Learning pipeline-invariant representation for robust brain phenotype prediction

Xinhui Li [1] **Alex Fedorov** [1] **Mrinal Mathur** [1] **Anees Abrol** [1] **Gregory Kiar** [2] **Sergey Plis** [1] **Vince Calhoun** [1]

Abstract

Deep learning has been widely applied in neuroimaging, including predicting brain-phenotype relationships from magnetic resonance imaging (MRI) volumes. MRI data usually requires extensive preprocessing prior to modeling but variation introduced by different MRI preprocessing pipelines may lead to different scientific findings, even when using identical data. Meanwhile, the machine learning community has emphasized the importance of shifting from model-centric to data-centric approaches considering the essential role of data quality in deep learning applications. Motivated by the recent data-centric perspective, we first evaluate how preprocessing pipeline selection can affect the downstream performance of a supervised learning model. We next propose two pipeline-invariant representation learning methodologies, *Multi-Pipeline Supervised Learning (MPSL)* and *Pipeline-based Contrastive Learning (PXL)*, to improve robustness in classification performance and to capture similar neural network representations. Using a wide range of sample sizes from the UK Biobank dataset, we demonstrate that two models present common advantages, in particular that MPSL and PXL can be used to improve within-sample prediction performance and out-of-sample generalization. Both PXL and MPSL can learn more similar between-pipeline representations. These results suggest that our proposed models can be applied to mitigate pipeline-related biases, and to improve prediction robustness in brain-phenotype modeling.

[1]The Georgia State University/Georgia Institute of Technology/Emory University Center for Translational Research in Neuroimaging and Data Science (TReNDS), Atlanta, GA, USA [2]Child Mind Institute, New York, NY, USA. Correspondence to: Xinhui Li <xinhuili@gatech.edu>, Alex Fedorov <afedorov@gatech.edu>.

Research paper presented at the Data-centric Machine Learning Research (DMLR) Workshop at the 40^{th} International Conference on Machine Learning (ICML) 2023, Honolulu, Hawaii, USA.

1. Introduction

Deep learning has been widely applied to establish novel brain-phenotype relationships and to advance our understanding of brain disorders, in part because of its effectiveness in learning nonlinear relationships from neuroimaging data (e.g., magnetic resonance imaging; MRI) (Abrol et al., 2021; Plis et al., 2014). MRI data usually requires extensive preprocessing, including brain extraction, tissue segmentation, and spatial normalization, among others. These steps are necessary to mitigate data collection artifacts and transform the data to standard spaces for performing statistical analyses and interpretation of results. In the past decade, a growing array of MRI preprocessing pipelines has been developed, but there remains no consensus standard for preprocessing methods. Though these pipelines share basic preprocessing components, the specific implementation at each step can be different. Recent studies have shown that pipeline-related variation may result in significantly different preprocessed results and may lead to conflicting scientific conclusions, even when using identical raw data (Botvinik-Nezer et al., 2020; Li et al., 2021). When used in the development of deep learning models, these pipeline-specific biases may be amplified if models learn shortcut strategies based on unique non-biological features (Torralba & Efros, 2011; Geirhos et al., 2020). However, there is little work in the literature assessing how preprocessing pipelines will affect downstream deep learning task performance.

Recently, the machine learning community has emphasized the importance of shifting from *model-centric* to *data-centric* approaches given that data quality plays an essential role in deep learning applications (Ng, 2021). Motivated by this *data-centric* perspective, we first evaluate how preprocessed data from three commonly-used pipelines affect the downstream performance of a supervised learning model. To this end, a uni-pipeline supervised learning (UPSL) model is trained on a combined age and gender classification task (Abrol et al., 2021), using a dataset preprocessed by each of three pipelines, respectively. We then compare models trained across pipelines through 1) within-sample test accuracy, 2) out-of-sample test accuracy from transfer learning and 3) representational similarity of network layers measured by minibatch centered kernel alignment (CKA) (Nguyen et al., 2020). Our results high-

light significant pipeline-related variation across learned UPSL models, and that learned models cannot generalize to other pipelines. Next, we propose two approaches to mitigate pipeline-related variation and improve between-pipeline representational similarity. First, we suggest a *multi-pipeline supervised learning (MPSL)* model trained on a dataset pairs to take features from both datasets into account. Second, we introduce a *pipeline-based contrastive learning (PXL)* model which integrates both supervised and contrastive learning paradigms. These approaches are evaluated similarly to the UPSL models, and our findings demonstrate that both techniques have common strengths. Specifically, MPSL and PXL can achieve competitive performance within a pipeline set and improve out-of-sample generalization to new pipelines. Notably, both MPSL and PXL can improve between-pipeline representational similarity.

The key contributions of this study include:

- Evaluation of the impact of neuroimaging preprocessing pipelines in a deep learning prediction task;
- Proposal of methodologies to evaluate learning performance including within-sample and out-of-sample test accuracy, and between-pipeline CKA;
- Development of two pipeline-invariant representation learning methodologies, MPSL and PXL, to capture pipeline-invariant latent representations and mitigate pipeline-related biases in the prediction task, including when applied to out-of-sample pipelines.

2. Methods

2.1. Data Preprocessing

Dataset We used the T1-weighted structural MRI (sMRI) images from the UK Biobank dataset (Miller et al., 2016; Abrol et al., 2021) (application number 34175). Subjects were grouped into 5 age groups ($45-52$, $53-59$, $60-66$, $67-73$, and $74-80$ years) and 2 sex groups (males and females), resulting in 10 labels in total. We chose this challenging 10-label classification task to investigate how the model performance is sensitive to pipeline-related variation. 2000 subjects were selected with balanced age and sex categories in the dataset. 1800 subjects with balanced labels were randomly selected and then evenly split into 9 folds for hyperparameter optimization and cross-validation. The remaining 200 subjects with balanced labels were used as a hold-out test set. We further studied data efficiency using a wide range of training sample sizes including 100, 200, 500 and 1000 subjects while keeping the same validation set and test set.

Preprocessing Workflow The same sMRI dataset was preprocessed by each of three commonly-used MRI preprocessing pipelines independently (Figure 1I): 1) the default pipeline in the Configurable Pipeline for the Analysis of Connectomes (C-PAC:Default) (Craddock et al., 2013), 2) the fMRIPrep-options pipeline in C-PAC (C-PAC:fMRIPrep) (Esteban et al., 2019), 3) the UK Biobank FSL pipeline followed by SPM (UKB FSL-SPM) (Alfaro-Almagro et al., 2018; Jenkinson et al., 2012; Friston et al., 1994).

These three pipelines differ at multiple preprocessing steps including brain extraction, tissue segmentation and registration. The detailed preprocessing workflow of each pipeline is as follows: 1) The *C-PAC:Default* structural preprocessing workflow performs brain extraction via AFNI 3dSkullStrip (Cox, 1996), tissue segmentation via FSL FAST (Zhang et al., 2001), and spatial normalization via ANTs SyN non-linear alignment (Avants et al., 2008). 2) The *C-PAC:fMRIPrep* structural pipeline applies ANTs N4 bias field correction (Tustison et al., 2010) on the raw images, followed by ANTs brain extraction, a custom thresholding and erosion algorithm to generate tissue segmentation masks (Esteban et al., 2019), and ANTs SyN alignment to transform the data to the standard space. ANTs registration is performed using skull-stripped images, unlike the C-PAC:Default pipeline which uses whole-head images. 3) The *UKB FSL-SPM* pipeline runs a gradient distortion correction and calculates linear and non-linear transformations via FSL FLIRT (Jenkinson & Smith, 2001; Jenkinson et al., 2002) and FNIRT (Andersson et al., 2007a;b), respectively. Then it performs brain extraction via FSL BET (Smith, 2002) and segments the sMRI data into tissue probability maps. The gray matter images are then warped to standard space, modulated and smoothed using a Gaussian kernel with an FWHM $=$ 10mm via SPM12 (Friston et al., 1994). The preprocessed gray matter volume, a known biomarker of aging and gender effects (Silva et al., 2021), is in MNI (2006) space (Grabner et al., 2006) with dimensions $91 \times 109 \times 91$, corresponding to a voxel size of $2 \times 2 \times 2$ mm^3.

2.2. Model Architectures

We applied a *uni-pipeline supervised learning (UPSL)* model to demonstrate pipeline-related variation in the within-sample and out-of-sample test performance of a downstream prediction task. We next proposed two models, *multi-pipeline supervised learning (MPSL)* and *pipeline-based contrastive learning (PXL)*, to mitigate pipeline-related biases and improve between-pipeline representational similarity.

Encoder Architecture The encoder network was developed based on AlexNet (Krizhevsky et al., 2012) because

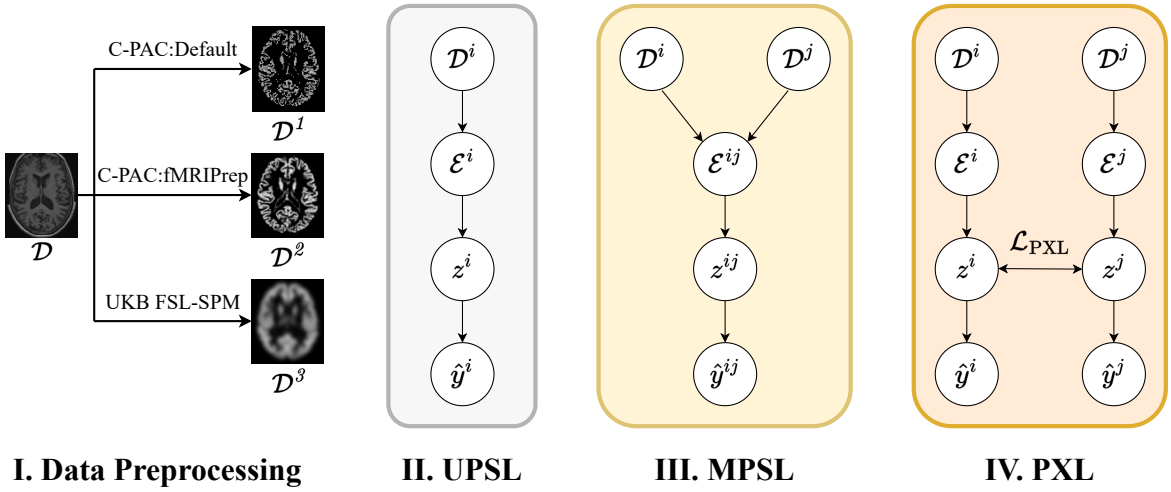

Figure 1. Experiment overview. **I: Data preprocessing.** The UK Biobank sMRI dataset $\mathcal{D}$ is independently preprocessed by each of three commonly-used pipelines (C-PAC:Default, C-PAC:fMRIPrep, UKB FSL-SPM), resulting in three preprocessed datasets $\mathcal{D}^i$ where $i \in \{1, 2, 3\}$. **II: Uni-pipeline supervised learning (UPSL).** Each dataset $\mathcal{D}^i$ is used to train an encoder $\mathcal{E}^i$ to learn representations z^i and predict labels $\hat{y}^i$. **III: Multi-pipeline supervised learning (MPSL).** Two datasets $\mathcal{D}^i$ and $\mathcal{D}^j$ are used to train an encoder $\mathcal{E}^{ij}$ to learn representations z^{ij} and predict labels $\hat{y}^{ij}$. **IV: Pipeline-based contrastive learning (PXL).** Two encoders $\mathcal{E}^i$ and $\mathcal{E}^j$ are trained and the PXL objective $\mathcal{L}_{\text{PXL}}$ maximizes representations z^i and z^j learned from two datasets $\mathcal{D}^i$ and $\mathcal{D}^j$.

it is widely-used in the neuroimaging literature (Lin et al., 2021; Zhang et al., 2020; Fedorov et al., 2019) and previous work (Abrol et al., 2021) provides a performance benchmark. The AlexNet encoder includes 5 convolutional layers and 1 average pooling layer. The convolutional layers have 64, 128, 192, 192, 64 output units, and 62^3, 18^3, 6^3, 6^3, 6^3 output dimensions, respectively. The last convolutional layer with 64 output units defines a 64 dimensional representation.

To further investigate whether pipeline-related variation persists across different encoder architectures, we replicated the UPSL experiment using an effective unsupervised representation learning encoder – deep convolutional generative adversarial network (DCGAN) (Radford et al., 2015).

Uni-Pipeline Supervised Learning To evaluate how data preprocessing affects model performance, we trained a supervised learning model in a combined age and gender prediction task for each preprocessed dataset separately, denoted as uni-pipeline supervised learning (UPSL). The UPSL model includes one encoder $\mathcal{E}^i$, taking each of three preprocessed datasets $\mathcal{D}^i$ as the input and producing the predicted labels $\hat{y}^i$ (Figure 1II).

Multi-Pipeline Supervised Learning Our first proposed architecture, multi-pipeline supervised learning (MPSL), includes one encoder $\mathcal{E}^{ij}$ taking the UK Biobank dataset preprocessed by two pipelines $\mathcal{D}^i$ and $\mathcal{D}^j$ to predict labels $\hat{y}^{ij}$ (Figure 1III). MPSL treats pipelines as unique data augmentation transformations and aims to learn pipeline-invariant representations. Such strategy doubles the dataset size, but the training process and the model implementation are identical to the UPSL.

Pipeline-based Contrastive Learning Our second proposed approach, pipeline-based contrastive learning (PXL), aims to learn pipeline-invariant representations by maximizing agreement between differently preprocessed views of data via a contrastive loss (Bachman et al., 2019). This approach consists of two encoders ($\mathcal{E}^i$, $\mathcal{E}^j$) using the dataset preprocessed by two different pipelines ($\mathcal{D}^i$, $\mathcal{D}^j$) as the inputs separately, and each producing their own sets of output labels ($\hat{y}^i$ and $\hat{y}^j$) (Figure 1IV). The novel contribution in PXL is the addition of a contrastive term to the supervised loss function. The goal of a contrastive loss is to bring the representations from different pipelines closer to each other in the latent space for the same subject, while pushing away the representations for different subjects.

Pipeline-based Contrastive Learning Objective The pipeline-based contrastive learning (PXL) objective function $\mathcal{L}_{\text{PXL}}$ is explained in detail as follows.

Let $\mathcal{D} = \{(x^i, x^j; y) \sim (\mathcal{D}^i, \mathcal{D}^j)\}$ be a dataset of paired samples $(x^i, x^j; y)$, where x^i is an input image from one dataset $\mathcal{D}^i$, x^j is an input image from another dataset $\mathcal{D}^j$, and y is a class label where $y \in \{1, \ldots, 10\}$ in our case. In PXL, two independent encoders $\mathcal{E}^i$ and $\mathcal{E}^j$ parameterized by convolutional neural networks map input images x^i and x^j to representations $z^i = \mathcal{E}^i(x^i)$ and $z^j = \mathcal{E}^j(x^j)$, respectively. To learn the parameters of the encoders, we optimize

the following PXL objective $\mathcal{L}_{\text{PXL}}$:

$$\mathcal{L}_{\text{PXL}} = \lambda \cdot \ell_{\text{supervised}} + (1-\lambda) \cdot \ell_{\text{contrastive}}, \quad (1)$$

where λ is a trade-off hyperparameter between the supervised loss $\ell_{\text{supervised}}$ and the contrastive loss $\ell_{\text{contrastive}}$. Note that the approach can become a fully self-supervised model by setting $\lambda = 0$, or a fully supervised model by setting $\lambda = 1$, equivalent to MPSL with two encoders.

The supervised loss $\ell_{\text{supervised}}$ is defined as the sum of cross-entropy losses ℓ_{CE} for pipelines i and j:

$$\ell_{\text{supervised}} = \ell_{\text{CE}}(g^i(z^i); y) + \ell_{\text{CE}}(g^j(z^j); y), \quad (2)$$

where g is a linear projection head from representations to class labels.

The contrastive loss $\ell_{\text{contrastive}}$ follows the Noise Contrastive Estimation (NCE) lower bound definition (Gutmann & Hyvärinen, 2010). For the n-th sample with a positive pair (z_n^i, z_n^j), the contrastive objective from pipeline i to pipeline j is defined as:

$$\ell_{i \to j}(z_n^i, z_n^j) = -\log \frac{e^{f(h^i(z_n^i), h^j(z_n^j))}}{\sum_{m=1}^{N} \mathbb{1}_{[m \neq n]} e^{f(h^i(z_n^i), h^j(z_m^j))}}, \quad (3)$$

where N is the total number of subjects in the training set, f is the critic function, h^i and h^j are the projection heads for pipelines i and j, respectively (Chen et al., 2020). We use scaled dot product as the critic function f to obtain critic scores and then apply L_2 regularization and soft $tanh$ clipping on the critic scores (Bachman et al., 2019).

The contrastive loss $\ell_{\text{contrastive}}$ is calculated in both directions to ensure its symmetry:

$$\ell_{\text{contrastive}} = \ell_{i \to j} + \ell_{j \to i}. \quad (4)$$

Hyperparameter Search We performed hyperparameter tuning by varying batch size (2, 4, 8, 16, 32, 64) and learning rate (10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}) options for all three models, and we selected the optimal batch size and learning rate according to the cross-validation performance. Additionally, we tuned two other parameters for PXL – the projection head h and the trade-off parameter λ between the supervised loss and the contrastive loss in the PXL objective (see Equation 1). For the choice of the projection head h, we searched over identity projection, linear projection and projection with 1, 2 or 3 hidden layers with dimensionality identical to the representation z. For the choice of the trade-off parameter λ, we evaluated 5 different options ($\lambda = 0, 0.25, 0.5, 0.75, 1$). According to the hyperparameter search result, we selected a batch size of 4 and a learning rate of 10^{-3} for UPSL; a batch size of 32 and a learning rate of 10^{-3} for MPSL; an identity projection head, a batch size of 4, a learning rate of 10^{-4}, a trade-off parameter λ of 0.75 for PXL.

Cross-Validation Each of three model was trained using the Adam optimizer (Kingma & Ba, 2014) for 200 epochs until convergence. We repeated the experiment across 9 folds of training and validation data with balanced labels. We reported the inference performance on the hold-out test set from models trained on 9 folds. All models were implemented in the PyTorch framework and trained with NVIDIA A40 GPUs.

2.3. Ablation Study

Data Efficiency To characterize the relationship between the model performance and the training sample size, we compared the inference performance on the hold-out test set using a wide range of training sample sizes ($n = 100, 200, 500, 1000, 1600$).

Smoothing Effect Spatial smoothing is a key preprocessing step. Among three pipelines used in this study, the UKB FSL-SPM pipeline implements spatial smoothing using a Gaussian kernel with an FWHM = 10mm while the other two pipelines do not perform smoothing. To investigate the smoothing effect on the other two pipelines, we applied a Gaussian kernel with an FWHM uniformly sampled from 0 to 10mm on the C-PAC:Default and C-PAC:fMRIPrep preprocessed datasets. To further assess how different levels of smoothing affect the model performance, we applied smoothing with different probability options ($p = 0, 0.25, 0.5, 0.75, 1$). According to the cross-validation performance, we reported the inference performance using MPSL with smoothing probability $p = 1$ and PXL with smoothing probability $p = 0.75$.

2.4. Evaluation Metrics

Prediction Performance We used two metrics to measure model performance: within-sample and out-of-sample inference performance. The within-sample test accuracy was obtained by applying the trained model on the hold-out test set preprocessed by the same pipeline. To measure out-of-sample generalizability, we trained a logistic regression model from *scikit-learn* (Pedregosa et al., 2011) using the training set from a different pipeline and reported the test accuracy.

Representational Similarity Minibatch centered kernel alignment (CKA) (Nguyen et al., 2020) was used to efficiently measure neural network representational similarity of high-dimensional neuroimaging features from pipeline pairs.

We describe minibatch CKA in detail as follows. Let $\mathbf{X} \in \mathbb{R}^{m \times u_1}$ and $\mathbf{Y} \in \mathbb{R}^{m \times u_2}$ denote representations of two layers, where m is the number of samples, and u_1 and u_2 are the number of neuron units in $\mathbf{X}$ and $\mathbf{Y}$, respectively. Here, m refers to 200 subjects in the test set. We flatten

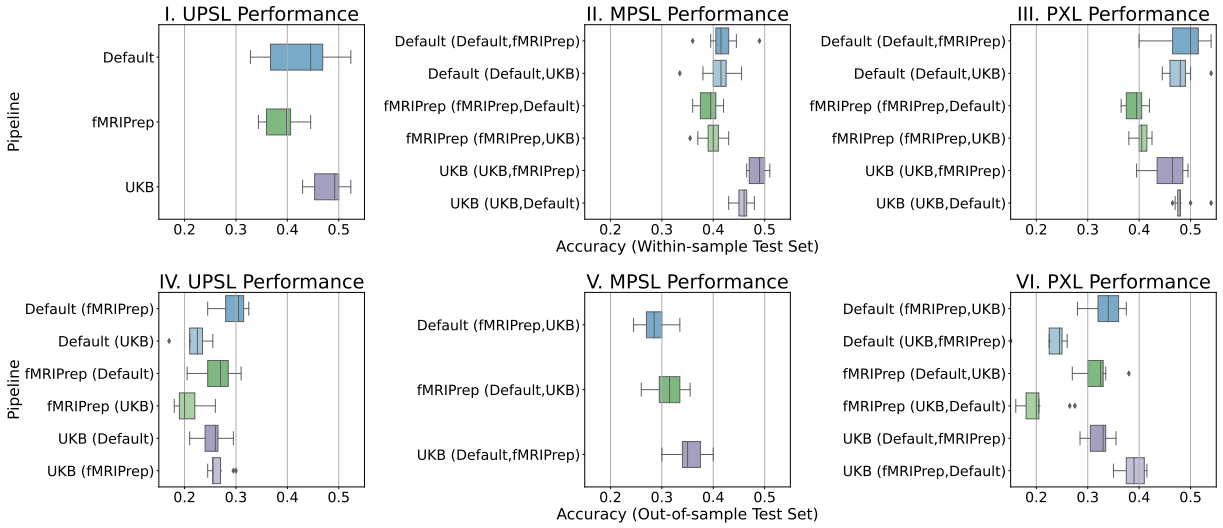

Figure 2. Within-sample and out-of-sample inference performance. The box plot shows the median and the interquartile range of the within-sample (top) and out-of-sample (bottom) test accuracies on the hold-out test set across 9 folds. The vertical y-axis label shows the test set for performance evaluation, with the training set (pair) indicated in brackets. The label denotes "the test set" in panel I and "the test set (the training set)" in panels II-VI. For example, in panel II, Default (Default,fMRIPrep) indicates that the test set is Default and the training set pair is Default and fMRIPrep. In panel VI, the first dataset in the training set pair indicates the encoder used for evaluation. For example, Default (fMRIPrep,UKB) indicates that the test set is Default, the training set pair is fMRIPrep and UKB, and the encoder is trained on fMRIPrep. UPSL shows significant within-sample prediction variation and poor out-of-sample generalization. MPSL achieves the best within-sample test performance on the UKB test set while PXL achieves competitive performance on the Default and fMRIPrep test sets. Both PXL and MPSL show more robust out-of-sample generalization compared to UPSL.

channels c and three spatial dimensions (width w, height h, depth d) of a convolutional layer into u neurons to compare representations of different layers, i.e. $u = c \times h \times w \times d$ (Raghu et al., 2017). We then randomly split m subjects into k minibatches and each minibatch contains n subjects. Here, we include $n = 8$ subjects in each minibatch. Let $\mathbf{X}_i \in \mathbb{R}^{n \times u_1}$ and $\mathbf{Y}_i \in \mathbb{R}^{n \times u_2}$ denote representations of two layers in the ith batch. We then compute the similarity matrices $\mathbf{K} = \mathbf{X}_i \mathbf{X}_i^\top$ and $\mathbf{L} = \mathbf{Y}_i \mathbf{Y}_i^\top$ and estimate the similarity of the similarity matrices using Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005).

Minibatch CKA is computed by averaging the linear CKA across k minibatches:

$$\text{CKA} = \frac{\frac{1}{k}\sum_{i=1}^{k} \text{H}(\mathbf{K}, \mathbf{L})}{\sqrt{\frac{1}{k}\sum_{i=1}^{k} \text{H}(\mathbf{K}, \mathbf{K})}\sqrt{\frac{1}{k}\sum_{i=1}^{k} \text{H}(\mathbf{L}, \mathbf{L})}}. \quad (5)$$

An unbiased estimator of HSIC (Song et al., 2012) is used in minibatch CKA:

$$\text{H}(\mathbf{K}, \mathbf{L}) = \frac{1}{n(n-3)}(\text{tr}(\tilde{\mathbf{K}}\tilde{\mathbf{L}}) + \frac{\mathbf{1}^\top \tilde{\mathbf{K}} \mathbf{1} \mathbf{1}^\top \tilde{\mathbf{L}} \mathbf{1}}{(n-1)(n-2)} - \frac{2}{n-2}\mathbf{1}^\top \tilde{\mathbf{K}}\tilde{\mathbf{L}}\mathbf{1}), \quad (6)$$

where $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$ are obtained by setting the diagonal entries of $\mathbf{K}$ and $\mathbf{L}$ to zeros.

Note that the CKA values are independent of the selection of batch sizes because of the unbiased estimator of HSIC. Detailed proof of the feasibility of using minibatch CKA to approximate CKA can be found in (Nguyen et al., 2020).

3. Results

3.1. UPSL shows significant within-sample prediction variation and poor out-of-sample generalization.

The median of UPSL within-sample test accuracies ranges from 39.8% (fMRIPrep) to 49.2% (UKB), with a difference of 9.4% (chance accuracy: 10%), highlighting notable prediction difference when applying different preprocessing pipelines on identical data (Figure 2I)[1]. The statistical analysis reveals that the UKB test result is significantly different from the Default test result (p-value $= 0.0078$, two-sided Wilcoxon signed-rank test, Bonferroni corrected critical value $\alpha = \frac{0.05}{3} = 0.0167$) and the fMRIPrep test result (p-value $= 0.0039$), though the performance difference between Default and fMRIPrep is not significant (p-value $= 0.3594$). To further investigate whether pipeline-related variation exists across different encoder architec-

[1]For conciseness, Default, fMRIPrep and UKB are used to denote C-PAC:Default, C-PAC:fMRIPrep and UKB FSL-SPM, respectively.

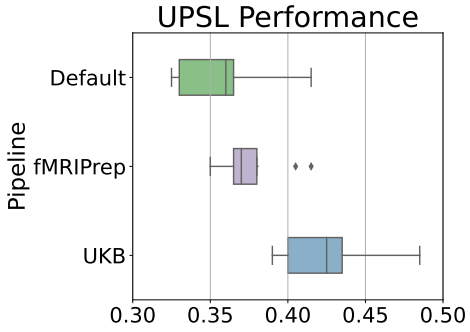

Figure 3. UPSL within-sample inference performance using the DCGAN encoder. The box plot shows the median and the interquartile range of the within-sample test accuracies on the hold-out test set across 9 folds. Similarly, significant UPSL within-sample prediction variation exists using the DCGAN encoder.

tures, we replicated the UPSL experiment using the DCGAN encoder. Similarly, we observe a similar pipeline-related effect using the DCGAN encoder – the median within-sample test accuracies across 9 folds are 36.0%, 37.0%, 42.5% for Default, fMRIPrep and UKB, respectively (Figure 3). The maximum difference of the medians of within-sample test accuracies is as high as 6.5% across pipelines. The statistical test also shows that UKB is significantly different from Default and fMRIPrep (p-value $= 0.0039$ and 0.0078). The UPSL result indicates that preprocessed data from different pipelines will significantly affect downstream prediction performance and variation persists across two encoder architectures.

Moreover, UPSL out-of-sample transfer learning performance is relatively poor with a median of 25.5% across all test accuracies (Figure 2IV), indicating that a naive UPSL approach cannot generalize well across pipeline populations.

3.2. MPSL and PXL achieve competitive within-sample inference performance.

To mitigate pipeline-related biases, we developed two pipeline-invariant representation learning methodologies MPSL and PXL. The MPSL design inspired by data augmentation shows the potential to improve prediction performance. Particularly, the mean of the UKB within-sample test accuracies from the UKB and fMRIPrep training set pair is 48.8%, the highest score across three models (Figure 2II, Table 1).

In PXL, we observe within-sample performance gains for the Default and fMRIPrep test sets compared to UPSL and MPSL (Figure 2III, Table 1). Specifically, PXL achieves the highest within-sample test accuracies on the Default test set (mean $\pm$ std: $48.9\% \pm 4.1\%$) and the fMRIPrep test set (mean $\pm$ std: $40.4\% \pm 1.3\%$). These results suggest that PXL can achieve competitive within-sample test performance by utilizing a contrastive objective.

Table 1. Within-sample inference performance. The mean $\pm$ the standard deviation of within-sample test accuracies ($\%$) across 9 folds. The test set order matches the vertical y-axis label order in Figure 2. UPSL shows significant within-sample prediction variation. MPSL and PXL achieve competitive within-sample inference performance.

	DEFAULT	FMRIPREP	UKB
UPSL	42.0 ± 6.4	39.0 ± 3.2	48.2 ± 3.0
MPSL	41.9 ± 3.4	39.3 ± 1.9	$\mathbf{48.8 \pm 1.5}$
	40.7 ± 3.2	39.7 ± 2.1	45.7 ± 1.6
PXL	$\mathbf{48.9 \pm 4.1}$	39.1 ± 1.7	45.7 ± 3.3
	48.1 ± 2.7	$\mathbf{40.4 \pm 1.3}$	48.5 ± 2.1

Table 2. Out-of-sample transfer learning inference performance. The mean $\pm$ the standard deviation of out-of-sample test accuracies ($\%$) across 9 folds. The test set order matches the vertical y-axis label order in Figure 2. UPSL shows poor out-of-sample generalization while PXL and MPSL demonstrate robust out-of-sample generalization.

	DEFAULT	FMRIPREP	UKB
UPSL	29.7 ± 2.7	26.5 ± 3.0	25.6 ± 2.4
	22.2 ± 2.3	20.7 ± 2.4	26.6 ± 1.9
MPSL	28.8 ± 2.7	31.3 ± 2.8	35.3 ± 3.0
PXL	$\mathbf{33.8 \pm 2.8}$	$\mathbf{32.0 \pm 2.9}$	32.4 ± 2.2
	22.2 ± 4.2	20.6 ± 3.7	$\mathbf{38.9 \pm 2.1}$

3.3. PXL and MPSL demonstrate robust out-of-sample generalization.

Compared to UPSL, MPSL models improve generalization performance with average out-of-sample transfer learning accuracies of 31.3% and 35.3% for fMRIPrep and UKB, respectively (Figure 2V, Table 2). PXL models also demonstrate robust out-of-sample performance (Figure 2VI, Table 2). Notably, PXL achieves the highest out-of-sample inference performance on all three preprocessed datasets (mean $\pm$ std: Default $33.8\% \pm 2.8\%$; fMRIPrep $32.0\% \pm 2.9\%$; UKB $38.9\% \pm 2.1\%$). These results highlight that the learned representations in MPSL and PXL models can generalize better across pipeline populations.

3.4. PXL and MPSL capture more similar between-pipeline representations.

We measured representational similarity between network layers using minibatch CKA (Nguyen et al., 2020). A higher CKA value indicates a more similar representation captured between layers. As shown in Figure 4, between-pipeline representations from the last three layers are more simi-

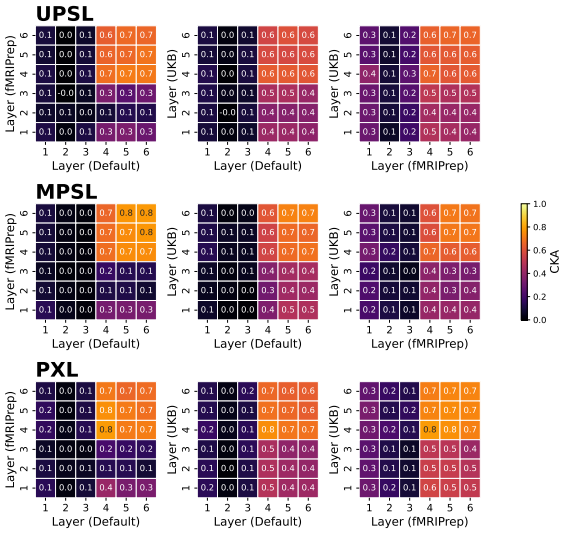

Figure 4. Between-pipeline CKA across network layers. PXL and MPSL capture more similar between-pipeline representations across the last three layers.

lar while those from the first three layers are less similar. Among three models, PXL and MPSL capture more similar between-pipeline representations across the last three layers than UPSL. Particularly, CKA values across the last three layers from PXL are significantly higher than those from MPSL (p-value $= 0.0039$), while MPSL shows a significant improvement over UPSL (p-value $= 0.0003$). PXL shows the highest average CKA value 0.718 across the last three layers.

Taken together, these results demonstrate that datasets preprocessed by different pipelines result in significantly different downstream performance, and that learned models may not generalize to other pipelines in UPSL. As extensions beyond naive UPSL approaches, both MPSL and PXL can achieve competitive within-sample and out-of-sample inference performance. Moreover, both MPSL and PXL can capture more similar between-pipeline representations compared to UPSL.

3.5. Ablation: Pipeline-related variation exists across a wide range of training sample sizes.

To evaluate how the number of training samples affects prediction performance, we compared the model performance using a wide range of training sample sizes ($n = 100, 200, 500, 1000, 1600$). As the training sample size increases from 100 to 1600 subjects, we observe that the prediction performance increases from 16.3%, 20.5%, 23.9% to 42.0%, 39.0%, 48.2% for Default, fMRIPrep and UKB, respectively. However, prediction variation induced by pipelines exists regardless of the number of training samples (Figure 5I). When comparing within-sample test performance across three models, we note that PXL achieves the best performance when the sample size is small (Figure 5III), suggesting that PXL can more efficiently learn predictive features from minimal training samples. MPSL out-of-sample test accuracy increases as the sample size increases, and three pipelines converge to statistically consistent performance when the sample size reaches more than 1000 subjects (p-value $> \frac{0.05}{3}$, Figure 5V). It might be practically meaningful that MPSL can lead to consistent generalization across pipelines when the sample size is sufficiently large.

As shown in Figure 6, when the training sample size is sufficiently large (> 1000 subjects), both PXL and MPSL show significantly higher average CKA values across the last three layers than UPSL (p-value $< \frac{0.05}{3}$). When the training sample size is small (100 subjects), we observe that CKA values from MPSL are relatively low, which is probably associated with its relatively poor within-sample test performance. This indicates that MPSL would require a relatively large training sample size to achieve competitive performance.

3.6. Ablation: Spatial smoothing improves MPSL and PXL performance.

In UPSL experiment (Figure 2), we observe that UKB within-sample test performance is significantly different from others. One potential reason could be that UKB is the only pipeline which implements spatial smoothing in this study. To investigate how different levels of smoothing affect model performance, we tested different smoothing probability options ($p = 0, 0.25, 0.5, 0.75, 1$). As shown in Figure 7, the application of spatial smoothing improves MPSL and PXL model performance to different extents. More specifically, an intermediate level of smoothing (e.g., $p = 0.5$) improves MPSL model performance while a high level of smoothing (e.g., $p = 1$) increases PXL model performance.

4. Discussion

The contributions of the present work are two-fold.

First, we evaluated the impact of neuroimaging preprocessing pipelines and preprocessed data quality in a prediction task, and demonstrated current limitations of UPSL. The UPSL performance demonstrates that the same dataset preprocessed by different pipelines can result in significantly different prediction performance. As shown in Figure 2I, the test accuracy difference from UPSL can be as high as 9.4% (chance accuracy: 10%) when using datasets preprocessed by different pipelines. The result emphasizes the importance of clear scientific communication surrounding decisions in

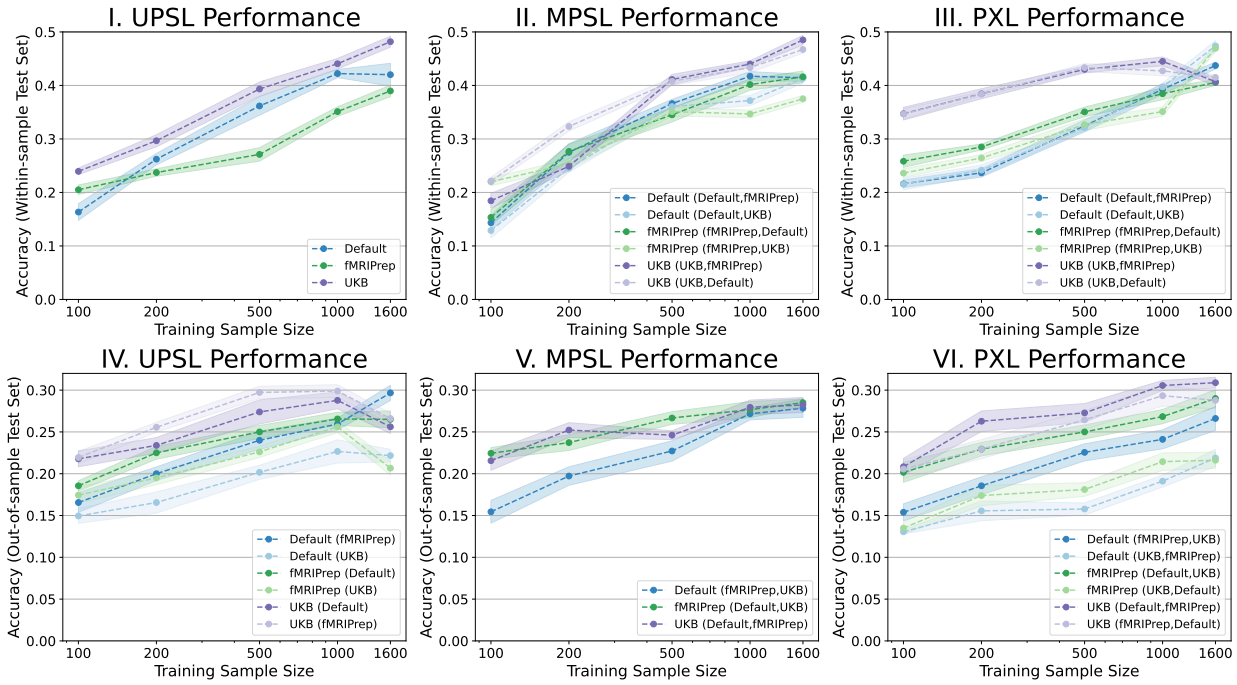

Figure 5. Within-sample and out-of-sample inference performance across different training sample sizes. The line plot shows the mean ± the standard error of test accuracies on the hold-out test set across 9 folds. Pipeline-related variation exists across different sample sizes.

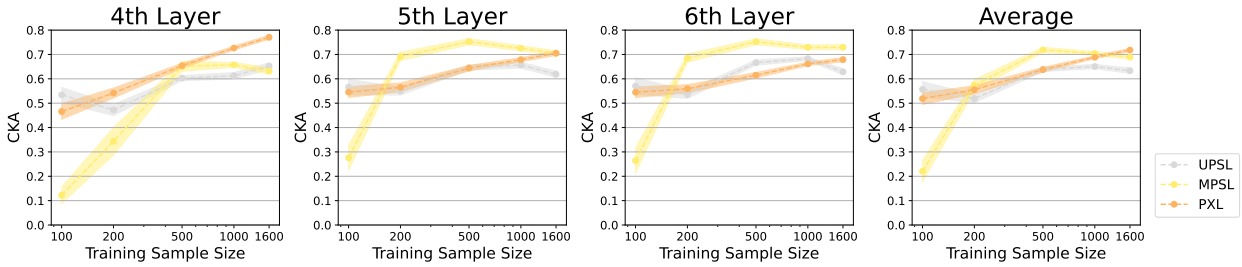

Figure 6. Between-pipeline CKA across different training sample sizes in the last three layers. The line plot shows the mean ± the standard error of between-pipeline CKA values across 9 folds. PXL and MPSL capture more similar between-pipeline representations across the last three layers when the training sample size is sufficiently large.

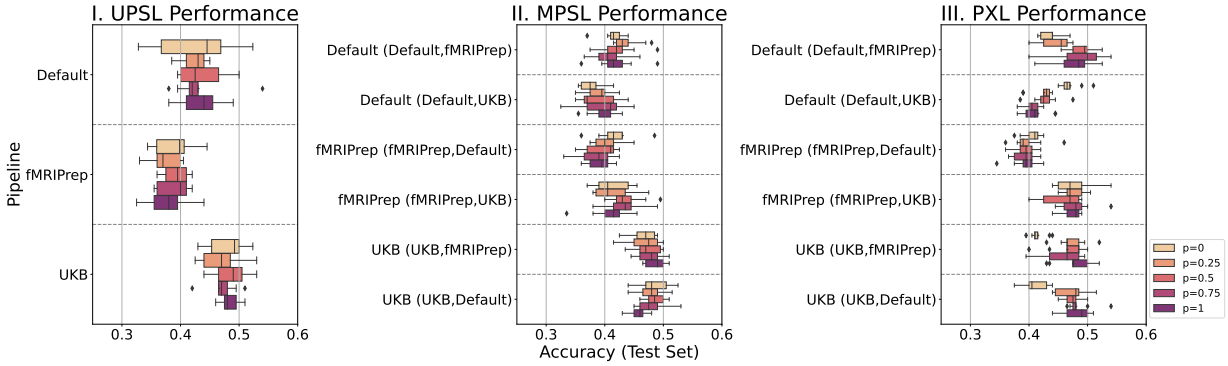

Figure 7. Within-sample inference performance across different smoothing probabilities (p). The box plot shows the median and the interquartile range of within-sample test accuracies on the hold-out test set across 9 folds. An intermediate level of smoothing (e.g., $p = 0.5$) improves MPSL model performance while a high level of smoothing (e.g., $p = 1$) increases PXL model performance.

neuroimaging preprocessing, and making pipelines publicly available to allow for evaluation, comparison, and reproduction in the context of downstream prediction tasks.

Next, we proposed two approaches, MPSL and PXL, to mitigate pipeline-related variation in the construction of deep learning models. While the MPSL approach is a naive extension of UPSL, we note that the MPSL approach leads to competitive within-sample performance (Figure 2II). One reason could be that the MPSL encoder learns features from datasets preprocessed by two pipelines, whereas the UPSL encoder is trained on a single dataset. Our novel approach PXL adopts a contrastive loss function leading to the improved within-sample performance and representational similarity across pipelines. Specifically, PXL is able to achieve the highest within-sample test accuracies on the C-PAC:Default and C-PAC:fMRIPrep datasets (Figure 2III). Additionally, it exhibits better performance when training sample size is small (Figure 5III) across three models. Both MPSL and PXL show competitive out-of-sample transfer learning performance, demonstrating their robust generalizability on new pipelines (Figure 2V, VI). Notably, both MPSL and PXL capture more similar representations in the last three layers (Figure 4), supporting their potentials to achieve pipeline-invariant learning. In practice, our results suggest that prediction performance can be improved by integrating datasets preprocessed by at least two pipelines as well as utilizing MPSL and PXL architectures.

It is important to recognize the limitations of the present study. Here, we used structural MRI due to the simplicity of building three-dimensional models and we only performed a combined age and sex prediction task. In future work, it is worth evaluating these approaches on other neuroimaging modalities such as functional MRI that incorporates temporal dynamics as well as on other tasks such as brain disorder prediction.

5. Conclusion

We show that pipeline-related variation can make a significant difference in the deep learning model performance of a downstream age and sex prediction task. We then propose two pipeline-invariant representation learning approaches, MPSL and PXL, to mitigate biases introduced by data preprocessing. Our results demonstrate that both MPSL and PXL can achieve competitive and robust inference performance and improve representational similarity of network layers. The proposed models can be applied to mitigate pipeline-related variation, and even site effects and data acquisition-related effects, as well as improve prediction robustness in brain-phenotype modeling.

References

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., and Calhoun, V. Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature Communications*, 12:353, 2021. doi: 10.1038/s41467-020-20655-6.

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166:400–424, 2018.

Andersson, J. L., Jenkinson, M., and Smith, S. Non-linear optimisation fmrib technical report tr07ja1. *Practice*, 2007a.

Andersson, J. L., Jenkinson, M., Smith, S., et al. Non-linear registration, aka spatial normalisation fmrib technical report tr07ja2. *FMRIB Analysis Group of the University of Oxford*, 2(1):e21, 2007b.

Avants, B. B., Epstein, C. L., Grossman, M., and Gee, J. C. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis*, 12(1): 26–41, 2008.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Botvinik-Nezer, R. et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, 2020. doi: 10.1038/s41586-020-2314-9.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

Cox, R. W. Afni: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical research*, 29(3):162–173, 1996.

Craddock, C., Sikka, S., Cheung, B., Khanuja, R., Ghosh, S. S., Yan, C., Li, Q., Lurie, D., Vogelstein, J., Burns, R., et al. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac). *Front Neuroinform*, 42:10–3389, 2013.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., and Gorgolewski, K. J. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

Fedorov, A., Hjelm, R. D., Abrol, A., Fu, Z., Du, Y., Plis, S., and Calhoun, V. D. Prediction of progression to alzheimer's disease with deep infomax. In *2019 IEEE EMBS International conference on biomedical & health informatics (BHI)*, pp. 1–5. IEEE, 2019.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, nov 2020. doi: 10.1038/s42256-020-00257-z.

Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., and Collins, D. L. Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 58–66. Springer, 2006.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pp. 63–77. Springer, 2005.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Teh, Y. W. and Titterington, M. (eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pp. 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

Jenkinson, M. and Smith, S. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, 2001.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. Fsl. *Neuroimage*, 62(2):782–790, 2012.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Li, X., Ai, L., Giavasis, S., Jin, H., Feczko, E., Xu, T., Clucas, J., Franco, A., Sólon Heinsfeld, A., Adebimpe, A., Vogelstein, J., Yan, C.-G., Esteban, O., Poldrack, R., Craddock, C., Fair, D., Satterthwaite, T., Kiar, G., and Milham, M. Moving beyond processing and analysis-related variation in neuroscience. *bioRxiv*, 2021. doi: 10.1101/2021.12.01.470790.

Lin, L., Zhang, G., Wang, J., Tian, M., and Wu, S. Utilizing transfer learning of pre-trained alexnet and relevance vector machine for regression for predicting healthy older adult's brain age from structural mri. *Multimedia Tools and Applications*, 80(16):24719–24735, 2021.

Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J., Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L., et al. Multimodal population brain imaging in the uk biobank prospective epidemiological study. *Nature neuroscience*, 19(11): 1523–1536, 2016.

Ng, A. Mlops: From model-centric to data-centric ai, 2021.

Nguyen, T., Raghu, M., and Kornblith, S. Do wide and deep networks learn the same things? uncovering how neural network representations vary with width and depth. *arXiv preprint arXiv:2010.15327*, 2020.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Plis, S. M., Hjelm, D. R., Salakhutdinov, R., Allen, E. A., Bockholt, H. J., Long, J. D., Johnson, H. J., Paulsen, J. S., Turner, J. A., and Calhoun, V. D. Deep learning for neuroimaging: a validation study. *Frontiers in neuroscience*, 8:229, 2014.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Silva, R. F., Damaraju, E., Li, X., Kochunov, P., Belger, A., Ford, J. M., Mathalon, D. H., Mueller, B. A., Potkin, S. G., and Preda, A. Direct linkage detection with multimodal iva fusion reveals markers of age, sex, cognition, and schizophrenia in large neuroimaging studies. *bioRxiv*, pp. 2021–12, 2021.

Smith, S. M. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. Feature selection via dependence maximization. *Journal of Machine Learning Research*, 13(5), 2012.

Torralba, A. and Efros, A. A. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528, 2011. doi: 10.1109/CVPR.2011.5995347.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

Zhang, J., Li, X., Li, Y., Wang, M., Huang, B., Yao, S., and Shen, L. Three dimensional convolutional neural network-based classification of conduct disorder with structural mri. *Brain imaging and behavior*, 14(6):2333–2340, 2020.

Zhang, Y., Brady, M., and Smith, S. Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE transactions on medical imaging*, 20(1):45–57, 2001.