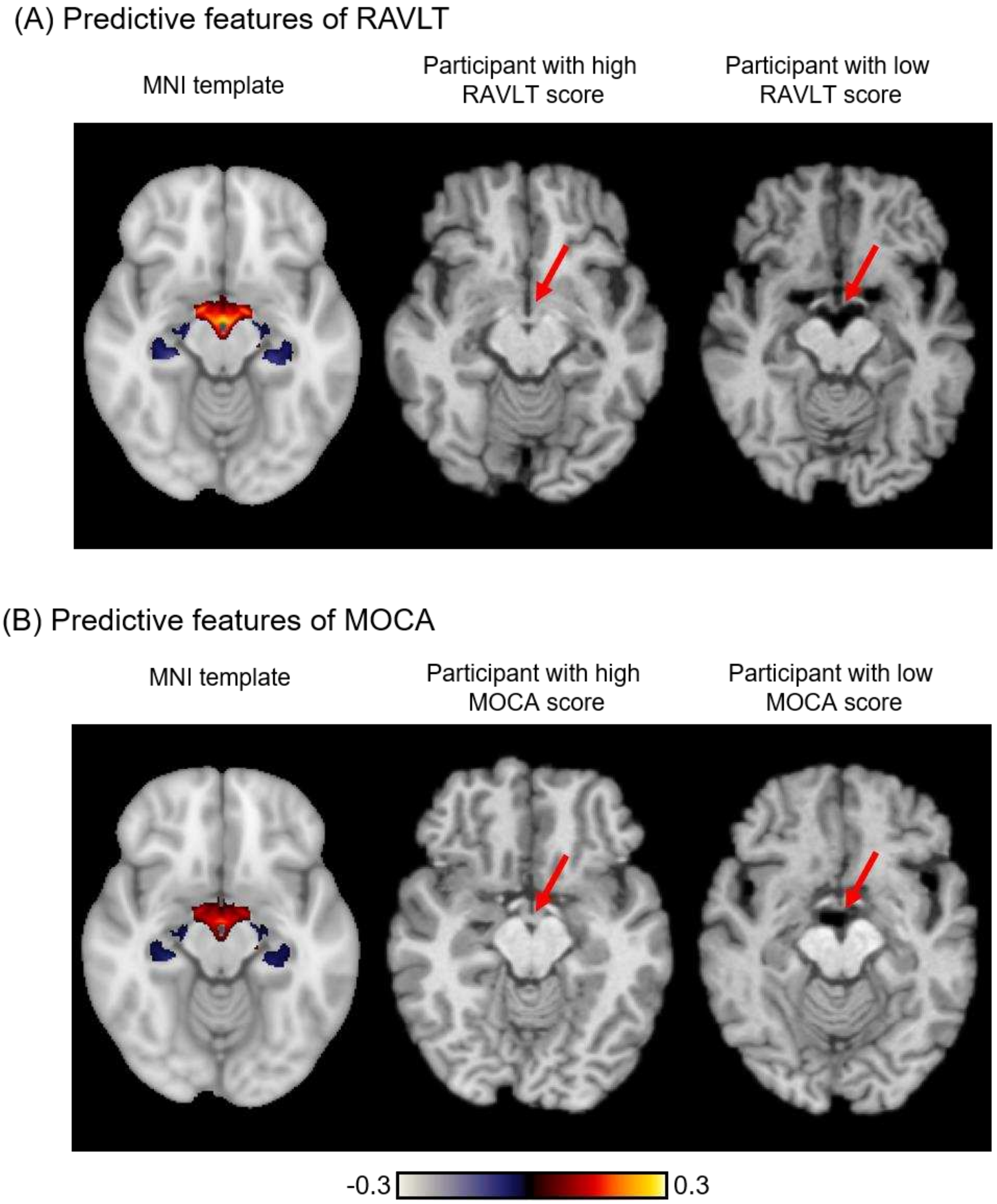
Figure 13 illustrates the feature importance maps obtained from the Haufe transform for predicting the Rey Auditory Verbal Learning Test (RAVLT) score and Montreal Cognitive Assessment (MOCA) in the HCP-Aging dataset for K = 100. We note that a higher RAVLT or MOCA scores indicated better cognition.

Since we are using T1 intensity for prediction, linking the feature importance values to the underlying biology needs to be done with care. For both RAVLT and MOCA, positive feature importance values were observed in the ventral diencephalon and the third ventricle (left panels of Figures 9A and 9B), which suggested that higher T1 value led to prediction of better cognition (higher RAVLT and MOCA scores). By observing participants who performed poorly (right panels of Figures 9A and 9B) and participants who performed well (middle panels of Figures 9A and 9B), we inferred that the prediction might be partially driven by enlarged ventricles in participants with worse cognition (arrows in Figure 9), yielding a lower T1 value in the region.

Similarly, we observed negative feature importance values at or near the left and right hippocampi, which suggested that higher T1 value led to prediction of worse cognition (lower RAVLT and MOCA scores). By observing participants who performed poorly (right panels

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

of Figures 9A and 9B) and participants who performed well (middle panels of Figures 9A and 9B), we inferred that the prediction might be partially driven by gray matter loss at or near the hippocampi, yielding a higher T1 value in the region, consistent with the aging literature (Apostolova et al., 2012; Ritter et al., 2017).



(A) Predictive features of RAVLT

MNI template
Participant with high RAVLT score
Participant with low RAVLT score

(B) Predictive features of MOCA

MNI template
Participant with high MOCA score
Participant with low MOCA score
-0.3
0.3

Figure 13. Feature importance of meta-matching stacking in the HCP-Aging dataset for K = 100 participants. (A) Feature importance map of meta-matching stacking from predicting Rey Auditory Verbal Learning Test (RAVLT) score. Left panel shows the feature importance map on the MNI152 template. A positive (or negative) feature importance value

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

indicates that higher intensity was associated with predicting greater (or lower) phenotypic values. Middle panel shows an example participant with high RAVLT score. Right panel shows an example participant with low RAVLT score. (B) Feature importance map of meta-matching stacking from predicting the Montreal Cognitive Assessment (MOCA) score. Left panel shows the feature importance map on the MNI152 template. Middle panel shows an example participant with high MOCA score. Right panel shows an example participant with low MOCA score.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4. Discussion

In this study, we adapted two meta-matching variants from our previous study (He et al., 2022) to translate prediction models trained from large-scale T1-weighted anatomical MRI datasets to predict new non-brain-imaging phenotypes in small-scale T1-weighted anatomical MRI datasets. We demonstrated that meta-matching finetune and meta-matching stacking greatly outperformed classical elastic net and classical transfer learning when the number of participants ≤ 200. Meta-matching performed well even when translating from a large-scale dataset (UK Biobank) to a small dataset (HCP-YA or HCP-Aging) with different scanners, acquisition, demographics, and pre-processing.

Across all analyses in the UK Biobank, HCP-YA and HCP-Aging datasets (Figures 5 to 10), meta-matching consistently outperformed elastic net across both evaluation metrics (correlation and COD). It is worth noting that the elastic net utilized thickness and volumetric measures generated by FreeSurfer, instead of the intensity values of T1 images (like meta-matching and transfer learning). Given that we are working in the small sample regime with K ≤ 200 training participants, we believe that the small number of less than 200 pre-defined morphometric features together with elastic net provides a powerful classical baseline.

When using Pearson's correlation as an evaluation metric, transfer learning performed poorly with substantially worse performance than both meta-matching variants and even elastic net (Figures 5, 7 and 9). On the other hand, when using COD as an evaluation metric, transfer learning was more competitive with respect to the other approaches (Figures 6, 8 and 10). More specifically, transfer learning was numerically better (but not statistically better) than meta-matching stacking for small values of K, while meta-matching stacking was numerically better (and sometimes statistically better) than transfer learning for larger values of K.

On the other hand, meta-matching finetune outperformed transfer learning for most values of K even in the case of COD. We note that meta-matching finetune is similar to classical transfer learning in the sense that the last two layers of the CNN were finetuned. However, while transfer learning initialized the last layer of the CNN from scratch (Section 2.3.2), meta-matching finetune retained the weights leading to the output node that predicted the K meta-test participants the best (for each meta-test phenotype). This further supported the importance of the meta-matching approach.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Overall, meta-matching stacking was the best for the Pearson's correlation metric, while meta-matching finetune was the best for COD. Pearson's correlation is a measure of relative prediction performance, while COD is a measure of absolute prediction performance (Finn et al., 2015; Scheinost et al., 2019; Poldrack et al., 2020). Therefore, researchers more focused on relative prediction performance might consider using meta-matching stacking, while researchers more focused on absolute prediction performance might consider using meta-matching finetune. Furthermore, all approaches achieved negative or close to zero COD when $K \leq 20$, suggesting that absolute prediction remains out of reach in the very small sample regime.

Meta-matching models can be interpreted at the level of imaging features by using the Haufe transform (Haufe et al., 2014). To illustrate this procedure, we applied the Haufe transform (Haufe et al., 2014) to the translated meta-matching stacking models in the HCP-Aging dataset (Figure 13). For a given meta-test phenotype, Haufe transform was calculated as the covariance between the phenotype's prediction based on the meta-matching stacking model and the intensity value of each T1 voxel (across K participants), yielding a 3D volume. We found that poorer cognitive performance in terms of worse RAVLT and MOCA scores were related to greater gray matter atrophy and larger ventricular size, which is consistent with the aging literature (Apostolova et al., 2012; Ritter et al., 2017). Meta-matching finetune can be interpreted in a similar fashion.

In addition to interpreting meta-matching models at the level of brain-imaging features, the meta-matching models can also be interpreted at the level of phenotypic traits. In the case of meta-matching stacking, this can again be achieved using the Haufe transform. To illustrate this, let us consider the pretrained 3D CNN model from the UK Biobank with 67 prediction outputs. This 3D CNN model can be translated to predict a new meta-test phenotype using K participants from the meta-test set using the stacking procedure. The Haufe transform can then be calculated as the covariance (across the K participants) between the phenotype's prediction from the final stacking model and the 67 inputs to the stacking model, yielding a vector of length 67, which indicates the relative importance of the original 67 meta-training phenotypes for predicting the meta-test phenotype.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Because meta-matching exploits correlations between the phenotypes of meta-training and meta-test sets, the amount of prediction improvement strongly relied on the strongest correlations between the meta-test phenotype and meta-training phenotypes (Figure 12). Consequently, not all phenotypes might benefit from meta-matching. However, we note that this limitation exists for all meta-learning and transfer learning algorithms – model transfer is easier if the source and target domains are more similar; performance will degrade if source and target domains are very different.

In future work, we seek to aggregate more and diverse large-scale population-level datasets targeting different populations, including children, lifespan and a variety of disorders. These additional datasets will not only provide a larger diversity of phenotypic measures but will also allow greater sampling of different MRI scanners, acquisition parameters and demographics. By pretraining on a wider range of datasets, we hope to see further improvements in the meta-matching approaches.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

5. Conclusion

In this study, we showed that meta-matching can be used to translate T1-based phenotypic prediction models from large source datasets to predict new phenotypes in small target datasets. By exploiting correlations between phenotypes, meta-matching greatly outperformed elastic net and classical transfer learning, both when translating models within the same dataset, as well as translating models across datasets with different MRI scanners, acquisition protocols and demographics. Overall, our results demonstrated the versatility of the meta-matching framework.

6. Acknowledgements

Our research is 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 (CTGIIT23jan-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 authors and do not reflect the views of the Singapore NRF, NMRC or MOH. This research has been conducted using the UK Biobank Resource under application number 25163, and the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research, and the National Institute On Aging of the National Institutes of Health under Award Number U01AG052564 and by funds provided by the McDonnell Center for Systems Neuroscience at Washington University in St. Louis. The HCP-Aging 2.0 Release data used in this report came from DOI: 10.15154/1520707.; and by the McDonnell Center for Systems Neuroscience at Washington University.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Author Contribution

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

Competing Interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

Aderghal, K., Khvostikov, A., Krylov, A., Benois-Pineau, J., Afdel, K., & Catheline, G. (2018). *Classification of Alzheimer disease on imaging modalities with deep CNNs using cross-modal transfer learning.* Paper presented at the 2018 IEEE 31st international symposium on computer-based medical systems (CBMS).

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L., Griffanti, L., Douaud, G., . . . Vallee, E. (2018). Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *NeuroImage, 166*, 400-424.

Apostolova, L. G., Green, A. E., Babakchanian, S., Hwang, K. S., Chou, Y.-Y., Toga, A. W., & Thompson, P. M. (2012). Hippocampal atrophy and ventricular enlargement in normal aging, mild cognitive impairment and Alzheimer's disease. *Alzheimer disease and associated disorders, 26*(1), 17.

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.

Bhagwat, N., Pipitone, J., Voineskos, A. N., Chakravarty, M. M., & Initiative, A. s. D. N. (2019). An artificial neural network model for clinical score prediction in Alzheimer disease using structural neuroimaging measures. *Journal of Psychiatry and Neuroscience, 44*(4), 246-260.

Bookheimer, S. Y., Salat, D. H., Terpstra, M., Ances, B. M., Barch, D. M., Buckner, R. L., . . . Elam, J. S. (2019). The lifespan human connectome project in aging: an overview. *NeuroImage, 185*, 335-348.

Breiman, L. (1996). Stacked regressions. *Machine Learning, 24*, 49-64.

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging, 3*(3), 223-230 https://doi.org/10.1016/j.bpsc.2017.11.007

Cohen, S. E., Zantvoord, J. B., Wezenberg, B. N., Bockting, C. L., & van Wingen, G. A. (2021). Magnetic resonance imaging for individual prediction of treatment response in major depressive disorder: a systematic review and meta-analysis. *Translational psychiatry, 11*(1), 168.

Dawud, A. M., Yurtkan, K., & Oztoprak, H. (2019). Application of deep learning in neuroradiology: brain haemorrhage classification using transfer learning. *Computational Intelligence and Neuroscience, 2019*.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., . . . Constable, R. T. (2015). Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature neuroscience, 18*(11), 1664-1671.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., . . . Klaveness, S. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron, 33*(3), 341-355.

Genon, S., Eickhoff, S. B., & Kharabian, S. (2022). Linking interindividual variability in brain structure to behaviour. *Nature Reviews Neuroscience, 23*(5), 307-318.

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., . . . Van Ginneken, B. (2017). *Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation.* Paper presented at the Medical Image Computing and Computer Assisted Intervention− MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20.

Gifford, G., McCutcheon, R., & McGuire, P. (2020). Neuroimaging studies in people at clinical high risk for psychosis. In *Risk Factors for Psychosis* (pp. 167-182): Elsevier.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Harms, M. P., Somerville, L. H., Ances, B. M., Andersson, J., Barch, D. M., Bastiani, M., . . . Burgess, G. C. (2018). Extending the Human Connectome Project across ages: Imaging protocols for the Lifespan Development and Aging projects. *NeuroImage, 183*, 972-984.

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.

He, T., An, L., Chen, P., Chen, J., Feng, J., Bzdok, D., . . . 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.

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.

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.

Jenkinson, M., & Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Medical Image Analysis, 5*(2), 143-156.

Klein, A., & Tourville, J. (2012). 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in neuroscience, 6*, 171.

Long, J., Shelhamer, E., & Darrell, T. (2015). *Fully convolutional networks for semantic segmentation.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., . . . Hendrickson, T. J. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature, 603*(7902), 654-660.

Masouleh, S. K., Eickhoff, S. B., Hoffstaedter, F., Genon, S., & Initiative, A. s. D. N. (2019). Empirical examination of the replicability of associations between brain structure and psychological variables. *elife, 8*.

Mehmood, A., Yang, S., Feng, Z., Wang, M., Ahmad, A. S., Khan, R., . . . Yaqub, M. (2021). A transfer learning approach for early diagnosis of Alzheimer's disease on MRI images. *Neuroscience, 460*, 43-52.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., . . . Andersson, J. L. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience, 19*(11), 1523-1536.

Ooi, L. Q. R., Chen, J., Zhang, S., Kong, R., Tam, A., Li, J., . . . Yeo, B. T. (2022). Comparison of individualized behavioral predictions across anatomical, diffusion and functional connectivity MRI. *NeuroImage, 263*, 119636.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., . . . Lerer, A. (2017). Automatic differentiation in pytorch.

Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A., & Smith, S. M. (2021). Accurate brain age prediction with lightweight deep neural networks. *Medical Image Analysis, 68*, 101871.

Pervaiz, U., Vidaurre, D., Woolrich, M. W., & Smith, S. M. (2020). Optimising network modelling methods for fMRI. *NeuroImage, 211*, 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.

Ritter, A., Hawley, N., Banks, S. J., & Miller, J. B. (2017). The association between Montreal Cognitive Assessment memory scores and hippocampal volume in a neurodegenerative disease sample. *Journal of Alzheimer's Disease, 58*(3), 695-699.

bioRxiv preprint doi: https://doi.org/10.1101/2023.12.31.573801; this version posted January 2, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Sabuncu, M. R., Konukoglu, E., & Initiative, A. s. D. N. (2015). Clinical prediction from structural brain MRI scans: a large-scale empirical study. *Neuroinformatics, 13*, 31-46.

Scheinost, D., Noble, S., Horien, C., Greene, A. S., Lake, E. M., Salehi, M., . . . Barron, D. S. (2019). Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage, 193*, 35-45.

SFCN.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, 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.

Talo, M., Baloglu, U. B., Yıldırım, Ö., & Acharya, U. R. (2019). Application of deep transfer learning for automated brain abnormality classification using MR images. *Cognitive Systems Research, 54*, 176-188.

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.

Wee, C.-Y., Liu, C., Lee, A., Poh, J. S., Ji, H., Qiu, A., & Initiative, A. D. N. (2019). Cortical graph neural network for AD and MCI diagnosis and transfer learning across populations. *NeuroImage: Clinical, 23*, 101929.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data, 3*(1), 1-40.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks, 5*(2), 241-259.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology), 67*(2), 301-320.