Exploring the potential of representation and transfer learning for anatomical neuroimaging: application to psychiatry

Benoit Dufumier, Pietro Gori, Sara Petiton, Robin Louiset, Jean-François Mangin, Antoine Grigis, Edouard Duchesnay

Exploring the potential of representation and transfer learning for anatomical neuroimaging: application to psychiatry

Benoit Dufumier[1,2*], Pietro Gori[2], Sara Petiton[1], Robin Louiset[1,2], Jean-François Mangin[1], Antoine Grigis[1], and Edouard Duchesnay[1]

[1]Université Paris-Saclay, CEA, CNRS, UMR9027 Baobab, NeuroSpin, Saclay, France
[2]LTCI, Télécom Paris, IPParis, Palaiseau, France
[*]benoit.dufumier@cea.fr

ABSTRACT

The perspective of personalized medicine for brain disorders requires efficient learning models for anatomical neuroimaging-based prediction of clinical conditions. There is now a consensus on the benefit of deep learning (DL) in addressing many medical imaging tasks, such as image segmentation. However, for single-subject prediction problems, recent studies yielded contradictory results when comparing DL with Standard Machine Learning (SML) on top of classical feature extraction. Most existing comparative studies were limited in predicting phenotypes of little clinical interest, such as sex and age, and using a single dataset. Moreover, they conducted a limited analysis of the employed image pre-processing and feature selection strategies. This paper extensively compares DL and SML prediction capacity on five multi-site problems, including three increasingly complex clinical applications in psychiatry namely schizophrenia, bipolar disorder and Autism Spectrum Disorder (ASD) diagnosis. To compensate for the relative scarcity of neuroimaging data on these clinical datasets, we also evaluate three pre-training strategies for transfer learning from brain imaging of the general healthy population: self-supervised learning, generative modelling and supervised learning with age. Overall, we find similar performance between randomly initialized DL and SML for the three clinical tasks and a similar scaling trend for sex prediction. This was replicated on an external dataset. We also show highly correlated discriminative brain regions between DL and linear ML models in all problems. Nonetheless, we demonstrate that self-supervised pre-training on large-scale healthy population imaging dataset ($N \approx$10k), along with Deep Ensemble, allows DL to learn robust and transferable representations to smaller-scale clinical datasets ($N \leq$ 1k). It largely outperforms SML on 2 out of 3 clinical tasks both in internal and external test sets. These findings suggest that the improvement of DL over SML in anatomical neuroimaging mainly comes from its capacity of learning meaningful and useful abstract representations of the brain anatomy, and it sheds light on the potential of transfer learning for personalized medicine in psychiatry.

1 Introduction

With the ever-growing availability of brain imaging data (*e.g.*, UKBioBank[5], HCP[6], ABIDE[7], etc.), Machine Learning (ML) and, in particular, Deep Learning (DL) models are starting to emerge for personalized medicine and biomarker discovery in psychiatry and neurology. Psychiatric disorders are complex and highly heterogeneous, gathering clinical, biological, and environmental variabilities[8], thus making their neurobiological characterization challenging. In this context, "Standard" ML (SML) models, including (regularized) linear models and kernel-based methods, have been broadly used in neuroimaging studies, where the number of available samples n is usually small ($n < 10^3$) and the number of imaging features p quite large (typically $p > 10^5$). One main drawback that limited their applicability in many medical imaging applications[9] (and more broadly in biomedicine) is their need for pre-selected features manually or automatically designed (*e.g.*, through feature engineering). As opposed to SML methods, DL, and in particular, ConvNets (CNN), can automatically learn from raw data a hierarchical representation of features relevant to the task at hand (*e.g.*, classification or regression). They have shown impressive results on supervised and unsupervised learning problems, both on natural and medical images, by learning a high abstraction of the data in a layer-wise manner. Several studies have started to benchmark such models on functional brain imaging[10,11] and cortical data[12] for phenotypes prediction, some of them showing improvement of DL over SML[11]. However, as noted in several recent studies[13–17], the benefit of using DL on anatomical brain MRI data for single-subject prediction (required for psychiatric disorder diagnosis or prognosis) is unclear, and a careful and extensive comparison with simple regularized linear models and kernel-methods is still missing. For bipolar disorder[18] and schizophrenia[19], several very large meta-analysis

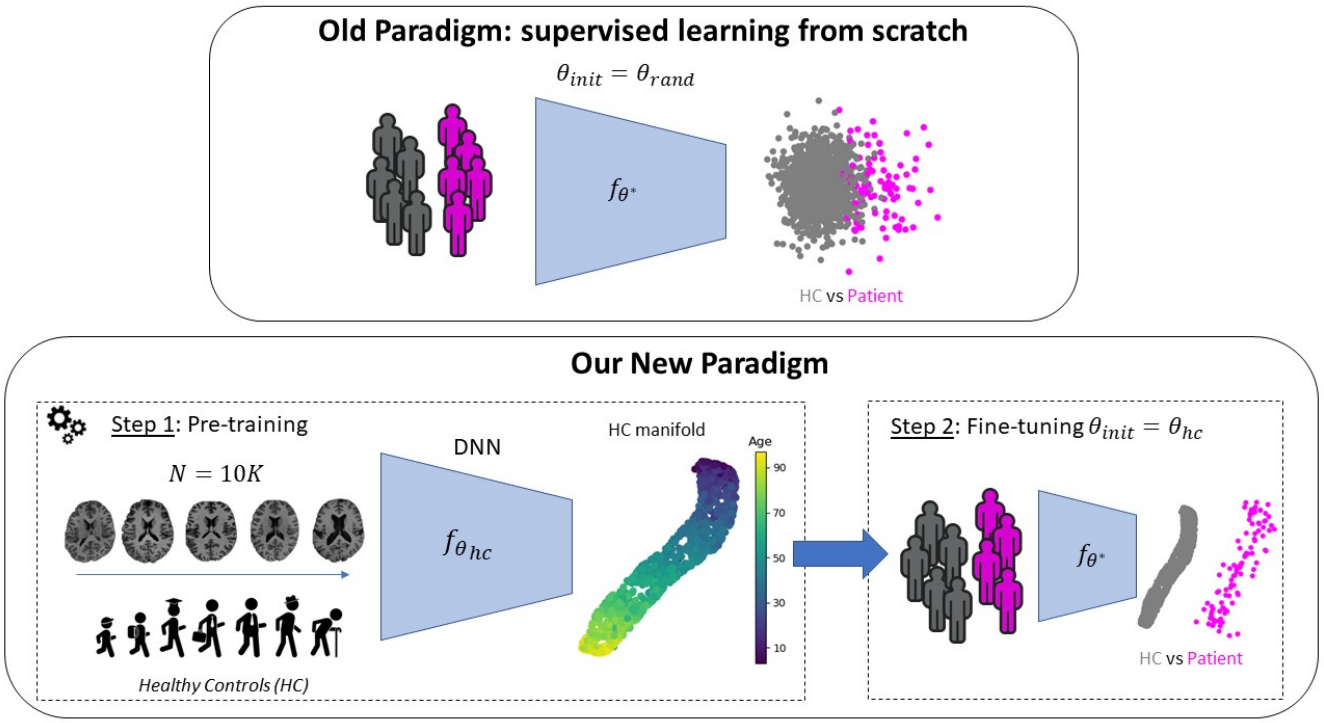

Figure 1. New paradigm for discriminating psychiatric disorders at the subject-level. In a pre-training phase, a non-linear DNN f_θ is trained to learn a low-dimensional embedding from a large brain imaging dataset of healthy controls, discovering the general variability associated with non-specific variables such as age and sex. This pre-training can be performed either with i) self-supervised tasks (e.g., contrastive learning[1,2]) ii) generative modeling (e.g., VAE[3]) or iii) discriminative tasks (e.g., age prediction[4]). In the second step, the model is initialized with pre-trained weights $\theta_{init} = \theta_{hc}$ and fine-tuned to discriminate between patients and controls. Our main hypothesis is that the representation learned during pre-training will allow easier discovery of the specific variability associated with the pathology of interest (e.g., abnormal cortical atrophy in temporal and pre-fontal regions for schizophrenia or ASD).

led by the ENIGMA consortium have shown significant variations in cortical regions including prefrontal, anterior temporal and insula cortices, visible in structural neuroimaging. More fine-grained analysis at the voxel-level is required to improve the diagnosis and prognosis accuracy of machine learning models at the subject-level[20]. For ASD[21], smaller subcortical volumes of the pallidum, putamen, amygdala, and nucleus accumbens, as well as increased cortical thickness in the frontal cortex and decreased thickness in the temporal cortex were observed from structural brain imaging. Only case-control study was performed in this case and more effort is required at the single-subject level to investigate anatomical brain abnormalities and its link to behavior.

In a recent study[14] based on the UK Biobank dataset[5] (UKB), Schulz et al. studied whether the two main priors encoded in current CNN, namely translational invariance (derived from the convolution operation) and compositionality (derived from its hierarchical structure), can be exploited to capture non-linear dependencies in structural/functional Magnetic Resonance Imaging (sMRI/fMRI) data for individual prediction tasks. In particular, they showed that SML and DL models have a similar scaling trend, even in the large-scale regime ($N_{train} = 8$k), on both modalities (sMRI and fMRI) for a variety of tasks (age and sex prediction but also fluid intelligence or household income prediction). However, these results contradict the ones obtained by Peng et al.[22] on both the Predictive Analytics Competition[23] and UKB, as noted by Abrol et al.[13]. Specifically, Abrol et al. pointed out some technical flaws in the work of Schulz et al. that heavily affected their conclusions. The main shortcomings were the feature selection step performed for SML and DL (with an arbitrary number of reduced dimensions) and using a single central brain slice in their main experiments, limiting DL representation capacity. On the contrary, in their study[13], Abrol et al. performed feature selection only for SML models, and they used a whole-brain approach for DL. They found a significantly better scaling trend for DL on UKB with training samples ranging from $N_{train} \sim 2000$ to $N_{train} = 10^4$, and they attributed the performance drop in the work of Schulz et al. to a coding bug. Moreover, they also found a small but significant increase in performance on the Mini-Mental State Examination (MMSE) regression task ($N_{train} = 428$, -0.07 of

MAE, Mean Absolute Error, for DL vs. SML), which might be in contradiction with a recent benchmark[24] on Alzheimer's detection that found no significant differences between SML and DL. While this score indicates Alzheimer's disease severity, it does not translate into Alzheimer's diagnosis[25], which may explain the different findings. Finally, they suggested that DL can consistently extract robust brain representations according to different saliency maps techniques, showing consistent patterns across runs and saliency methods for age and sex prediction.

Nonetheless, the past literature comparing DL and SML with neuroimaging data has still several limitations that we highlight here.

Limited number of prediction tasks. First, most recent papers[13,14,22] have mainly focused their analysis on age and sex prediction in the healthy population. While studying age regression has become an important research field for many research questions (new biomarkers discovery for psychiatric disorders or neurocognitive impairment with brain age gap[26–29] or normative modeling[8,30,31]), DL evaluation on psychiatric disorder classification is (also) urgently required. The advances made in the ML field are remarkable, and the availability of large-scale neuroimaging data previously inaccessible to research gives a unique opportunity to study these clinical tasks. The question of whether non-linearities can be captured in highly heterogeneous clinical cohorts, including patients with schizophrenia[8,26] (SCZ), Bipolar Disorder[8] (BD), and Autism Spectrum Disorders[31] (ASD) is still debated, and no clear consensus arises[16,24,32]. This is mainly due to the small sample size of the current datasets (typically $N < 10^3$), which causes ML models to over-fit and bias the neuroimaging community towards over-optimistic results[33–37]. These disorders involve subtle anatomical atrophies/hypertrophies in cortical and subcortical structures, and their identification is still a difficult challenge.

No replication on external multi-site data. Second, both Abrol et al. and Schulz et al. have based their analysis mostly on a unique homogeneous (*i.e.* single-site and single-scanner model) dataset (UKB) that does not reflect the inevitable heterogeneity in emerging large multi-site and multi-scanner clinical data collections (*e.g.*, ABIDE, ABCD, SCHIZCONNECT, etc). As such, a comprehensive complementary benchmark on phenotype prediction with large-scale multi-site datasets is required. As noted in a recent study[38], since DL has an exceptional capacity to learn any function (even random noise[39]), it can also learn "disease-irrelevant site-specific characteristics," and its generalization capacity on data acquired on never-seen sites must also be reported.

No evaluation on "raw" data. Third, previous studies[13] argued that DL models should be evaluated on voxel-level brain imaging data rather than ROI-based or slice-based MRI, as they are originally conceived to extract features to perform complex tasks[9] automatically. Previous studies[13,22] have concentrated their effort on fully preprocessed voxel-based MRI. However, much less research has been devoted to the pre-processing pipeline and its impact on DL performance. Recent findings on brain age[22,40–42] suggest that DL models perform similarly between raw images (with only linear registration and eventually non-brain tissue removal) and fully pre-processed ones (with gray matter extraction, non-linear diffeomorphic registration, and several bias correction steps as performed with Voxel-Based Morphometry (VBM)), suggesting that CNN do not extract extra-information from raw data. This is a major difference with classical vision tasks (*e.g.*, ImageNet classification) since we know that automatic feature extraction of color, shape, and texture is the cornerstone of today's CNN performance. As a result, a fundamental question is whether usual non-linear computationally demanding pre-preprocessing steps can remove non-linear discriminative information for brain disorders that could have been leveraged by DL (*e.g.*, cortical folding patterns). This problem has been rarely addressed for mental disorders such as schizophrenia, bipolar disorder, and autism, especially with large multi-site studies (*e.g.*, ABIDE, SCHIZCONNECT). Furthermore, several recent works[24,40–43] showed that the prediction capacity of CNNs on images from never-seen sites is worse when using raw data than VBM as pre-processing for both age prediction[40–42], Alzheimer's diagnosis[24] and sex prediction[42,43]. This suggests that CNNs probably overfit acquisition sites using raw data rather than extracting discriminative information. This point is critical since most large-scale clinical datasets that arise in the neuroimaging field are highly multi-centric (*e.g.*, ABIDE, SCHIZCONNECT, ENIGMA[20]).

No evaluation of transfer learning strategies. Finally, probably the most important difference between DL and SML models is the ability of the former to learn a generalizable representation from a large dataset that can be transferred to other tasks they were not trained on (*i.e.*, Transfer Learning and Self-Supervised Learning). Initiated by the work of Caruana et al.[44] on transfer learning and multi-task learning, this idea has been first successfully applied to natural images[45,46] (by re-using features first learned on ImageNet[47], a large-scale dataset with $N > 10^6$ images and 1000 categories), and then to medical datasets[48,49] (by pre-training a CNN on unlabeled medical images in a self-supervised manner). While this idea has been discussed in recent works[13,38] (considering the availability of large brain MRI datasets of healthy subjects, *e.g.*, HCP[6] or UKB[5]), very few studies[1] have evaluated this approach on brain disorder classification tasks, remaining mainly limited to age and sex prediction tasks[50].

To summarise previous studies, there is no consensus on the superiority of deep learning for individual prediction tasks.

While Schulz et al.[14] only provided a partial analysis on age and sex prediction, Abrol et al.[13] extended their findings on these two tasks, arguing that DL was able to outperform SML. Both works remained mainly limited to the same prediction tasks (age and sex prediction), and they provide empirical evidence from the same benchmarking resource (UKB). In this work, we propose to investigate more clinically relevant tasks using a different neuroimaging data set for comparing DL learning capacity against SML. We also aim to explore new learning strategies for DL based on Transfer Learning (TL) that was not investigated in previous studies.

Contributions. We propose to revisit and extend the analysis initiated in recent works[13,14] to large multi-site datasets. We perform extensive experiments to compare DL vs. linear and kernel-SVM (i.e., SML) models on five supervised tasks (age, sex prediction and brain disorder diagnosis) using one of the largest multi-site clinical dataset to date. We investigate pre-processing of anatomical data (VBM and quasi-raw), data augmentation for DL models, features reduction for SML models (Gaussian Random Projection, Univariate Feature Selection, Recursive Feature Elimination) and cross-site generalization both in the medium-scale ($n \approx 1$k) and large-scale ($n \approx 10$k) data regime. Unlike previous literature, we also consider three main transfer learning strategies for mental disorders classification with DL based on self-supervised pre-training, generative modelling and supervised pre-training (see Fig. 1). Finally, we consider Deep Ensemble technique to quantify uncertainty in deep models and we analyze its impact on prediction.

In summary, in this work, we are interested in digging into key questions for neuroimaging: can current SOTA DL models extract non-linearities from highly multi-center brain disorder datasets? How do they scale compared to standard machine learning models? Can we transfer a brain representation of the healthy population to better discriminate patients with mental disorders?

2 Methods

2.1 Data

All data have been collected through various data-sharing initiatives, consortiums, and platforms that can be consulted in the dedicated papers and webpages accessible through hyperlinks in Table 1. We have reported the most important demographic information in Table 1 for all datasets. Importantly, since we acknowledged that reproducibility is critical for all ML/DL studies, we have also integrated the OpenBHB dataset recently released[51] that can be found here. The testing splits used for both age and sex prediction are defined using only data from OpenBHB, for reproducibility purpose, as described in section 2.4.

2.2 VBM and quasi-raw pre-processing

VBM pre-processing is performed with CAT12[53] from the SPM toolbox, essentially consists of noise and bias-field correction followed by Gray Matter (GM), White Matter (WM), and Cerebrospinal Fluid (CSF) segmentation. Images are non-linearly aligned to the MNI template with DARTEL[54] and modulated using the Jacobian deformation field map. All sMRI scans are re-sampled to have an isotropic 1.5mm^3 spatial resolution with dimension $121 \times 145 \times 121$ using a linear spline interpolation. Going to a higher spatial resolution would have induced a bigger computational burden and considering the difference in scanner parameters in our cohorts (e.g., permanent magnetic field), we decided to fix this resolution for all images. We also normalized all images using the Total Intracranial Volume (TIV) estimated by CAT12 to account for the (irrelevant) differences in head size.

As opposed to VBM, quasi-raw pre-processing was designed to be minimal. Only essential steps have been kept to map the images from different sites and scanners to the same space with the same resolution, and only important image correction steps have been applied. Specifically, each scan is rigidly re-oriented to the MNI space and then re-sampled to a 1.5mm^3 spatial resolution through a linear spline interpolation. The bias field is corrected using the N4ITK algorithm[55] from ANTs[56], and the brain is extracted with BET2[57] (the skull and non-brain tissues are removed). Each image is linearly registered (9 degrees of freedom) to the MNI template with FLIRT from FSL[58]. During the training of DL models, we normalize each quasi-raw image by subtracting its mean and dividing by its standard deviation computed across the voxels in each volume.

For all pre-processed images, we applied a visual quality check and removed images poorly segmented or with obvious MRI artefacts.

VBM images provide volumetric information about gray matter density in each voxel which are good predictors of phenotype[13,14,22,29]. However, original raw MR images may contain more information than VBM, in particular related to cortical folding patterns, which may be predictive of psychiatric disorders (e.g. gyrification index[59]). This suggests that raw images could bring more discriminative information than VBM images. We aim to elucidate whether DNN can extract such complementary patterns and consequently achieve better performance.

	Datasets	Disease	# Subjects	# Scans	Age	Sex (%F)	# Sites	Accessibility
OpenBHB[51]	IXI	-	559	559	48 ± 16	55	3	Open
	CoRR	-	1366	2873	26 ± 16	50	19	Open
	NPC	-	65	65	26 ± 4	55	1	Open
	NAR	-	303	323	22 ± 5	58	1	Open
	RBP	-	40	40	22 ± 5	52	1	Open
	GSP	-	1570	1639	21 ± 3	58	5	Open
	ABIDE I	ASD	567	567	17 ± 8	12	20	Open
		HC	566	566	17 ± 8	17	20	Open
	ABIDE II	ASD	481	481	14 ± 8	15	19	Open
		HC	542	555	15 ± 9	30	19	Open
	Localizer	-	82	82	25 ± 7	56	2	Open
	MPI-Leipzig	-	316	317	37 ± 19	40	2	Open
	HCP	-	1113	1113	29 ± 4	45	1	Restricted
	OASIS 3	Only HC	578	1166	68 ± 9	62	4	Restricted
	ICBM	-	606	939	30 ± 12	45	3	Restricted
BIOBD[52]		BD	306	306	40 ± 12	55	8	Private
		HC	356	356	40 ± 13	55	8	Private
SCHIZCONNECT-VIP		SCZ	275	275	34 ± 12	28	4	Open
		HC	329	329	32 ± 13	47	4	Open
PRAGUE		HC	90	90	26 ± 7	55	1	Private
BSNIP		HC	198	198	32 ± 12	58	5	Private
		SCZ	190	190	34 ± 12	30	5	Private
		BD	116	116	37 ± 12	66	5	Private
CANDI		HC	25	25	10 ± 3	41	1	Open
		SCZ	20	20	13 ± 3	45	1	Open
CNP		HC	123	123	31 ± 9	47	1	Open
		SCZ	50	50	36 ± 9	24	1	Open
		BD	49	49	35 ± 9	43	1	Open
Total			10882	**13412**	32 ± 19	50	101	

Table 1. Demographic information about the datasets used throughout this study. We integrated OpenBHB, a large multi-site sMRI dataset freely available here from which we have drawn our training set until $N_{train} = 5000$ and our internal and external testing sets for all our experiments on age and sex prediction.

2.3 Machine learning pipeline for phenotype prediction

First, we wanted to confirm the results obtained by several studies[13,22] on age and sex prediction from anatomical data, as we increase the number of training samples N_{train}, for both DL and SML, but with several key differences: i) we apply no feature selection[32,60] strategy on *both* DL and SML, as we observed a strong degradation in performance with the experimental design previously used in[13,14] (see section E in Supplementary, in line with[32,60]); ii) we separately predict age and sex to avoid arbitrary age discretization ; iii) we assess the generalization performance on an external test, including never-seen sites, and an internal test set stratified on age, sex, and site (see section 2.4 hereafter). Using an external test allows us to give unbiased results since the model cannot make predictions based on confounding variables related to site information[42]. Then, we explore DL performance compared to SML models on three increasingly difficult binary classification tasks for psychiatric diagnosis, including patients with schizophrenia, bipolar disorder, and ASD. Importantly, these three tasks do not have the same difficulty (at least with SML[32,61]), and one might expect improvement with non-linear models on harder tasks where SML models under-perform (*e.g.*, in ASD[62]). We pooled a large number ($n = 19$) of datasets covering a wide age range (from childhood to elderhood) and balanced between males and females (see section 2.1).

2.3.1 SML models

We considered two linear models with ℓ_2 and $\ell_1 + \ell_2$ penalization to promote parsimonious and shrunk solutions, along with Radial-Basis Function Kernel SVM (rbf-SVM). These models have been commonly used in the literature[13,14,32] and

consistently resulted in similar performance, even when additional kernel functions were included during cross-validation (e.g., polynomial or sigmoidal). We also explore three feature selection strategies (Gaussian Random Projection, Univariate Feature Selection, Recursive Feature Elimination) described in Supplementary E but they systemically result in lower performance so we do not apply them in our main analysis.

2.3.2 DL architectures

Due to the lack of standard benchmarks in the neuroimaging field, there is still no consensus about the DL architectures adapted to our downstream tasks. We focused our analysis on SOTA CNN architectures and Transformers as they consistently resulted in top performance on image recognition tasks. Specifically, we chose a classical 3D-AlexNet[63] architecture, as defined by Abrol et al.[13], consisting of five convolutional layers. This network was called "DL1" by Abrol et al. and was used in most of their experiments. To use recent advances in the DL field, we also retained 3D-ResNet18[64] and 3D-DenseNet121[65], similar to recent works that have used structural neuroimaging data[1]. The latter network has 121 layers and is the deepest network used in this paper. Finally, we also compared Transformer-based architecture and smaller CNN backbones in Supplementary D but they systematically under-performed compared to the three models selected in this study. All networks have been implemented in Python and the code is available here.

2.4 Cross-validation procedure and training splits

For age regression and sex prediction, we have built a multi-site dataset including both OpenBHB (see Table 1) - a public dataset that can be accessed without further authorizations- along with more restricted datasets: HCP[6], OASIS 3[66] (only Healthy Controls, HC), ICBM[67], BIOBD[52] (only HC), SCHIZCONNECT-VIP[1] (only HC), and BSNIP[68] (only HC). Eventually, we gathered $N = 11210$ scans from 8679 participants and $n = 99$ sites. We first derived an external test dataset with MPI-Leipzig and NAR ($N_{test}^{inter} = 640$ from 619 participants distributed across the lifespan from $n = 3$ sites). Then, from OpenBHB, we derived an age/sex/site-stratified internal test dataset and a stratified validation dataset with respectively $N_{test}^{intra} = 662$ scans from 480 participants and $N_{val} = 655$ scans from 482 participants. The remaining training set includes $N_{train} = 9253$ scans from 7098 participants. Importantly, each participant appears in only one split so that we avoid any data leakage from the validation/test set. We chose to use validation/test set only from OpenBHB to promote reproducibility in our work[2]. Finally, we sub-sampled this training set in a stratified manner (on age, sex and site) in order to compute performance at varying training sample size ($N \in [100, 500, 1000, 3000, 5000, 9253]$) for both age and sex prediction using a Monte-Carlo Cross Validation (CV) procedure, similarly to[13,14]. We repeated this sub-sampling 5 times for $N \leq 500$ and three times otherwise to keep a reasonable computational budget while still deriving a consistent estimator of classifiers' performance. About schizophrenia, bipolar disorder, and autism detection, we detailed the splits used in Table 2. We used the same splits for all models (SML and DL) and repeated each experiment 30 times, using different random initialization and reporting the average and standard deviation.

Task	Split	Datasets	# Subjects	#Scans	Age	Sex(%F)
SCZ vs. HC	Training	SCHIZCONNECT-VIP, CNP PRAGUE, BSNIP, CANDI	933	933	33 ± 12	43
	Validation		116	116	32 ± 11	37
	External Test		133	133	32 ± 12	45
	Internal Test		118	118	33 ± 13	34
BD vs. HC	Training	BIOBD, BSNIP CNP, CANDI	832	832	38 ± 13	56
	Validation		103	103	37 ± 12	51
	External Test		131	131	37 ± 12	52
	Internal Test		107	107	37 ± 13	56
ASD vs. HC	Training	ABIDE 1+2	1488	1526	16 ± 8	17
	Validation		188	188	17 ± 10	17
	External Test		207	207	12 ± 3	30
	Internal Test		184	186	17 ± 9	18

Table 2. Training/Validation/Test splits used for the 3 mental illness disorders detection. Out-of-site images always make the external test set, and each participant falls into only one split, avoiding data leakage. The internal testing set is always stratified according to age, sex, site, diagnosis, and training and validation set. All models use the same splits.

[1] schizconnect.org
[2] A first version is available here

2.5 DL and SML training

We performed a grid search for SML models to choose the best values of the hyperparameters using the full training set for all tasks. Specifically, for Logistic and Ridge Regression, we tuned the regularization term α within the values $[10^{-1}, 1, 10, 10^2, 10^3]$ and for ElasticNet, we also tuned the ℓ_1 ratio term within the values $[0.1, 0.5, 0.9]$. As for rbf-SVM, we tuned the *gamma* parameter within the values $[10^{-1}, 1, 10, 100]$ for both classification and regression problems.

We implemented all DNN networks with the PyTorch[69] library and SML models with the scikit-learn library[70]. Similarly to Abrol et al.[13], we used the Adam[71] optimizer to perform Stochastic Gradient Descent (SGD) with a weight decay fixed to 10^{-5}. We tuned the learning rate α within the values $[10^{-3}, 10^{-4}, 10^{-5}]$ for all regression and classification tasks with the maximum number of training samples each time, finding that $\alpha = 10^{-4}$ was a good value for all DNN. We then cross-validated the hyper-parameter $\gamma \in [0.2, 0.4, 0.8]$ by decreasing the initial learning rate α every 10 epochs for all DNN and tasks. For computational reasons, we set the batch size b equal to $b = 32$. We optimized all DNN for 300 epochs on age and sex prediction and 100 epochs for diagnosis classification. While we did our best to cross-validate critical hyper-parameters for DL models, we could not reasonably test all hyper-parameters with grid-search (e.g., non-linearities, optimizers, etc). This is a fundamental challenge when working with DL since we optimize highly non-convex functions with many local minima. It motivated the apparition of standard benchmarks in computer vision (such as ImageNet) that allowed easy reproductibility and comparison between SOTA models. Yet, such a benchmark is urgently required for the neuroimaging community, but we did our best to obtain strong baselines for all SML and DL models (in line with recent studies on the same topic[13,14,20,32]).

2.6 ComBat and Linear Adjusted Regression

As reported in several multi-site studies[42,43], the high heterogeneity between scanners and acquisition protocols leads ML models to under-perform on cross-site images (i.e., coming from other sites than the ones used during training). This also explains why we carefully introduced an external test to evaluate models generalization performance in this study. Here, we leverage two SOTA harmonization methods to remove non-biological variance: ComBat[72,73] and Linear Adjusted Regression. These two methods directly harmonize the data without changing the model (as opposed to recent methods[74] that act on DL representations), allowing for a fair comparison between SML and DL methods. Both ComBat and Linear Adjusted Regression need image statistics on all sites to remove site information. However, in our case, only the training and internal test set contain the same sites, so we only residualized these two sets, leaving the external test set unchanged.

Linear Adjusted Regression is a linear harmonization method that tries to preserve biological variability from the data while removing non-biological effects (such as site effect). The model itself can be expressed as[42]:

$$\begin{cases} Y_{ijf} = \alpha_f + \gamma_{if} + \boldsymbol{\beta_f}^T \mathbf{k}_j + \varepsilon_{ijf} \\ \mathscr{Y}_{ijf} = Y_{ijf} - \hat{\gamma}_{if} \end{cases}$$

where Y_{ijf} is the voxel value for site i, subject j, voxel f; α_f is an average measure for voxel f, γ_{if} is the site effect, $\mathbf{k}_j$ is the vector of biological variables we want to keep for subject j (i.e. age, sex, and diagnosis eventually), β_f are parameters estimated by linear regression and ε_{ijf} the residual noise. $\mathscr{Y}_{ijf}$ is the residualized voxel value, where $\hat{\gamma}_{if}$ is the estimated site effect. The parameters γ_{if} and β_f are estimated during training.

Differently, ComBat[73] adds a multiplicative non-linear effect δ_{if} on the residual noise, which brings to a different residualization scheme that also requires the biological variables $\mathbf{k}_j$:

$$\begin{cases} Y_{ijf} = \alpha_f + \gamma_{if} + \boldsymbol{\beta_f}^T \mathbf{k}_j + \delta_{if}\varepsilon_{ijf} \\ \mathscr{Y}_{ijf} = \frac{Y_{ijf} - \hat{\alpha}_f - \hat{\boldsymbol{\beta}}_f^T \mathbf{k}_j - \hat{\gamma}_{if}}{\hat{\delta}_{if}} + \hat{\alpha}_f + \hat{\boldsymbol{\beta}}_f^T \mathbf{k}_j \end{cases}$$

These models generally require to have access to all imaging sites during training. In our experimental design, this was possible only when using the internal test set but not when using the independent external test set. To avoid possible data leakage during residualization, we propose to set $\delta_{if} = 1$ and $\gamma_{if} = 0$ for all unknown test sites i in both linear adjusted regression and ComBat. This is not ideal, and other DL-based[74,75] solutions are starting to emerge in the literature but there is still no consensus, and most of the current studies use ComBat or Linear Adjusted Regression[76,77].

2.7 DL and SML models interpretation

While DL models are often considered "black box" models, several interpretability methods have been proposed over the years to highlight the image areas that have been important for the model to make its decision (see this recent paper[78] for a comprehensive survey). Here, we aim at elucidating whether DL (trained from scratch) and linear models make their decision based on the same brain region patterns, which is a critical question for precision psychiatry.

In this regard, linear models are much simpler to interpret since we have direct access to the weighted maps (or importance

maps[77]). In a weighted map, each weight is associated to a unique input feature. Higher absolute weight values indicate stronger importance of the corresponding input features on the final prediction score. In particular, in a clinical context with anatomical images, hypertrophy (resp. atrophy) in regions with high positive (resp. negative) weights translates into a stronger brain signature for a given pathology, i.e., a higher predictive score.
To generalize to the non-linear case, we have chosen a gradient-based method[79] for DL model interpretability. This sensitivity analysis computes the gradient of predicted output w.r.t. each input voxel (*i.e.*, it quantifies how much output prediction value varies depending on input voxel value). More sophisticated gradient-based models have been proposed over the years, but they do not necessarily result in more accurate saliency maps[80]. Similarly to Abrol et al.[13], we compute brain region importance maps using the Automated Anatomical Labeling atlas[81] (AAL) for each model trained with the maximum number of samples on each task. Specifically, a weighted map is computed through sensitivity analysis for each input image, and all absolute values are summed per region. The resulting importance map is normalized so that it sums to one. Finally, all importance maps for each test set (internal and external) are averaged. We compute the correlation matrix between all averaged maps to compare region importance obtained with SML and DL models.

2.8 Deep ensemble learning

Deep Ensemble for DNN uncertainty quantification. In a real-world scenario where an AI tool is implemented in a hospital, knowing the uncertainty associated with a prediction allows the clinician to trust (or not) the system. It is crucial, especially for computer-aided diagnosis or clinical trial design, as an over-confident system could highly bias an expert's opinion over incorrect predictions based on MRI screening. Additionally, knowing when the prediction is likely to be incorrect (e.g., for out-of-domain images) may improve performance since it allows the system to "*go beyond binary statements on existence vs. non-existence of an effect; and afford credibility estimates around all model parameters at play, which thus enable single-subject predictions with rigorous uncertainty intervals*[82]". In this regard, Bayesian models (such as MC-Dropout[83]) and Deep Ensemble learning[84] have been developed for quantifying predictive DNN uncertainty. A recent benchmark[85] has shown the superiority of the latter over the former and, considering its simplicity, we adopted this framework for brain disorder classification.

Previous deep models do not integrate any notion of uncertainty inside their prediction. Once trained, they estimate the predictive distribution $p(y|x,\mathscr{D})$ for any input image x, given a training set $\mathscr{D}$ (where y represents the clinical status). However, modern DNN tend to be over-confident in their prediction[86], highly limiting their reliability and their clinical use. Yarin Gal[87] introduced the notion of *epistemic* uncertainty to quantify the uncertainty associated to model's weights θ inside DNN. Lakshminarayanan et al.[84] showed that Deep Ensemble provides a simple way to quantify this uncertainty by aggregating several DNN output $p(y|x,\theta^{(t)})$ trained with Stochastic Gradient Descent (SGD) from different random initialization. The averaged distribution $\hat{p}(y|x,\mathscr{D}) = \frac{1}{T}\sum_{t=1}^{T} p(y|x,\theta^{(t)})$ for T trained DNN can be seen, from a Bayesian perspective, as a posterior distribution estimation of $p(y|x,\mathscr{D})$ through Monte-Carlo sampling $\theta^{(t)} \sim p(\theta|\mathscr{D})$.

Implementation. As shown by Lee et al.[88], deep ensemble learning with independently trained neural networks on the whole dataset benefits much more than bagging regarding accuracy and calibration. As a result, in our study, we use the standard deep ensemble strategy often used in DL: we train each network with a different random seed each time and perform stochastic gradient descent on the whole training set. Then, for the regression task (resp. classification task), the output values (resp. probabilities computed after *softmax*) of all networks are averaged. This strategy encourages diversity in learned DL representations without sharing weight between networks. While increasing the number T of independently trained networks can increase this diversity, it is computationally costly. As a trade-off between performance, computational time, and memory, we fixed $T = 3$ in our experiments.

2.9 Pre-training strategies

Deep models have several key advantages over SML besides leveraging raw data. Since DL should be able to learn both low- and high-level imaging features relevant to a given task, it has been hypothesized that at least part of this information could be important for other tasks or domains. Transfer Learning[44–46,89] was grounded on this idea, and it has achieved good performance using both natural and medical images[46,48,90]. Closely related to this idea, in a recent study on resting-state fMRI, He et al.[91] showed how an ML system trained to predict a large bank of phenotypes (e.g., cognition or blood biomarkers) can boost the prediction of correlated, but distinct, set of phenotypes on UKBioBank[5]. As suggested by a recent study[13], predicting phenotype or demographic information in the large-scale data regime may be achieved by a DNN to significantly outperform SML (e.g., for age regression). It suggests that non-linear patterns related to variables non-specific to a pathology are discovered from brain imaging. The discovery of these non-specific axes of variance should allow the learning, in a second phase, of the specific variability associated with mental disorders.

We propose to use a new paradigm depicted in Fig. 1 to train a DNN to discriminate mental disorders from controls. In the first pre-training step, we pre-train a DNN on brain MRI of the healthy population (from childhood to elderhood) to learn a representation capturing the biological and environmental variability of the healthy brain. This can be achieved with a

large-scale dataset. Then, in the second step, the network is fine-tuned to predict the mental condition from brain MRI. Our main assumption is that the representation learned during pre-training will help to discover the pathological variability related to specific mental conditions.

We explore five pre-training strategies to learn anatomical features from the healthy population before applying transfer learning to clinical datasets: 1) our proposed weakly self-supervised model that integrates participant's age as auxiliary information–namely Age-Aware Contrastive Learning[1], 2) self-supervised contrastive learning (SimCLR[2]) 3) SOTA self-supervised model for medical imaging based on context-based restoration (Model Genesis[49]) 4) Variational AutoEncoder (VAE[3]) considered as SOTA generative model (easier to train than GAN[92] and integrating an encoder that can be fine-tuned) and 5) a discriminative supervised model trained on age prediction. Importantly, age information is only used during pre-training of age-aware CL and supervised models but it is never used during fine-tuning. All these models are pre-trained on OpenBHB (with also HCP, ICBM and OASIS3 to increase the dataset size and without ABIDE to avoid data leakage on ASD prediction). This dataset is international, lifespan, and highly multi-centric, promoting heterogeneity in the population under study as well as in image quality. To cross-validate the hyper-parameters, we derived the same validation set as we did for age and sex prediction (stratified on age, sex and site). We provide a detailed description of these five strategies hereafter.

2.9.1 Self-supervised learning

Age-aware contrastive learning. To learn a brain representation of the healthy population, we have developed a new self-supervised algorithm[1], built on the recent development in contrastive learning[2,93]. In particular, this algorithm is able i) to encode invariance to a set of image transformations $\mathscr{T}$ and ii) integrate phenotype information (in our case, participant's chronological age) to enforce images with close phenotype to have close representation in the DL space. The set $\mathscr{T}$ is chosen according to the exploratory work[1] we performed on psychiatric disorders. In our case, $\mathscr{T}$ consists of random cutout, i.e., a black patch covering $1/16$ of the input image is applied to a random location. Two brain images with small missing parts from the same individual still share most of their anatomical features. Consequently, property (i) enforces the encoder to map these two images to the same point in the representation space. To ensure property (ii) is fulfilled, we used a Radial Basis Function kernel to measure the similarity between two chronological ages. We optimized Age-Aware InfoNCE loss as described in[1]. σ^2 was cross-validated in $\{1,2,3,5\}$. Similarly to our previous work[1], we used DenseNet121 as DL encoder, and a 2-layers MLP as non-linear projection head (see our code). We set the batch size to $b = 64$.

After pre-training with Age-Aware InfoNCE loss, we fine-tuned the encoder on each downstream task by cross-validating the learning rate α and scheduler hyper-parameters γ in the same way as before with DL models trained from scratch (see section 2.5). A randomly initialized linear layer is added on top of the pre-trained encoder and trained end-to-end on each downstream task.

Contrastive learning. As a fair comparison with the previous algorithm developed, we have also explored SimCLR[2], a SOTA contrastive learning model adapted for brain MRI. Specifically, we used the same transformations $\mathscr{T}$ (based on cutout) during pre-training, and we trained it for 100 epochs. Since the pretext task is solved quickly (reaching 99% accuracy in less than 10 epochs), we have fine-tuned the pre-trained model after i) 10 epochs, ii) 30 epochs, iii) 100 epochs, and we have cross-validated the optimal γ during fine-tuning and setting the learning rate $\alpha = 10^{-4}$. The best results were obtained using the model pre-trained for 10 epochs, suggesting a rapid over-fit on the training set (even if we reach $\approx$ 10k samples).

Context-based restoration. Context-based restoration is a distinct category of self-supervised models that emerged recently for medical imaging. It can be seen as a special case of denoising autoencoder[94] for representation learning (like inpainting[95]) where the idea is to retrieve the original image from its artificially degraded version using an encoder-decoder neural network. This method mainly requires defining the degrading module and transforming an input image into a degraded, transformed version. It is worth noting that degraded images need not be realistic but rather hide/transform important semantic information that could be deduced from its surrounding context (by analogy with Natural Language Processing where typical self-supervised task consists in retrieving a missing word in a sentence[96]). Model Genesis[49] defines such a module and introduces different strategies to learn context, texture, and appearance. The original formulation leverages UNet backbone (with skip connections between the encoder and decoder) to learn 3D image representations from medical images. We take the same original transformations and backbone to pre-train the network on the same brain MRI dataset as the other methods. We train it for 200 epochs using a learning rate 10^{-4} and Adam optimizer.

2.9.2 Variational Auto-Encoder

VAE[3] is a generative model that uses an encoder-decoder architecture to i) reconstruct an input image from its latent representation and ii) impose a prior distribution in the latent space (generally a Gaussian distribution). Once trained, the VAE can be used either to generate new samples from the known prior distribution or to encode input images through its encoder. One main difficulty encountered during training is to avoid posterior collapse where the posterior latent variable is equal to the prior (thus ignoring the input signal). This is notably due to the non-identifiability issue of the latent variable[97], caused

partly by the model architecture. We used two methods to avoid such behavior: 1) the encoder-decoder architecture is light including only 5 convolutional blocks in the encoder (and a symmetric decoder with transposed convolutions); 2) a β-VAE[98] objective function to restrict the parameters space. β is chosen small ($\beta = 10^{-5}$), and the pre-trained model is validated using linear probing. Linear probing is a simple tool coming from the representation learning field. Here, it consists in training a linear layer on top of the pre-trained VAE encoder to predict the phenotype (age and sex). We hypothesize that if the biological variables can be successfully predicted from the latent representation, the VAE model has learned transferable anatomical brain patterns. Ridge regression is used for predicting age and logistic regression for sex prediction with a regularization term $\alpha \in \{10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}, 10^{3}\}$ cross-validated on the validation set.

2.9.3 Supervised learning

This pre-training strategy is the simplest but also the most widely used in transfer learning[46]: the network is trained to predict a rich signal in a supervised manner on a large-scale database, and we assume that high-level semantic features will be re-used on downstream tasks. In our context, it consists in modeling normal brain aging by training a DNN to predict the age from our large-scale dataset of HC (DenseNet121 in this case). It has two crucial advantages over ImageNet pre-training: i) we do not have a domain gap between natural and medical images, and ii) we can directly transfer to 3D data using 3D DNN. Recent studies[48,99] on transfer learning with medical images suggest that domain gap can hurt performances.

3 Results

3.1 Comparable performance between DL trained from scratch and linear models on psychiatric disorders prediction

We start by evaluating the performance of DL and SML models on the five phenotype prediction tasks across multi-site datasets on VBM data. From Fig. 2, we observe very similar performance on all classification tasks (both sex prediction and diagnosis classification) across all models and even in the very large data regime ($N_{train} > 9000$ for sex prediction). Specifically, all models achieve almost perfect AUC score (Area Under the Curve) on sex prediction on both test sets (AUC $= 98.32\%$ for Logistic Regression and AUC $= 98.47\%$ for DenseNet with $N_{train} = 9253$ on the external test set). While DenseNet is almost always the best-performing network for detecting schizophrenia, bipolar disorder, and autism, it achieves performance on par with Logistic ℓ_2 and rbf-SVM, *i.e.* AUC $\approx 85\%$ on SCZ vs. HC, $\approx 76\%$ for BD vs. HC, and $\approx 65\%$ on ASD vs. HC, on the internal test. DenseNet (like other models) shows poor generalization performance on the external test, losing -10%, -5%, and -1% AUC for SCZ vs. HC, VD vs. HC, and ASD vs. HC, respectively.

As for age regression, we observe that DL outperforms SML only in the large-scale data regime $N_{train} > 9$k on the external test, e.g., ΔMAE $= 0.82$ between AlexNet and ElasticNet. On internal test, DL always outperforms SML, in line with[13,22]. We obtain SOTA performance compared to previous studies (MAE=$2.36_{\pm 0.04}$)[3], which validates the architectural design of DL models (see Supplementary D for more experiments with Transformers). This discrepancy between internal and external tests suggests poor generalization performance on cross-site images due to a large over-fitting on the acquisition site (discussed hereafter).

To further validate our results on age prediction, we replicate our SML analysis pipeline on UKBioBank dataset. Several studies[13,22] have already reported better results from DL models on these data, suggesting an exceptional scaling trend compared to SML. In Supplementary F, we indeed show that DL largely outperforms SML on age regression by 0.9 MAE with $N_{train} = 9253$, but it requires more data than $N_{train} = 100$ samples to achieve such results on the external test, as opposed to what was found on the internal test[22]. It confirms that our SML pipeline is competitive with the current literature on UKBioBank and it extends previous results reported in the literature to cross-site generalization for age regression.

3.1.1 Data augmentation does not improve performance

Considering the small sample size (typically $N \approx 1$k) and high input dimensionality of brain images (> 1M voxels) in previous experiments on psychiatric disorders, data augmentation should provide a simple way to artificially increase the dataset size, limit the over-fit and improve the performance. From the vinicial risk minimization point-of-view, Chapelle et al.[100] showed that it could be seen as a regularization technique that imposes invariance to given transformations for a prediction task. We evaluate five standard augmentations, including geometrical transformations, random noise, cropping, and cutout for all psychiatric disorder classification tasks. We report the results in Supplementary (section A) for VBM and quasi-raw data. Surprisingly, we do not observe significant improvement in performance for DL models for both pre-processings. Therefore, in the rest of this study, we only apply weight decay as regularization technique without data augmentation.

[3] We emphasize that, even if the data size is comparable with previous works, it is not a direct comparison since previous studies used a different test set stratified on UKBioBank.

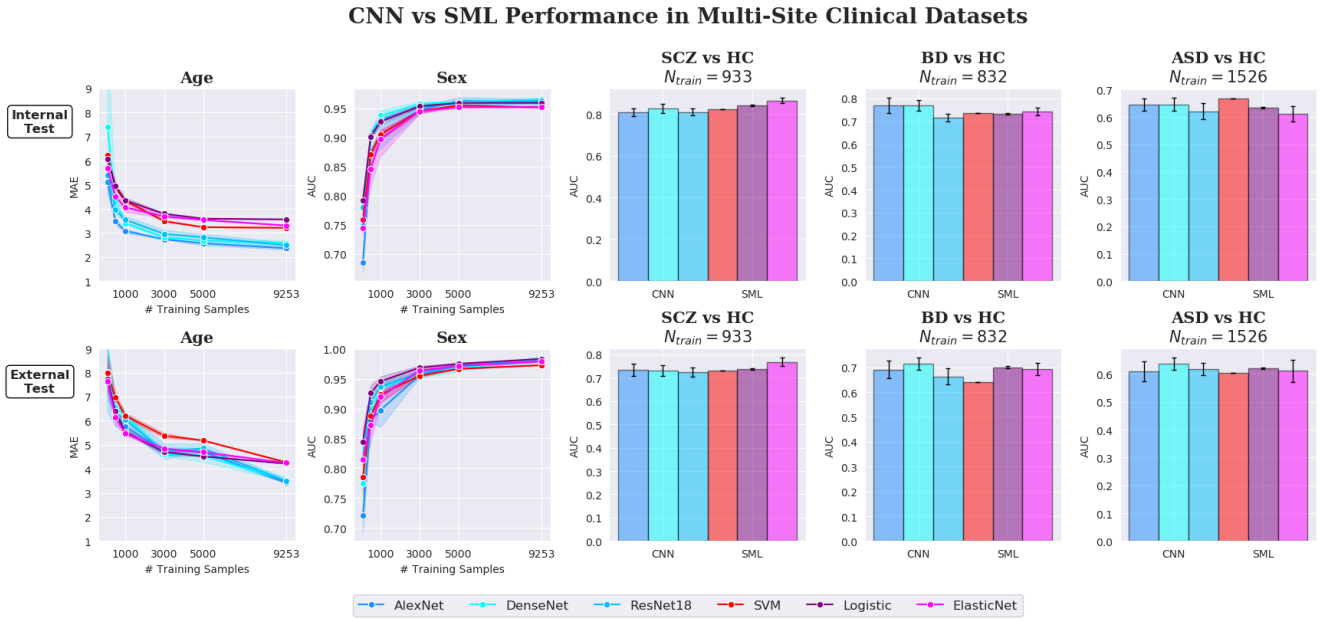

Figure 2. DL vs. SML performance on phenotype prediction and increasingly difficult diagnosis classification tasks on highly multi-site datasets. For SML methods, 2 linear models with ℓ_1 (Logistic) or $\ell_1 + \ell_2$ (ElasticNet) penalization are evaluated, as well as non-linear Radial Basis Function (rbf) SVM. As for DL, vanilla AlexNet[63] (previously introduced by Abrol et al.[13] with 2.5M parameters and 5 layers) and more advanced ResNet18[64] (33.2M parameters, 18 layers) and DenseNet121[65] (11.2M parameters, 121 layers taking advantage from skip-connections and feature re-using) are considered. Both DL and SML algorithms are trained on whole-brain 3D anatomical images. All models are evaluated on two different test sets: an internal test stratified on age, sex, and site ($N^{pheno}_{test} = 662$, $N^{pheno}_{val} = 655$), and diagnosis for clinical datasets ($N^{scz}_{test} = 118$, $N^{scz}_{val} = 116$, $N^{bd}_{test} = 107$, $N^{bd}_{val} = 103$, $N^{asd}_{test} = 184$, $N^{asd}_{val} = 188$); an external test including sites never seen during training ($N^{pheno}_{test} = 640$, $N^{scz}_{test} = 133$, $N^{bd}_{test} = 131$, $N^{asd}_{test} = 207$). Models cannot use site-specific information for their prediction on this test set, eliminating a strong bias reported in the literature. For age and sex prediction, we performed a 5-fold (resp. 3-fold) Monte Carlo Cross-Validation sub-sampling procedure for $N_{train} \in \{100, 500\}$ (resp. $N_{train} \in \{1000, 3000, 5000, 9253\}$). As for diagnosis classification tasks, each model is trained 30 times with different random initialization, and average and standard deviations are reported. Mean Absolute Error (MAE) is the reference measure for age prediction while Area Under the Curve (AUC) is the preferred metric for binary classification tasks since it does not depend on a particular threshold (it only measures a classifier discriminative power). Overall, SML models perform equally well with DL models for sex prediction (up to $N_{train} = 9253$), SCZ vs. HC, BD vs. HC and ASD vs. HC. SML and DL performance keeps improving for age prediction when increasing the number of training subjects N_{train} on the external test. On the other hand, performance increases very slowly (it is almost a plateau) on the internal test starting from $N_{train} \approx 3k$ with an important improvement for non-linear DL models over SML.

3.1.2 Data harmonization produces mitigated results

From Table 9 in Supplementary, we observe that data residualization does not bring improvement for DL models while it marginally improves performance for SML with $N_{train} = 9253$ on age regression. It is not reproducible on external tests (in line with results obtained by Fortin et al. in the original ComBat study[73] on age prediction). However, the difference is more pronounced on psychiatric datasets with a gain of $1-3\%$ AUC overall on the three tasks with SML on internal tests. On external tests, improvement is mitigated especially for BD vs. HC and ASD vs. HC. As for DL models, we observe a significant degradation in performance on both internal and external test sets, indicating that current residualization methods fail to preserve non-linear biological variability extracted by DL models (in line with a recent study on Alzheimer's disease[101]). We perform additional experiments on DenseNet and ResNet, clearly supporting these conclusions; see Table 9 and 10 in Supplementary. Data harmonization techniques for anatomical MRI have been mainly crafted for SML models, and their adaptation to DL is still in its infancy (e.g.[74,102]). Overall, applying data harmonization does not significantly change our main conclusions in the previous section 3.1.

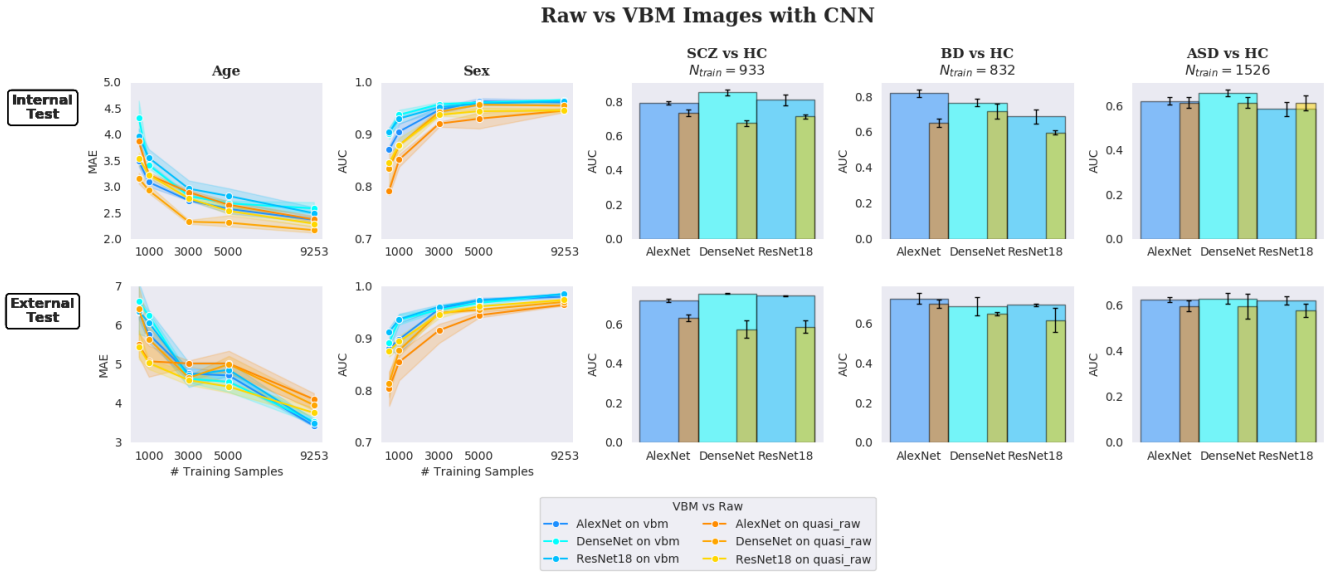

Figure 3. DL performances are evaluated on both raw brain images and extensively pre-processed, non-linearly registered, anatomical Gray Matter (GM) brain images (namely VBM). Raw measurements bring additional geometrical information about cortical folding patterns that may be predictive of psychiatric disorders (e.g., increased gyrification index during childhood for ASD and during adolescence for schizophrenia[59]). Results indicate that DL models fail at extracting more discriminative features from raw images than fully pre-proprocessed ones, even in the large-scale data regime. This observation contrasts with their exceptional automatic feature extraction capacity on natural images.

3.1.3 DL models under-perform on raw data

Fig. 3 shows that DL models under-perform on raw images compared to VBM data for all tasks and testing sets at the current sample size $N_{train} \leq 10$k. The only exception is age prediction on the internal test, but they still poorly generalize to external data compared with models trained on VBM data. It suggests that DL models overfit more on acquisition sites with raw images than VBM. This would prevent them from learning additional geometrical patterns because of the noise inside brain images. We hypothesize that the domain gap between internal and external test for age prediction is more pronounced for raw data than for VBM pre-processed data. To check this hypothesis, we plotted both raw and VBM pre-processed images (from internal and external the test set) encoded by a DenseNet trained on age prediction with $N_{train} = 9253$ (see Supplementary Fig. 7). We used t-SNE[103] visualization to map the embedded images to 2D representations. In the embedded space, we observe a clear difference between raw images coming from either the internal or external test set (especially for middle-aged participants between 20 and 40 years old). This is not the case for VBM images, where inter- and intra-site images overlap correctly in the embedded space for a given age range (blue/orange and yellow/cyan). This greater difference (i.e., domain gap) between internal and external test sets for raw encoded images could explain the differences shown in Fig. 3 for age prediction, supporting the site over-fitting hypothesis.

Additionally, we make an indirect test to check whether noise induced by the scanner explains the discrepancy in results between VBM and raw measurements on psychiatric disorders. From a network trained to predict a given psychiatric condition with a given pre-processing (VBM or raw), we train a linear classifier to predict the acquisition site from the network representation. In Supplementary Table 11, we notably show an increase $> 40\%$ in balanced accuracy (Bacc) on site prediction when the network is trained on raw data rather than VBM to classify psychiatric conditions. From an information bottleneck point-of-view, it suggests that the network fails at compressing disease-related features from raw images and tends to rapidly over-fit on scanner-induced noise.

3.2 Deep and linear models make their decision based on the same brain regions for psychiatric disorders, aging, and gender

Fig. 4 shows two clear patterns, both reproducible across the testing set. First, all DL models generate similar saliency maps to logistic regression with ℓ_2 regularization for all tasks (correlation $r > 0.70$ between the linear model and all DL models for all tasks). This is in line with recent studies[32,77] on SML models applied to age prediction, schizophrenia, and bipolar disorder detection. Both linear and non-linear models resulted in similar final weighted maps with various degrees of noise and sparsity. Second, ElasticNet generates extremely sparse maps (which is expected) but with regions overall poorly correlated with other

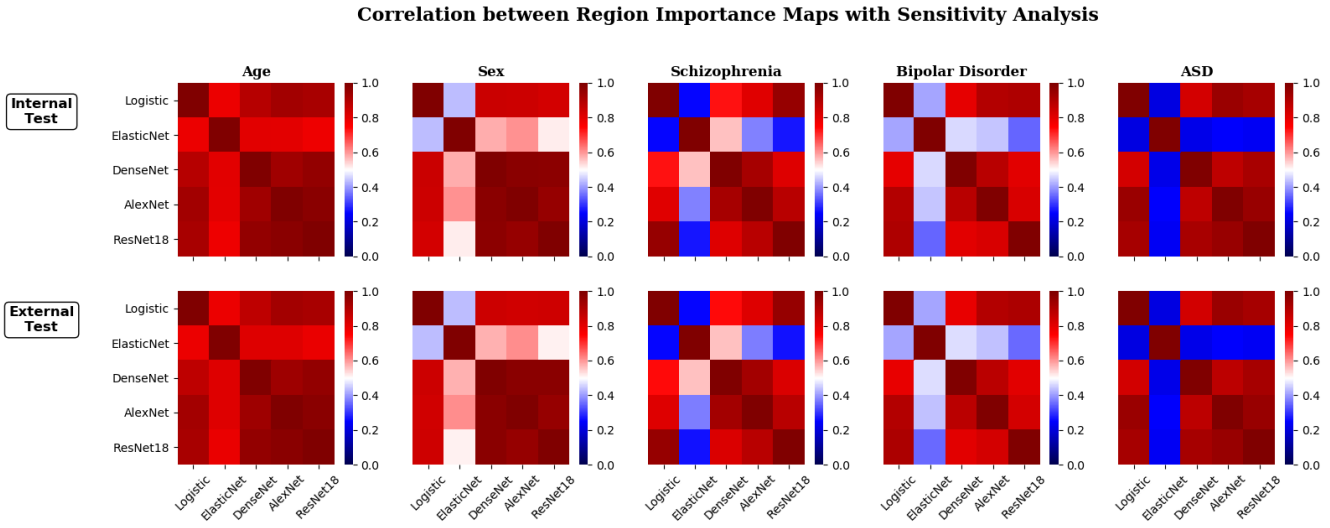

Figure 4. Correlation matrix computed between brain region importance maps obtained for each task and model. A strong correlation indicates a good agreement between two models for a given task. Each brain region importance map is obtained through sensitivity analysis (i.e., using a gradient-based method) for both DL and linear models. All models considered have been trained with the maximum number of training samples. Brain regions are defined through the AAL atlas.

models ($r = 0.21$, $r = 0.22$, $r = 0.25$ and $r = 0.24$ between ElasticNet and Logistic ℓ_2, DenseNet, ResNet and AlexNet resp. on ASD detection). This is more pronounced as we increase the task difficulty (e.g., age or sex prediction with $> 95\%$AUC v.s. ASD detection with $\approx 60\%$AUC). Furthermore, for completeness, we also used an occlusion-based method[104] to compare the saliency maps given by sensitivity analysis and occlusion. Occlusion consists of monitoring the model prediction variation while occluding each brain region independently (defined by the AAL atlas). We reported in Supplementary (Fig. **??**) the correlation between the saliency maps obtained from occlusion vs. sensitivity analysis. Overall, we found an excellent agreement between these two methods ($r > 0.70$ for all models and tasks except AlexNet with sex prediction and DenseNet on bipolar detection).

3.3 Transfer learning and Deep Ensemble improve DL representation to outperform SML for psychiatric disorders

3.3.1 Exploring pre-training strategies

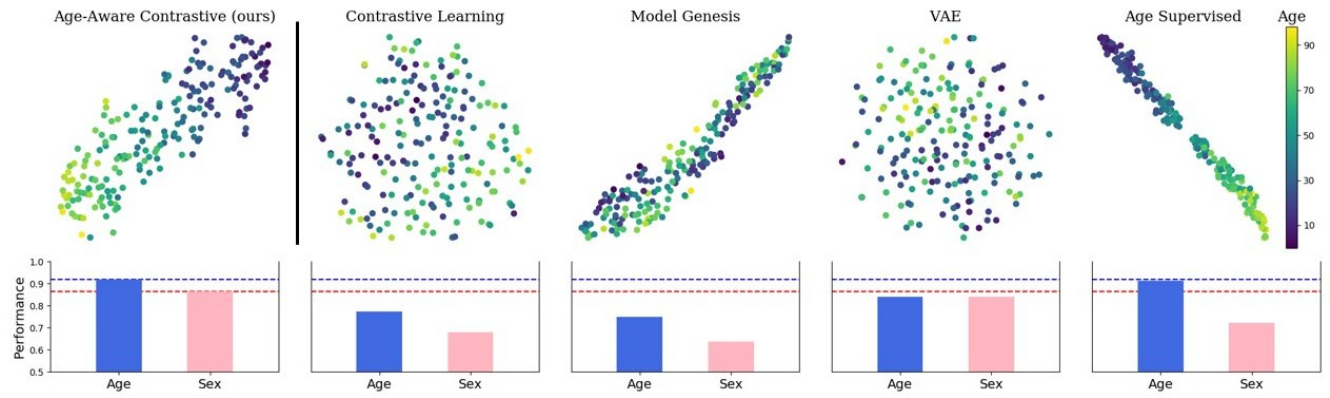

Figure 5. We explore several pre-training strategies based on representation learning applied to brain MRI, among which our proposed model Age-Aware contrastive[1]. We plot t-SNE representation (top) of latent features encoded from new healthy brain images in the external BSNIP dataset (unseen during training). Below, we report the decoding performance to predict demographic information (age/sex) from the latent features (Pearson's correlation for age and balanced accuracy for sex), using linear probing. While Age-Aware contrastive[1] and Age Supervised both use age as weak signal during pre-training, all other models are unsupervised. All models use DenseNet121 backbone except VAE (using a smaller CNN architectures with 5 layers to avoid posterior collapse) and Model Genesis (UNet backbone as in the original formulation[49]).

In Fig. 5, we plot the latent representation of healthy brain images encoded through the pre-trained models described in section 2.9. These brain images come from the external dataset BSNIP, unseen during training. We also report the decoding performance to predict demographic information from latent features, using linear probing. Interestingly, our proposed model Age-Aware contrastive[1] is the only one that captures well both age/sex phenotype, while it has not been trained with sex information. It also has a better decoding performance for age, even compared to fully age supervised model. This can be explained from previous results showing poor generalization performance to new external images with this model i.e., DenseNet121 (see Fig. 2). It suggests that Age-Aware contrastive model encodes robust features independent from scanner. Additionally, VAE also captures well demographic information while it has not been trained with weak supervision. Nonetheless, it still under-performs compared to our proposed model.

Task	**Test Set**		**Pre-training Strategies**				
			Weakly Self-Supervised	Self-Supervised		Generative	Discriminative
		Baseline	Age-Aware Contrastive[1]	Model Genesis[49]	Contrastive Learning[2]	VAE[3]	Age Sup.
SCZ vs. HC ↑	Internal Test	$85.27_{\pm 1.60}$	$\mathbf{85.17}_{\pm \mathbf{0.37}}$	$76.31_{\pm 1.77}$	$82.31_{\pm 2.03}$	$82.56_{\pm 0.68}$	$83.05_{\pm 1.36}$
$N_{train} = 933$	External Test	$75.52_{\pm 0.12}$	$\mathbf{77.00}_{\pm \mathbf{0.55}}$	$67.40_{\pm 1.59}$	$75.48_{\pm 2.54}$	$75.11_{\pm 1.65}$	$74.36_{\pm 2.28}$
BD vs. HC ↑	Internal Test	$76.49_{\pm 2.16}$	$\mathbf{78.81}_{\pm \mathbf{2.48}}$	$76.25_{\pm 1.48}$	$72.71_{\pm 2.06}$	$71.61_{\pm 0.81}$	$77.21_{\pm 1.00}$
$N_{train} = 832$	External Test	$68.57_{\pm 4.72}$	$\mathbf{77.06}_{\pm \mathbf{1.90}}$	$65.66_{\pm 0.90}$	$71.23_{\pm 3.05}$	$71.70_{\pm 0.23}$	$73.02_{\pm 2.66}$
ASD vs. HC ↑	Internal Test	$65.74_{\pm 1.47}$	$66.36_{\pm 1.14}$	$63.58_{\pm 4.35}$	$61.92_{\pm 1.67}$	$59.67_{\pm 2.04}$	$\mathbf{67.11}_{\pm \mathbf{1.76}}$
$N_{train} = 1526$	External Test	$62.93_{\pm 2.40}$	$\mathbf{68.76}_{\pm \mathbf{1.70}}$	$54.95_{\pm 3.58}$	$61.93_{\pm 1.93}$	$57.45_{\pm 0.81}$	$62.07_{\pm 2.98}$

Table 3. Fine-tuning results of models pre-trained with the five previous strategies. All models are pre-trained with only healthy brains. We reported average AUC(%) for all models and the standard deviation by repeating each experiment three times. Baseline is reported from DenseNet121 backbone, giving the best results for mental disorder classification and thus providing strong results.

Transfer to clinical datasets. To further compare these strategies, we fine-tune the different models on the three classification tasks and we report the performance in Table 3. We observe that Age-Aware contrastive model gives the best performance by a large margin (+2%, +4%, +8% AUC resp. on SCZ vs. HC, BD vs. HC and ASD vs. HC, sorted by task difficulty) compared to all other pre-training strategies. Interestingly, adding phenotype information (in particular age) during pre-training (either with discriminative or weakly self-supervised models) allows a boost in performance compared to completely unsupervised models (self-supervised and generative). It notably implies that 1) anatomical knowledge related to age can be transferred to discriminate a wide range of psychiatric disorders and 2) decoder-free self-supervised models provide more robust, reproducible features across sites. Interestingly, a discriminative approach with age prediction as pre-training can well improve performance for ASD. However, it does not replicate on the external test, suggesting an over-fit on scanner. In the following, we have thus used our Age-Aware contrastive model as pre-training.

3.3.2 Knowing what you don't know helps: quantifying DNN uncertainty with Deep Ensemble

Task	**Measure**	*Baseline*	**Deep Ensemble**	
			$T = 2$	$T = 3$
SCZ vs. HC	Calibration (ECE) ↓	$21.05_{\pm 2.27}$	$18.52_{\pm 1.61}$	$\mathbf{15.01}_{\pm 1.06}$
$N_{train} = 933$	Performance (AUC) ↑	$75.52_{\pm 0.12}$	$76.15_{\pm 0.78}$	$\mathbf{77.47}_{\pm 0.71}$
BD vs. HC	Calibration (ECE) ↓	$33.11_{\pm 4.97}$	$27.01_{\pm 2.66}$	$\mathbf{23.56}_{\pm 2.37}$
$N_{train} = 832$	Performance (AUC) ↑	$68.57_{\pm 4.72}$	$74.40_{\pm 1.72}$	$\mathbf{76.11}_{\pm 0.53}$
ASD vs. HC	Calibration (ECE) ↓	$36.54_{\pm 0.87}$	$24.69_{\pm 1.64}$	$\mathbf{22.48}_{\pm 0.92}$
$N_{train} = 1526$	Performance (AUC) ↑	$62.93_{\pm 2.40}$	$63.23_{\pm 2.27}$	$\mathbf{64.48}_{\pm 1.51}$

Table 4. Deep Ensemble improves calibration and performance for all clinical tasks. Calibration is measured by the Expected Calibration Error (ECE) and performance is measured by ROC-AUC. In this experiment, Deep Ensemble model takes the average representation (given after softmax layer) of T models trained with supervision with different random initializations.

In Table 4, we show that quantifying DNN uncertainty through Deep Ensemble allows i) to drastically improve DNN calibration (quantifying whether DNN confidence score for a given prediction can be trusted) and ii) to improve performance

for all psychiatric disorder prediction tasks. We report the results with DenseNet backbone on the external test set and an increasing number of ensemble models T. We observe a significant improvement in calibration for all tasks as we increase the number of ensemble models with -6%, -10% and -14% ECE respectively for SCZ vs. HC, BD vs. HC and ASD vs. HC. Interestingly, calibration was higher for harder task (e.g. ASD) with the baseline model, suggesting that DNN was indeed largely over-confident even when making a high number of mistakes. Additionally, the improvement in calibration systematically goes with an improvement in performance.

3.3.3 Coupling Deep Ensemble and Transfer Learning outperforms SML and achieves SOTA results

Task	Test Set	Deep Learning Models				SML		
		Baseline	Deep Ensemble	Transfer	Transfer + Deep Ensemble	rbf-SVM	Logistic ℓ_2	ElasticNet
SCZ vs. HC ↑	Internal Test	$85.27_{\pm 1.60}$	$85.73_{\pm 0.53}$	$85.17_{\pm 0.37}$	$\mathbf{86.28}_{\pm \mathbf{0.44}}$ (+1.01)	$82.06_{\pm 0.00}$	$84.03_{\pm 0.00}$	$85.98_{\pm 1.9}$
$N_{train} = 933$	External Test	$75.52_{\pm 0.12}$	$\mathbf{77.47}_{\pm \mathbf{0.71}}$	$77.00_{\pm 0.55}$	$76.36_{\pm 0.61}$ (+0.84)	$72.88_{\pm 0.95}$	$73.60_{\pm 0.00}$	$76.42_{\pm 1.68}$
BD vs. HC ↑	Internal Test	$76.49_{\pm 2.16}$	$79.49_{\pm 1.36}$	$78.81_{\pm 2.48}$	$\mathbf{79.59}_{\pm \mathbf{1.77}}$ (+3.10)	$73.63_{\pm 0.00}$	$72.96_{\pm 0.25}$	$73.85_{\pm 0.28}$
$N_{train} = 832$	External Test	$68.57_{\pm 4.72}$	$76.11_{\pm 0.53}$	$77.06_{\pm 1.90}$	$\mathbf{78.01}_{\pm \mathbf{1.97}}$ (+9.44)	$63.92_{\pm 0.00}$	$70.12_{\pm 0.26}$	$70.26_{\pm 1.75}$
ASD vs. HC ↑	Internal Test	$65.74_{\pm 1.47}$	$67.67_{\pm 0.74}$	$66.36_{\pm 1.14}$	$\mathbf{68.48}_{\pm \mathbf{1.45}}$ (+2.74)	$66.84_{\pm 0.00}$	$63.40_{\pm 0.18}$	$60.62_{\pm 2.63}$
$N_{train} = 1526$	External Test	$62.93_{\pm 2.40}$	$64.48_{\pm 1.51}$	$68.76_{\pm 1.70}$	$\mathbf{69.68}_{\pm \mathbf{1.70}}$ (+6.75)	$60.28_{\pm 0.00}$	$61.85_{\pm 0.05}$	$54.96_{\pm 4.94}$

Table 5. Combining Deep Ensemble learning and Transfer Learning improve DL representation over SML models, especially on complex tasks such as ASD and BD detection. We report average AUC for all models and the standard deviation by repeating each experiment three times. We used DenseNet121 as backbone for all DL models. The baseline corresponds to a single network trained from scratch on VBM images. For Deep Ensemble, we aggregate $T = 3$ networks trained from different random initialization. For Transfer Learning, we pre-train a single network with Age-Aware contrastive learning[1] and fine-tune all weights on each clinical task. For Transfer+Deep Ensemble, we aggregate three networks, all pre-trained with Age-Aware contrastive learning (only once) and fine-tuned on each downstream task. The randomness thus comes from the gradient descent optimization on each downstream task. Green numbers indicate improvement over DL baselines.

We present here the results on mental disorder classification when we combine the new paradigm presented in Fig. 1 and the Deep Ensemble strategy previously described. We compare them to SML trained on VBM data (results on residualized data are reported in Supplementary Sec. 3.1.2).

From Table 5, we observe a consistent increase in performance when combining both Deep Ensemble learning and Transfer Learning w.r.t. baseline on the external test ($+0.84\%$, $+9.44\%$, $+6.75\%$ AUC resp. on schizophrenia, bipolar disorder, and autism spectrum disorders detection). For Deep Ensemble learning, it supports the hypothesis that different random initialization leads to different representations after training. For Transfer Learning, it shows that anatomical features learnt from the healthy population during brain maturation and aging can be re-used, in particular to drastically improve DL generalization performance on the external test for hard clinical tasks (i.e bipolar disorder and autism spectrum disorders).

Nonetheless, DL performance is still on par with SML models on easier tasks (e.g., schizophrenia), the task difficulty being measured by linear performance.

Variance analysis. To better explain the performance of TL and Deep Ensemble, we hypothesize that pre-trained models do not escape from the initial basin landscape as randomly initialized model do[105], leading to less variance during model optimization. We have tested this hypothesis on SCZ vs. HC and BD vs. HC by training $n = 30$ independent DNN on each task using the same training set each time but different initialization (random for baseline and pre-trained for transfer and transfer + deep ensemble). We then computed the variance of the performance every 50 epochs across models and we report the standard deviation. We did not run this experiment for ASD vs. HC considering the computational cost (ASD is the largest clinical dataset in this study). Standard deviation is estimated using 30 independent measures for all tasks and models, except for transfer+deep ensemble where it is estimated with 10 measures (since we aggregate three DNN for each measure).

From Table 6, we observe that Transfer+Deep Ensemble offers the lowest variance in all cases (while also being the best performing model, see Table 5). Interestingly, transfer learning drastically lower SD for SCZ vs HC, favouring our hypothesis that solutions are constrained in the same basin landscape, thus confirming previous findings on natural and medical images[105]. It is more mitigated for BD vs HC where Deep Ensemble seems a crucial component to achieve low variance of the models.

4 Discussion

In this study, we have investigated the potential of DL models to extract non-linearities on large-scale and medium-scale multi-site datasets for key problems in neuroimaging including single-subject psychiatric disorders and age/sex prediction, as compared to standard linear and kernel machine learning methods (SML).

Task	Epoch	Baseline	Transfer	Transfer + Deep Ensemble
SCZ vs. HC	10	3.33	2.63	1.50
	50	2.28	1.55	1.11
	100	2.08	1.32	0.98
	150	2.13	1.35	**0.95**
BD vs. HC	10	3.12	3.26	1.58
	50	2.92	2.24	1.68
	100	2.27	2.53	**1.01**
	150	2.05	2.04	1.13

Table 6. Standard Deviation (SD) of AUC performance reported during models optimization, depending on their initialization. TL and TL+Deep Ensemble drastically reduces SD, suggesting that they do not escape much from the initial basin landscape of the loss function. SD is estimated using 30 measures for all pairs (task, model), excepted Transfer+Deep Ensemble where it is estimated with 10 measures (3 models are used for Deep Ensemble).

We first confirm recent findings[14] raising doubts on a universal usage of DL models in anatomical neuroimaging. In particular, we found overall no difference in performance between DL methods trained from scratch and SML for both simple and more complex single subject neuroimaging classification tasks including: 1) sex prediction, 2) schizophrenia detection, 3) bipolar disorder, and 4) autism spectrum disorders classification. Our results on psychiatric disorders extend the ones found in a recent benchmark on Alzheimer's detection[24], showing that DL is on par with simple linear SVM trained on ADNI[106] – the largest neuroimaging initiative to date for Alzheimer's disease (in their case, they comprised $N_{train} = 666$ participants with several time points per participant). Nonetheless, we did find that DL outperforms SML on age regression task, confirming recent studies on this topic[13,22], but it needs a very large number of samples ($N_{train} > 9000$) to extract a better representation than simple regularized linear model when images come from sites never seen during training.

A question then arises: why does DL outperform SML in computer vision on challenging image classification tasks and not on single subject neuroimaging tasks?

A first reason explaining this phenomenon is the highly complex pre-processing pipeline engineered for years in neuroimaging, allowing for noise reduction, spatial alignment, and data harmonization. In particular, diffeomorphic spatial registration as well as brain tissue segmentation and other non-linear image corrections (e.g., bias field correction, intensity rescaling, etc.) have been developed over the last two decades[53,54] for statistical analysis and allow powerful statistical learning with simple linear models. This whole pipeline can be viewed as a complex non-linear function mapping brain raw images to nicely aligned and denoised anatomical images, thus explaining the success of SML (including both linear and kernel methods) in the neuroimaging community. A second obvious reason is clinical data scarcity. Brain imaging produces very large, yet limited, input volumes with $>$ 300k dimensions across no more than a few thousands subjects. It is 1000 times less than ImageNet and with potentially less diversity. A third reason could be related to very high inter-individual heterogeneity in the anatomy of various patients labelled with the same diagnosis, e.g. bipolar disorder or autism[8,20,31]. This last hypothesis is further supported by the current re-conceptualization of major disorders in psychiatry (for instance through the RDoC initiative).

Are brain images too noisy ? The main hypothesis made by Schulz et al.[14] regarding the similar scaling trend between DL and linear models is the linearization of decision boundaries when input images are over-whelmed by noise (e.g. MRI artefacts) unrelated to underlying neurobiological changes related to the pathology. It was well illustrated on the MNIST dataset[107] (grayscale images dataset with handwritten digits ranging from 0 to 9) with a simple experiment: authors[14] added Gaussian noise to the images and, the stronger was the noise, the closer the learning curves were between DNN and linear model. We argue that our experiments on VBM vs. raw images supports this hypothesis. We showed how site-related noise was well preserved in the representation space of DNN trained to predict age/sex or mental condition, especially with raw measurements while we know that more discriminative signal is present. This hypothesis was also supported in the experiments on age prediction in section 3.1: while the learning curve for SML was significantly worse than DL on internal test (reaching a plateau early with $N_{train} = 3k$), it was not the case on external test. These findings suggest that current site-related noise inside MRI prevent DNN models from exploiting non-linear signal, thus somehow linearizing its decision boundary for psychiatric conditions classification.

Transfer Learning from large-scale healthy dataset to medium-scale clinical studies. Crucially, we propose a new transfer learning paradigm for discriminating patients with mental disorders from controls, achieving new SOTA for ASD classification and bipolar disorder detection. This paradigm is versatile and does not specify a particular pre-training strategy.

It mainly relies on the hypothesis that capturing the biological variability in the healthy population related to non-specific variables (e.g. age, sex, etc.) with large-scale dataset allows easier discovery of specific pathological variability (e.g. subtle cortical atrophy in pre-fontal and temporal lobe for ASD detection) during fine-tuning on small-scale cohorts. Our findings with our proposed Age-Aware contrastive strategy suggests that age-related features are also implicated in BD and ASD diagnosis, supporting previous findings on this topic[108,109] (e.g. related to brain overgrowth during childhood). In this regard, integrating other phenotypes (e.g. cognition) during pre-training using y-Aware contrastive learning opens up a new avenue for transfer learning and representation learning. It would allow to shape brain imaging representation according to non-imaging variables and possible learn a richer manifold from large-scale healthy dataset.

Additionally, we also show how uncertainty quantification ("knowing what you don't know") is crucial for DL model, and it can be solved with Deep Ensemble. Considering their over-confidence for solving complex tasks even with noisy data, modelling and quantifying a predictive uncertainty is essential for computer-aided diagnosis and clinical trial design.

Quantitatively, we found that DL, combined with TL, establishes the new state-of-the-art prediction performance on bipolar disorder detection from brain anatomical imaging ($>78\%$ AUC on both internal and external test, with 1173 subjects and 471 patients with BD), in light of recent results from the ENIGMA consortium[20] (the largest to date with 3020 subjects and 853 patients with BD). In their experiments[20], they achieved $\approx 70\%$ AUC (resp.$\approx 75\%$) on external (resp. internal) test after linear residualization adjusted on age, sex and site.

These findings suggest that i) discriminative transferable anatomical non-linear patterns can be learned with DL through pre-training from brain imaging of the healthy population; ii) different DL initialization converge to different solutions after training that, if aggregated together, can outperform SML; iii) DL models tend to learn simple features on easy tasks (such as schizophrenia detection), falling into the Simplicity Bias[110], which encourages DNN to find the simplest features to perform the task (and thus hurting generalization power on external test sets).

Interestingly, for schizophrenia, the easiest clinical task among the three tackled in this paper (relatively to ML diagnosis accuracy), DL struggles to find better representation than simple regularized linear models, even when performing TL or Deep Ensemble learning. We hypothesized that this might be due to the simplicity bias[110] where DL trained with standard training procedures, such as Stochastic Gradient Descent (SGD), tends to rely on the simplest features even if more complex ones could bring more discriminative information. We saw that aggregating different DL representations trained from scratch on SCZ detection leads to marginal improvement ($+0.46\%$ AUC on internal test), as opposed to BD and ASD classification ($+3\%$ and $+2.92\%$ AUC respectively), suggesting that different DL models extract dissimilar (potentially non-linear) features only on complex tasks. This would also explain the performance drop on external test for SCZ vs. HC (-9.92% AUC compared to internal test) viewed as out-of-domain dataset since the simplicity bias leads to poor out-of-domain generalization[110]. This performance drop was only observed on SCZ vs. HC after performing TL and Deep Ensemble. Simplicity bias is a relatively new concept, and removing this bias in current DL models is still an open challenge. We hypothesize that, by avoiding simplicity in DL, we may also benefit from the powerful representation capacity of DL on simpler clinical tasks such as schizophrenia detection.

We acknowledge that current DL architectures may not be ideal for brain anatomical data. On natural images, DL architectures (in particular CNN) bring a strong inductive bias (e.g. translation invariance, hierarchical representation) that seems very beneficial for challenging computer vision tasks, which could partly explain their success. In particular, on MNIST[107] (a highly popular benchmarking image dataset containing handwritten digits), CNN are able to outperform SML (by $>+15\%$ accuracy[14]) with as few as $N_{train} = 100$ samples. Another work[111] also showed that the representation space of a CNN randomly initialized can be used as such to achieve accurate results on MNIST ($>90\%$ accuracy). More remarkably, CNNs randomly initialized (i.e. not trained) can be used as a "handcrafted prior" for image denoising, inpainting, image reconstruction[112], and object localization[113] on ImageNet to achieve SOTA results. On the other hand, we hypothesize that current inductive bias in CNN may not be sufficient for brain anatomical data where all images are already aligned, and share same colors and textures (in line with a recent review[114]). Other recent DL architectures such as Transformers[115], integrating attention modules at its core and relaxing the inductive bias constraints present in CNN, might be another exciting research direction for neuroimaging. While Transformers still require massive amount of data on natural images (because of their flexibility[116]), first works in neuroimaging are starting to appear[117] and should receive special attention.

Our findings demonstrate that DL and SML tend to rapidly over-fit the acquisition sites, even in the large-scale data regime. With age regression problem, we observe a significant performance drop of all DL and SML models between internal and external tests (average drop of MAE: $\Delta\text{MAE(DL)} = 1.00$, $\Delta\text{MAE(SML)} = 0.88$ with $N = 10$k images acquired on 17 sites). Similar drop of classification performances is found with schizophrenia detection (with 1300 samples) $\Delta\text{AUC(DL)} = 7.81\%$, $\Delta\text{AUC(SML)} = 9.72\%$. Such a decrease in performances might be mainly attributed to site acquisition settings. Moreover, this suggests a systematic bias with results obtained on test images that stem from sites that have been seen during the training phase. DL models appear to over-fit even more with raw data than VBM on age regression, explaining their higher performance drop between internal and external test, observed in Fig. 3 and confirmed in Supplementary Fig. 7. This is in line with

the inter-scanner reliability test performed by Cole et al.[40] on DL models. Our results again favor the handcrafted VBM pre-processing also for DL, since it seems to limit the site bias (at least on age regression). Interestingly, similar results were obtained on Alzheimer's detection[24], with poor DL generalization when using raw images coming from never-seen sites.

Overall, this shines a light on a recurrent issue in neuroimaging with multi-site studies related to data harmonization and debiasing in DL. While SOTA data harmonization techniques (Combat[73] and Linear Adjusted Regression) have been partially beneficial for SML on clinical applications, it was not the case for DL (see Table 9 in Supplementary). It suggests that current harmonization techniques still fail at preserving non-linear input relationships leveraged by DL to perform the downstream task. Removing site information from DL representation while protecting for variables of interest (e.g., biological such as diagnosis, age, sex, or sensitive attributes in the context of trustworthy AI) is an open challenge both in computer vision[118] and neuroimaging[74,119,120]. It is still a relatively new research area with no benchmarking datasets nor metrics in neuroimaging.

Often considered as a "black box," we provide empirical evidence that DL models randomly initialized take their decision based on very similar brain regions as compared to linear models. We observed these agreements between DL and linear models on internal and external test sets. This consistency across DL and linear models is reassuring and suggests the reliability of features extracted by DL models. It should also be noted that different CNN models based their decision on highly similar importance maps for all evaluated tasks. DL reliability is crucial in the context of precision medicine for psychiatry as a first step towards building models accepted and trusted by clinicians.

Overall, our study confirms that DL utility over SML on challenging clinical applications in psychiatry comes from TL and Deep Ensemble learning. Coupling these two strategies outperforms SML on both BD and ASD and achieves new state-of-the-art BD results. While DL trained from scratch did not dominate simple linear models on psychiatric disorders, we showed that recent advances in contrastive learning[1,121,122] applied on a large healthy population ($N \approx 10$k) allow DL models to learn strong re-usable features. Aggregating other modalities (e.g., functional and diffusion MRI and genetics) to perform representation learning remains an exciting challenge that might be solved with contrastive learning. It would improve our understanding of brain disorders and possibly pave the way towards personalized medicine in psychiatry through prognostic models of clinical outcome, where only small longitudinal cohorts are, and will be, available in the near future.

References

1. Dufumier, B. *et al.* Contrastive learning with continuous proxy meta-data for 3d mri classification. In *International conference on medical image computing and computer-assisted intervention* (Springer, 2021).
2. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607 (PMLR, 2020).
3. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *ICLR* (2014).
4. Bashyam, V. M. *et al.* Mri signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide. *Brain* **143**, 2312–2324 (2020).
5. Bycroft, C. *et al.* The uk biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
6. Van Essen, D. C. *et al.* The wu-minn human connectome project: an overview. *Neuroimage* **80**, 62–79 (2013).
7. Di Martino, A. *et al.* The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. psychiatry* **19**, 659–667 (2014).
8. Wolfers, T. *et al.* Mapping the heterogeneous phenotype of schizophrenia and bipolar disorder using normative models. *JAMA psychiatry* **75**, 1146–1155 (2018).
9. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
10. He, T. *et al.* Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage* **206**, 116276 (2020).
11. Mellema, C. J., Nguyen, K. P., Treacher, A. & Montillo, A. Reproducible neuroimaging features for diagnosis of autism spectrum disorder with machine learning. *Sci. reports* **12**, 3057 (2022).
12. Dahan, S. *et al.* Surface vision transformers: Attention-based modelling applied to cortical analysis. In *International Conference on Medical Imaging with Deep Learning*, 282–303 (PMLR, 2022).
13. Abrol, A. *et al.* Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nat. communications* **12**, 1–17 (2021).
14. Schulz, M.-A. *et al.* Different scaling of linear models and deep learning in ukbiobank brain images versus machine-learning datasets. *Nat. communications* **11**, 1–15 (2020).

15. He, T. *et al.* Do deep neural networks outperform kernel regression for functional connectivity prediction of behavior? *BioRxiv* 473603 (2018).

16. Quaak, M., van de Mortel, L., Thomas, R. M. & van Wingen, G. Deep learning applications for the classification of psychiatric disorders using neuroimaging data: systematic review and meta-analysis. *NeuroImage: Clin.* **30** (2021).

17. Vieira, S. *et al.* Using machine learning and structural neuroimaging to detect first episode psychosis: reconsidering the evidence. *Schizophr. bulletin* **46**, 17–26 (2020).

18. Hibar, D. *et al.* Cortical abnormalities in bipolar disorder: an mri analysis of 6503 individuals from the enigma bipolar disorder working group. *Mol. psychiatry* **23**, 932–942 (2018).

19. Van Erp, T. G. *et al.* Subcortical brain volume abnormalities in 2028 individuals with schizophrenia and 2540 healthy controls via the enigma consortium. *Mol. psychiatry* **21**, 547–553 (2016).

20. Nunes, A. *et al.* Using structural mri to identify bipolar disorders–13 site machine learning study in 3020 individuals from the enigma bipolar disorders working group. *Mol. psychiatry* **25**, 2130–2143 (2020).

21. Van Rooij, D. *et al.* Cortical and subcortical brain morphometry differences between patients with autism spectrum disorder and healthy individuals across the lifespan: results from the enigma asd working group. *Am. J. Psychiatry* **175**, 359–369 (2018).

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

23. Fisch, L. *et al.* Predicting chronological age from structural neuroimaging: The predictive analytics competition 2019. *Front. Psychiatry* **12** (2021).

24. Wen, J. *et al.* Convolutional neural networks for classification of alzheimer's disease: Overview and reproducible evaluation. *Med. Image Analysis* 101694 (2020).

25. Dinomais, M. *et al.* Anatomic correlation of the mini-mental state examination: a voxel-based morphometric study in older adults. *PloS one* **11**, e0162889 (2016).

26. Koutsouleris, N. *et al.* Accelerated brain aging in schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. *Schizophr. bulletin* **40**, 1140–1153 (2014).

27. Cole, J. H. & Franke, K. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends neurosciences* **40**, 681–690 (2017).

28. Cole, J. H. *et al.* Brain age predicts mortality. *Mol. psychiatry* **23**, 1385–1392 (2018).

29. Jonsson, B. A. *et al.* Brain age prediction using deep learning uncovers associated sequence variants. *Nat. Commun.* **10**, DOI: 10.1038/s41467-019-13163-9 (2019).

30. Marquand, A. F. *et al.* Conceptualizing mental disorders as deviations from normative functioning. *Mol. psychiatry* **24**, 1415–1424 (2019).

31. Zabihi, M. *et al.* Fractionating autism based on neuroanatomical normative modeling. *Transl. psychiatry* **10**, 1–10 (2020).

32. Salvador, R. *et al.* Evaluation of machine learning algorithms and structural features for optimal mri-based diagnostic prediction in psychosis. *PLoS One* **12**, e0175683 (2017).

33. Varoquaux, G. Cross-validation failure: Small sample sizes lead to large error bars. *Neuroimage* **180**, 68–77 (2018).

34. Schnack, H. G. & Kahn, R. S. Detecting neuroimaging biomarkers for psychiatric disorders: sample size matters. *Front. psychiatry* **7**, 50 (2016).

35. Kambeitz, J. *et al.* Reply to: sample size, model robustness, and classification accuracy in diagnostic multivariate neuroimaging analyses. *Biol. psychiatry* **84**, e83–e84 (2018).

36. Pulini, A. A., Kerr, W. T., Loo, S. K. & Lenartowicz, A. Classification accuracy of neuroimaging biomarkers in attention-deficit/hyperactivity disorder: effects of sample size and circular analysis. *Biol. Psychiatry: Cogn. Neurosci. Neuroimaging* **4**, 108–120 (2019).

37. Flint, C. *et al.* Systematic misestimation of machine learning performance in neuroimaging studies of depression. *Neuropsychopharmacology* **46**, 1510–1517 (2021).

38. Koppe, G., Meyer-Lindenberg, A. & Durstewitz, D. Deep learning for small and big data in psychiatry. *Neuropsychopharmacology* **46**, 176–190 (2021).

39. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM* **64**, 107–115 (2021).

40. Cole, J. H. *et al.* Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage* **163**, 115–124 (2017).

41. Hwang, I. *et al.* Prediction of brain age from routine t2-weighted spin-echo brain magnetic resonance images with a deep convolutional neural network. *Neurobiol. Aging* **105**, 78–85 (2021).

42. Wachinger, C., Rieckmann, A., Pölsterl, S., Initiative, A. D. N. *et al.* Detect and correct bias in multi-site neuroimaging datasets. *Med. Image Analysis* **67**, 101879 (2021).

43. Glocker, B., Robinson, R., Castro, D. C., Dou, Q. & Konukoglu, E. Machine learning with multi-site imaging data: An empirical study on the impact of scanner effects. *Med. Imaging meets NeurIPS Work.* (2019).

44. Caruana, R. Multitask learning. *Mach. learning* **28**, 41–75 (1997).

45. Bengio, Y. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, 17–36 (JMLR Workshop and Conference Proceedings, 2012).

46. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. & Weinberger, K. Q. (eds.) *Advances in Neural Information Processing Systems*, vol. 27 (Curran Associates, Inc., 2014).

47. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255 (Ieee, 2009).

48. Azizi, S. *et al.* Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3478–3488 (2021).

49. Zhou, Z., Sodha, V., Pang, J., Gotway, M. B. & Liang, J. Models genesis. *Med. Image Analysis* **67**, 101840 (2021).

50. Malik, N. & Bzdok, D. From youtube to the brain: Transfer learning can improve brain-imaging predictions with deep learning. *Neural Networks* **153**, 325–338 (2022).

51. Dufumier, B. *et al.* Openbhb: a large-scale multi-site brain mri data-set for age prediction and debiasing. *NeuroImage* **263**, 119637 (2022).

52. Sarrazin, S. *et al.* Neurodevelopmental subtypes of bipolar disorder are related to cortical folding patterns: An international multicenter study. *Bipolar disorders* **20**, 721–732 (2018).

53. Gaser, C. & Dahnke, R. Cat-a computational anatomy toolbox for the analysis of structural mri data. *HBM* **2016**, 336–348 (2016).

54. Ashburner, J. A fast diffeomorphic image registration algorithm. *Neuroimage* **38**, 95–113 (2007).

55. Tustison, N. J. *et al.* N4itk: improved n3 bias correction. *IEEE transactions on medical imaging* **29**, 1310–1320 (2010).

56. Avants, B. B., Tustison, N. & Song, G. Advanced normalization tools (ants). *Insight j* **2**, 1–35 (2009).

57. Jenkinson, M., Pechaud, M., Smith, S. *et al.* Bet2: Mr-based estimation of brain, skull and scalp surfaces. In *Eleventh annual meeting of the organization for human brain mapping*, vol. 17, 167 (Toronto., 2005).

58. Jenkinson, M. & Smith, S. A global optimisation method for robust affine registration of brain images. *Med. image analysis* **5**, 143–156 (2001).

59. Sasabayashi, D., Takahashi, T., Takayanagi, Y. & Suzuki, M. Anomalous brain gyrification patterns in major psychiatric disorders: a systematic review and transdiagnostic integration. *Transl. psychiatry* **11**, 1–12 (2021).

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

61. Eslami, T., Almuqhim, F., Raiker, J. S. & Saeed, F. Machine learning methods for diagnosing autism spectrum disorder and attention-deficit/hyperactivity disorder using functional and structural mri: A survey. *Front. neuroinformatics* **14**, 62 (2021).

62. Hoogman, M. *et al.* Consortium neuroscience of attention deficit/hyperactivity disorder and autism spectrum disorder: The enigma adventure. *Hum. brain mapping* **43**, 37–55 (2022).

63. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105 (2012).

64. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

65. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708 (2017).

66. LaMontagne, P. J. *et al.* Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv* (2019).

67. Mazziotta, J. *et al.* A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (icbm). *Philos. Transactions Royal Soc. London. Ser. B: Biol. Sci.* **356**, 1293–1322 (2001).

68. Tamminga, C. A. *et al.* Bipolar and schizophrenia network for intermediate phenotypes: outcomes across the psychosis continuum. *Schizophr. bulletin* **40**, S131–S137 (2014).

69. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019).

70. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. machine Learn. research* **12**, 2825–2830 (2011).

71. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *ICLR* (2015).

72. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8**, 118–127 (2007).

73. Fortin, J.-P. *et al.* Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage* **167**, 104–120 (2018).

74. Dinsdale, N. K., Jenkinson, M. & Namburete, A. I. Deep learning-based unlearning of dataset bias for mri harmonisation and confound removal. *NeuroImage* **228**, 117689 (2021).

75. Torbati, M. E. *et al.* Multi-scanner harmonization of paired neuroimaging data via structure preserving embedding learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 3284–3293 (2021).

76. Radua, J. *et al.* Increased power by harmonizing structural mri site differences with the combat batch adjustment method in enigma. *NeuroImage* **218**, 116956 (2020).

77. Ball, G., Kelly, C. E., Beare, R. & Seal, M. L. Individual variation underlying brain age estimates in typical development. *NeuroImage* **235**, 118036 (2021).

78. Zhang, Y., Tiňo, P., Leonardis, A. & Tang, K. A survey on neural network interpretability. *IEEE Transactions on Emerg. Top. Comput. Intell.* (2021).

79. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

80. Adebayo, J. *et al.* Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292* (2018).

81. Rolls, E. T., Huang, C.-C., Lin, C.-P., Feng, J. & Joliot, M. Automated anatomical labelling atlas 3. *Neuroimage* **206**, 116189 (2020).

82. Bzdok, D., Floris, D. L. & Marquand, A. F. Analysing brain networks in population neuroscience: a case for the bayesian philosophy. *Philos. Transactions Royal Soc. B* **375**, 20190661 (2020).

83. Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059 (PMLR, 2016).

84. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, 6402–6413 (2017).

85. Gustafsson, F. K., Danelljan, M. & Schon, T. B. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 318–319 (2020).

86. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, 1321–1330 (2017).

87. Gal, Y. Uncertainty in deep learning. *Univ. Camb.* **1**, 3 (2016).

88. Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D. & Batra, D. Why m heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint arXiv:1511.06314* (2015).

89. Raina, R., Battle, A., Lee, H., Packer, B. & Ng, A. Y. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, 759–766 (2007).

90. Mustafa, B. *et al.* Supervised transfer learning at scale for medical imaging. *arXiv preprint arXiv:2101.05913* (2021).

91. He, T. *et al.* Meta-matching as a simple framework to translate phenotypic predictive models from big to small data. *Nat. Neurosci.* 1–10 (2022).

92. Goodfellow, I. *et al.* Generative adversarial nets. *Adv. neural information processing systems* **27** (2014).

93. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738 (2020).

94. Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, 1096–1103 (2008).

95. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T. & Efros, A. A. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2536–2544 (2016).

96. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

97. Wang, Y., Blei, D. & Cunningham, J. P. Posterior collapse and latent variable non-identifiability. *Adv. Neural Inf. Process. Syst.* **34**, 5443–5455 (2021).

98. Higgins, I. *et al.* beta-vae: Learning basic visual concepts with a constrained variational framework. *Int. Conf. on Learn. Represent.* (2016).

99. Raghu, M., Zhang, C., Kleinberg, J. & Bengio, S. Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems*, 3347–3357 (2019).

100. Chapelle, O., Weston, J., Bottou, L. & Vapnik, V. Vicinal risk minimization. *Adv. neural information processing systems* 416–422 (2001).

101. An, L. *et al.* Goal-specific brain mri harmonization. *NeuroImage* **263**, 119570 (2022).

102. Bashyam, V. M. *et al.* Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. *J. Magn. Reson. Imaging* **55**, 908–916 (2022).

103. Van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. machine learning research* **9** (2008).

104. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833 (Springer, 2014).

105. Neyshabur, B., Sedghi, H. & Zhang, C. What is being transferred in transfer learning? *Adv. neural information processing systems* **33**, 512–523 (2020).

106. Jack Jr, C. R. *et al.* The alzheimer's disease neuroimaging initiative (adni): Mri methods. *J. Magn. Reson. Imaging: An Off. J. Int. Soc. for Magn. Reson. Medicine* **27**, 685–691 (2008).

107. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).

108. Courchesne, E., Carper, R. & Akshoomoff, N. Evidence of brain overgrowth in the first year of life in autism. *Jama* **290**, 337–344 (2003).

109. Greimel, E. *et al.* Changes in grey matter development in autism spectrum disorder. *Brain Struct. Funct.* **218**, 929–942 (2013).

110. Shah, H., Tamuly, K., Raghunathan, A., Jain, P. & Netrapalli, P. The pitfalls of simplicity bias in neural networks. *Adv. Neural Inf. Process. Syst.* **33**, 9573–9585 (2020).

111. Alain, G. & Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644* (2016).

112. Ulyanov, D., Vedaldi, A. & Lempitsky, V. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9446–9454 (2018).

113. Cao, Y.-H. & Wu, J. A random cnn sees objects: One inductive bias of cnn and its applications. In *Proceedings Of The AAAI Conference On Artificial Intelligence*, vol. 36, 194–202 (2022).

114. Eitel, F., Schulz, M.-A., Seiler, M., Walter, H. & Ritter, K. Promises and pitfalls of deep neural networks in neuroimaging-based psychiatric research. *Exp. Neurol.* 113608 (2021).

115. Vaswani, A. *et al.* Attention is all you need. In *Advances in neural information processing systems*, 5998–6008 (2017).

116. Lin, T., Wang, Y., Liu, X. & Qiu, X. A survey of transformers. *AI Open* **3**, 111–132 (2022).

117. He, S., Grant, P. E. & Ou, Y. Global-local transformer for brain age estimation. *IEEE Transactions on Med. Imaging* **41**, 213–224 (2021).

118. Barbano, C. A., Tartaglione, E. & Grangetto, M. Bridging the gap between debiasing and privacy for deep learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3806–3815 (2021).

119. Barbano, C. A., Dufumier, B., Tartaglione, E., Grangetto, M. & Gori, P. Unbiased Supervised Contrastive Learning. In *International Conference on Learning Representations (ICLR)* (2023).

120. Barbano, C. A., Dufumier, B., Duchesnay, E., Grangetto, M. & Gori, P. Contrastive learning for regression in multi-site brain age prediction. In *IEEE 20th International Symposium on Biomedical Imaging (ISBI)* (2023).

121. Dufumier, B., Gori, P., Victor, J., Grigis, A. & Duchesnay, E. Conditional Alignment and Uniformity for Contrastive Learning with Continuous Proxy Labels. In *MedNeurIPS, Workshop NeurIPS* (2021).

122. Dufumier, B., Barbano, C. A., Louiset, R., Duchesnay, E. & Gori, P. Integrating Prior Knowledge in Contrastive Learning with Kernel. In *International Conference on Machine Learning (ICML)* (2023).

123. Hernández-García, A., Mehrer, J., Kriegeskorte, N., König, P. & Kietzmann, T. C. Deep neural networks trained with heavier data augmentation learn features closer to representations in hit. In *Conference on Cognitive Computational Neuroscience*, vol. 1 (2018).

124. Chadebec, C., Thibeau-Sutre, E., Burgos, N. & Allassonnière, S. Data augmentation in high dimensional low sample size setting using a geometry-based variational autoencoder. *IEEE Transactions on Pattern Analysis Mach. Intell.* **45**, 2879–2896 (2022).

125. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal machine learning research* **15**, 1929–1958 (2014).

126. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, 448–456 (PMLR, 2015).

127. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR* (2021).

128. Steiner, A. *et al.* How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270* (2021).

129. Chen*, X., Xie*, S. & He, K. An empirical study of training self-supervised vision transformers. *arXiv preprint arXiv:2104.02057* (2021).

130. Bethlehem, R. A. *et al.* Brain charts for the human lifespan. *Nature* **604**, 525–533 (2022).

Acknowledgements

This work received funding from French National Research Agency for the project Big2Small (Chair in AI, ANR-19-CHIA-0010-01), the project RHU-PsyCARE (French government's "Investissements d'Avenir" program, ANR-18-RHUS-0014), and European Union's Horizon 2020 for the project R-LiNK (H2020-SC1-2017, 754907). This work was granted access to the HPC resources of IDRIS under the allocation 2023-AD011011854R2 made by GENCI.

5 Author contributions statement

B.D. conceived the experiments, B.D. and S.P. conducted the experiment(s). E.D. and P.G. supervised the project. J.-F.M., A.G., and R.L. provided critical feedbacks. A.G. pre-processed all data. All authors reviewed the manuscript.

Additional information

All code implementation for this project is available at https://github.com/Duplums/SMLvsDL. The data are available in the different web platforms described Table 1. The authors declare no competing interests.

A Data Augmentation

In the small-scale data regime, data augmentations provide a simple way to artificially increase the dataset size, assuming that all augmentations preserve the semantic information inside images (including the label). It is especially relevant for clinical datasets (including patients with schizophrenia, bipolar disorder and autism) since their size rarely exceeds 1-2k. We have tested several standard augmentation strategies, namely affine transformation (with both rotation and translation), crop, flip and adding Gaussian noise. For each strategy, we tested both strong and light augmentations. As noted by Hernandez-Garcia[123], strong augmentations produce more biologically plausible representations compared to light augmentations (maybe because it generates examples that should be explored by DNN for good generalization on test images, exploiting domain knowledge). The hyper-parameters cross-validated are indicated in Table 7. For completeness, we have evaluated these augmentations both on VBM and quasi-raw data on the clinical tasks using DenseNet as backbone.

Augmentations	Affine	Crop	Gaussian Noise	Cutout
Strong	rot(-45deg, 45deg) trans(0, 50vox) zoom(0, 0.2)	$0.5 \times (h,w,d)$	$\mathscr{U}([0, 5\sigma_0])$	$0.5 \times (h,w,d)$
Light	rot(-5deg, 5deg) trans(0, 10vox) zoom(0, 0.1)	$0.75 \times (h,w,d)$	$\mathscr{U}([0, \sigma_0])$	$0.25 \times (h,w,d)$

Table 7. Hyper-parameters cross-validated to evaluate the benefit of data augmentation, viewed as regularization, on final performance.

A.1 Augmentations on VBM data

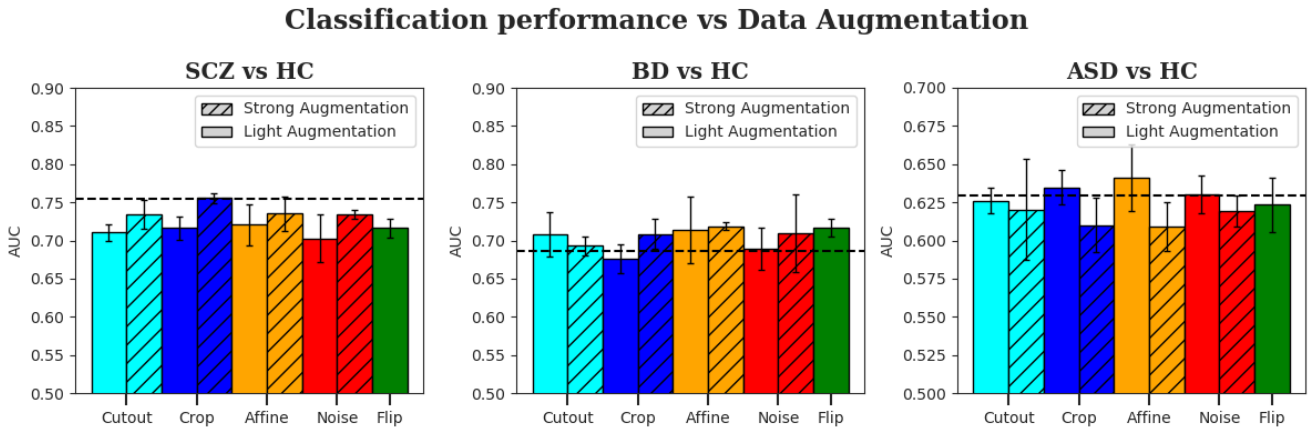

Figure 6. Data augmentation does not bring a significant improvement with clinical datasets. In the rest of this study, we did not perform any particular augmentations when training deep models.

In Fig. 6, we observe overall no improvement when using data augmentation. It can degrade performance using either strong or light augmentation depending on the task, which suggest that current augmentations are highly class-dependent. We hypothesize that standard geometrical transformations (e.g. rotation, translation, flip) are not adapted to our data since all images are non-linearly registered to the same template. As for Gaussian noise, it also appears unnecessary since we already applied a smoothing kernel to regularize our data. We acknowledge that new methods are emerging for generating meaningful synthetic images through non-linear deep generative models[124] and we leave this research axis as future works. In this study, we did not perform data augmentation with VBM to avoid the additional computational cost.

A.2 Augmentations on quasi-raw data

Results from Fig. 6 suggests that current augmentations are not adapted to VBM data. We might hypothesize that this kind of augmentations could be more suited to quasi-raw images, since they are only linearly registered to the MNI template and they could be thus more noisy than VBM. To test this hypothesis, we apply a random combination of all previous augmentations (cutout, crop, affine, Gaussian noise, flip) with probability 50% for each transformation. We report the performances for light and strong augmentations with the same hyper-parameters as in Table 7 and we compare them to baseline results without augmentations on VBM and quasi-raw data.

Task	Test set	Baseline VBM	Baseline Quasi-Raw (QR)	QR+Augmentations	
				Light	Strong
SCZ vs. HC ↑	Internal Test	**85.27**$_{\pm 1.60}$	67.30$_{\pm 1.88}$	73.51$_{\pm 2.73}$	56.13$_{\pm 0.35}$
$N_{train} = 933$	External Test	**75.52**$_{\pm 0.12}$	57.40$_{\pm 4.40}$	55.15$_{\pm 5.01}$	67.30$_{\pm 0.94}$
BD vs. HC ↑	Internal Test	**76.49**$_{\pm 2.16}$	74.62$_{\pm 0.45}$	70.08$_{\pm 1.30}$	55.12$_{\pm 3.00}$
$N_{train} = 832$	External Test	68.57$_{\pm 4.72}$	68.83$_{\pm 0.55}$	63.67$_{\pm 1.58}$	56.08$_{\pm 4.35}$
ASD vs. HC ↑	Internal Test	**65.74**$_{\pm 1.47}$	59.21$_{\pm 2.16}$	57.92$_{\pm 2.14}$	57.06$_{\pm 1.07}$
$N_{train} = 1526$	External Test	**62.93**$_{\pm 2.40}$	61.92$_{\pm 1.62}$	53.39$_{\pm 3.39}$	46.40$_{\pm 4.90}$

Table 8. Data augmentation evaluation on quasi-raw data. We apply a random combination of 5 augmentations (cutout, crop, affine, Gaussian noise, flip) during training on the quasi-raw images with DenseNet backbone. We systematically cross-validate $\gamma \in \{0.2, 0.4, 0.8\}$ value in the *LRStep* scheduler and we set the initial learning rate to $\alpha = 10^{-4}$.

From Table 8, we observe no improvement with the tested augmentations except for SCZ vs HC on the internal test but it always remains far below the baselines on VBM data. Our main conclusions regarding data augmentation on VBM data also hold for quasi-raw images.

B Standard data harmonization improves SML but not DL representations

B.1 Multi-site data harmonization for SML and DL

Task	Model	Internal Test			External Test		
		Linear Adj. Res.	ComBat	No Res.	Linear Adj. Res.	ComBat	No Res.
Age ↓ $N_{train} = 9253$	AlexNet	$2.79_{\pm 0.07}$	$2.98_{\pm 0.06}$	$\mathbf{2.36_{\pm 0.04}}$	$4.59_{\pm 0.08}$	$6.92_{\pm 1.03}$	$\mathbf{3.43_{\pm 0.02}}$
	rbf-SVM	$3.34_{\pm 0.00}$	$3.67_{\pm 0.00}$	$\mathbf{3.21_{\pm 0.00}}$	$4.59_{\pm 0.00}$	$5.74_{\pm 0.00}$	$\mathbf{4.27_{\pm 0.00}}$
	Ridge	$\mathbf{3.08_{\pm 0.00}}$	$3.33_{\pm 0.00}$	$3.56_{\pm 0.00}$	$4.93_{\pm 0.00}$	$4.39_{\pm 0.00}$	$\mathbf{4.21_{\pm 0.00}}$
	ElasticNet	$\mathbf{3.14_{\pm 0.00}}$	$3.21_{\pm 0.02}$	$3.31_{\pm 0.00}$	$4.62_{\pm 0.00}$	$4.38_{\pm 0.03}$	$\mathbf{4.25_{\pm 0.00}}$
Sex ↑ $N_{train} = 9253$	AlexNet	$93.88_{\pm 0.64}$	$95.24_{\pm 0.55}$	$\mathbf{96.13_{\pm 0.42}}$	$94.54_{\pm 0.34}$	$95.58_{\pm 0.65}$	$\mathbf{97.91_{\pm 0.15}}$
	rbf-SVM	$\mathbf{96.09_{\pm 0.00}}$	$95.86_{\pm 0.00}$	$95.16_{\pm 0.00}$	$97.88_{\pm 0.00}$	$98.03_{\pm 0.00}$	$\mathbf{97.28_{\pm 0.00}}$
	Logistic	$95.88_{\pm 0.04}$	$95.63_{\pm 0.03}$	$\mathbf{95.95_{\pm 0.04}}$	$98.26_{\pm 0.00}$	$98.23_{\pm 0.03}$	$\mathbf{98.32_{\pm 0.00}}$
	ElasticNet	$95.09_{\pm 0.05}$	$94.83_{\pm 0.01}$	$\mathbf{95.23_{\pm 0.01}}$	$\mathbf{98.04_{\pm 0.04}}$	$97.95_{\pm 0.65}$	$97.93_{\pm 0.05}$
SCZ vs. HC ↑ $N_{train} = 933$	AlexNet	$71.53_{\pm 0.71}$	$\mathbf{82.35_{\pm 1.45}}$	$79.13_{\pm 0.96}$	$68.50_{\pm 0.90}$	$\mathbf{74.14_{\pm 1.13}}$	$72.07_{\pm 0.95}$
	rbf-SVM	$\mathbf{83.55_{\pm 0.00}}$	$82.06_{\pm 00}$	$82.06_{\pm 0.00}$	$\mathbf{76.39_{\pm 0.00}}$	$72.88_{\pm 0.00}$	$72.88_{\pm 0.95}$
	Logistic	$\mathbf{85.31_{\pm 0.07}}$	$84.25_{\pm 0.02}$	$84.03_{\pm 0.03}$	$\mathbf{76.45_{\pm 0.15}}$	$73.76_{\pm 0.46}$	$73.60_{\pm 0.00}$
	ElasticNet	$\mathbf{88.81_{\pm 1.03}}$	$86.96_{\pm 0.82}$	$85.98_{\pm 1.9}$	$78.98_{\pm 0.98}$	$\mathbf{79.02_{\pm 1.08}}$	$76.42_{\pm 1.68}$
BD vs. HC ↑ $N_{train} = 832$	AlexNet	$62.41_{\pm 3.03}$	$66.77_{\pm 5.44}$	$\mathbf{74.16_{\pm 3.25}}$	$61.67_{\pm 1.26}$	$65.58_{\pm 1.73}$	$\mathbf{72.46_{\pm 2.74}}$
	rbf-SVM	$\mathbf{75.00_{\pm 0.00}}$	$70.92_{\pm 0.00}$	$73.63_{\pm 0.00}$	$\mathbf{67.74_{\pm 0.00}}$	$63.36_{\pm 0.00}$	$63.92_{\pm 0.00}$
	Logistic	$\mathbf{74.07_{\pm 0.09}}$	$73.17_{\pm 0.38}$	$72.96_{\pm 0.25}$	$69.54_{\pm 0.33}$	$69.36_{\pm 0.28}$	$\mathbf{70.12_{\pm 0.26}}$
	ElasticNet	$71.19_{\pm 2.29}$	$72.27_{\pm 1.60}$	$\mathbf{73.85_{\pm 0.28}}$	$\mathbf{70.33_{\pm 2.47}}$	$68.14_{\pm 0.93}$	$70.26_{\pm 1.75}$
ASD vs. HC ↑ $N_{train} = 1526$	AlexNet	$59.06_{\pm 1.96}$	$58.55_{\pm 1.34}$	$\mathbf{62.07_{\pm 1.77}}$	$54.25_{\pm 2.06}$	$60.51_{\pm 1.09}$	$\mathbf{62.46_{\pm 1.21}}$
	rbf-SVM	$66.78_{\pm 0.00}$	$64.64_{\pm 0.00}$	$\mathbf{66.84_{\pm 0.00}}$	$59.10_{\pm 0.00}$	$58.94_{\pm 0.00}$	$\mathbf{60.28_{\pm 0.00}}$
	Logistic	$\mathbf{64.71_{\pm 0.22}}$	$63.11_{\pm 0.09}$	$63.40_{\pm 0.18}$	$\mathbf{63.98_{\pm 0.15}}$	$61.98_{\pm 0.30}$	$61.85_{\pm 0.05}$
	ElasticNet	$\mathbf{63.30_{\pm 4.78}}$	$60.30_{\pm 3.76}$	$60.62_{\pm 2.63}$	$57.98_{\pm 4.71}$	$\mathbf{60.21_{\pm 3.19}}$	$54.96_{\pm 4.94}$

Table 9. Effect of SOTA residualization methods (ComBat or Linear Adjusted on age, sex and diagnosis) on multi-site datasets. Both SML and CNN performance are evaluated on residualized data to assess whether task-related features have been retained after site removal. AlexNet is reported as representative of CNN models. All models are trained 3 times with different random initialization and standard deviation is reported. AUC is reported for binary classification tasks, while MAE is reported for age prediction. Residualization never improves performance for CNN models (see also Fig. 10 in Supplementary for more results with DenseNet121 and ResNet18) with only minor improvement for phenotype prediction with SML. More consistent improvements (between 1% and 3% AUC) appear with less training samples ($N_{train} < 2000$) on diagnosis classification tasks, only with SML.

B.2 More extensive comparisons for DL Models

In order to account for site-related effects on neuroimaging data, we explore the benefit of data harmonization for both SML and DL. Here, we report the results of Linear Adjusted Regression for protecting age, sex and diagnosis while removing site-related variability with deep networks. In Table 10, we observe a constant decrease in performance when performing residualization.

Task	Model	Internal Test		External Test	
		Linear Adj. Res.	No Res.	Linear Adj. Res.	No Res.
Age ↓ $N_{train} = 9253$	AlexNet	$2.79_{\pm 0.07}$	$\mathbf{2.36}_{\pm \mathbf{0.04}}$	$4.59_{\pm 0.00}$	$\mathbf{3.43}_{\pm \mathbf{0.02}}$
	DenseNet	$2.75_{\pm 0.06}$	$\mathbf{2.58}_{\pm \mathbf{0.09}}$	$4.24_{\pm 0.01}$	$\mathbf{3.53}_{\pm \mathbf{0.07}}$
	ResNet18	$2.75_{\pm 0.06}$	$\mathbf{2.49}_{\pm \mathbf{0.08}}$	$3.76_{\pm 0.03}$	$\mathbf{3.49}_{\pm \mathbf{0.08}}$
Sex ↑ $N_{train} = 9253$	AlexNet	$93.88_{\pm 0.64}$	$\mathbf{96.13}_{\pm \mathbf{0.42}}$	$94.54_{\pm 0.34}$	$\mathbf{97.91}_{\pm \mathbf{0.15}}$
	DenseNet	$94.55_{\pm 0.03}$	$\mathbf{96.57}_{\pm \mathbf{0.25}}$	$95.48_{\pm 0.16}$	$\mathbf{98.47}_{\pm \mathbf{0.11}}$
	ResNet18	$95.46_{\pm 0.40}$	$\mathbf{96.33}_{\pm \mathbf{0.34}}$	$96.72_{\pm 0.40}$	$\mathbf{98.39}_{\pm \mathbf{0.26}}$
SCZ vs. HC ↑ $N_{train} = 933$	AlexNet	$71.53_{\pm 0.71}$	$\mathbf{79.13}_{\pm \mathbf{0.96}}$	$68.50_{\pm 0.90}$	$\mathbf{72.07}_{\pm \mathbf{0.95}}$
	DenseNet	$73.09_{\pm 1.32}$	$\mathbf{85.27}_{\pm \mathbf{1.60}}$	$63.34_{\pm 1.10}$	$\mathbf{75.52}_{\pm \mathbf{0.12}}$
	ResNet18	$78.12_{\pm 1.82}$	$\mathbf{80.93}_{\pm \mathbf{3.16}}$	$73.07_{\pm 2.15}$	$\mathbf{74.31}_{\pm \mathbf{0.12}}$
BD vs. HC ↑ $N_{train} = 832$	AlexNet	$62.41_{\pm 3.03}$	$\mathbf{74.16}_{\pm \mathbf{3.25}}$	$61.67_{\pm 1.26}$	$\mathbf{65.49}_{\pm \mathbf{0.91}}$
	DenseNet	$62.91_{\pm 2.20}$	$\mathbf{76.49}_{\pm \mathbf{2.16}}$	$61.70_{\pm 3.50}$	$\mathbf{68.57}_{\pm \mathbf{4.72}}$
	ResNet18	$62.59_{\pm 0.85}$	$\mathbf{68.63}_{\pm \mathbf{3.82}}$	$67.31_{\pm 1.09}$	$\mathbf{69.33}_{\pm \mathbf{0.60}}$
ASD vs. HC ↑ $N_{train} = 1526$	AlexNet	$59.06_{\pm 1.96}$	$\mathbf{62.07}_{\pm \mathbf{1.77}}$	$54.25_{\pm 2.06}$	$\mathbf{62.46}_{\pm \mathbf{1.21}}$
	DenseNet	$61.33_{\pm 3.25}$	$\mathbf{65.74}_{\pm \mathbf{1.47}}$	$54.70_{\pm 2.07}$	$\mathbf{62.93}_{\pm \mathbf{2.40}}$
	ResNet18	$\mathbf{59.02}_{\pm \mathbf{2.37}}$	$58.52_{\pm 3.25}$	$58.64_{\pm 1.66}$	$\mathbf{62.09}_{\pm \mathbf{1.75}}$

Table 10. DL performance on VBM data residualized with linear adjusted residualization (adjusted on age, sex, site and eventually diagnosis). DL performance on VBM data not residualized is indicated for comparison purposes. Linear residualization hurts performance for all models and tasks, indicating that it removes discriminative features used by DL models.

C Raw vs. VBM pre-processing for DL representations

Pre-processing		SCZ vs. HC	BD vs. HC	ASD vs. HC
VBM	Site Pred.(%)	$29.07_{\pm 3.73}$	$26.43_{\pm 2.07}$	$7.01_{\pm 1.53}$
Raw	Site Pred.(%)	$70.71_{\pm 3.36}$ (+41%)	$82.92_{\pm 3.86}$ (+56%)	$48.74_{\pm 5.88}$ (+41%)
	Random Level	10.0	7.69	3.45
	Δ AUC	14%	4%	3%

Table 11. Site prediction balanced accuracy (in %) from latent representation of DenseNet trained on psychiatric disorder classification. We reported the random level when predicting random sites ($= 1/n_{sites}$) as well as the difference ΔAUC between performance on psychiatric classification from VBM and raw data. It clearly shows a much higher over-fitting effect on site (viewed as noise) for raw data compared to VBM even when the model is not trained on this task. This could be a partial explanation for the drop in performance between VBM and raw data.

To explain the difference in performance obtained between VBM and raw measurements, we have represented the latent representation of DenseNet trained on age prediction in Fig. 7. We observe a clear over-fitting effect on site that correlates well with the performance drop observed between internal and external test. As for psychiatric disorders, to quantitatively assess how much site-related information has been captured during training, we have performed a linear evaluation on DenseNet representation. Specifically, we train a linear classifier to predict acquisition site on top of the penultimate layer of DenseNet trained to predict psychiatric condition. Importantly, DenseNet's weights are frozen so its representation is fixed. We have reported the balanced accuracy obtained on site prediction task in Table 11. We observe that site-related features have been very well captured when the network is trained on raw data, even if it has *not* been trained for this task. Differently, this behaviour is much less pronounced for VBM images, which is somewhat expected since the highly non-linear pre-processing acts as a noise-reduction module on raw measurements.

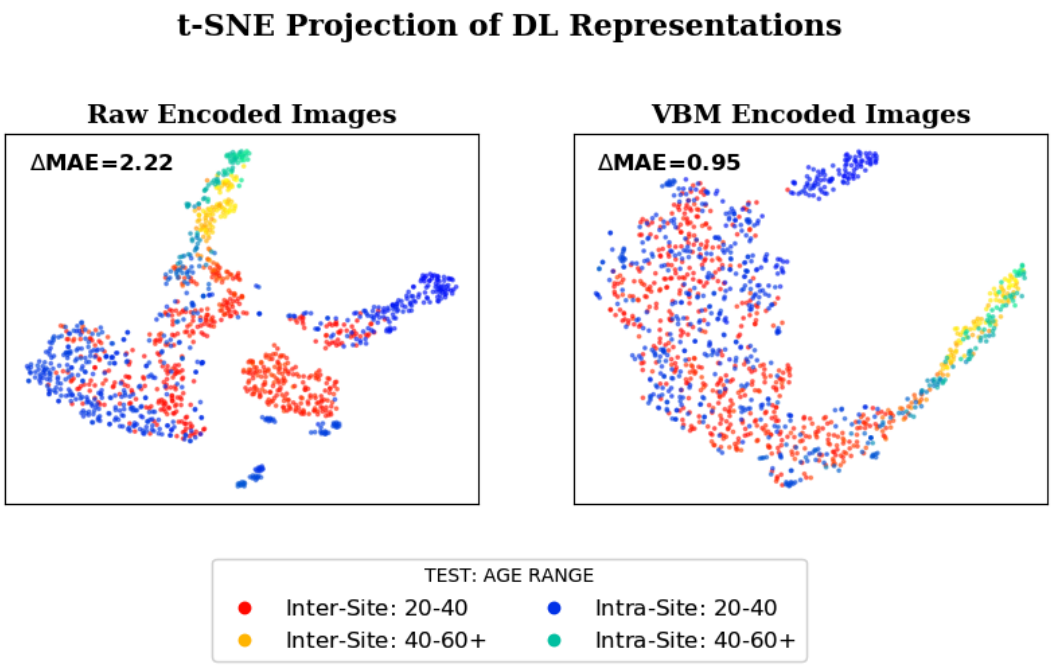

Figure 7. t-SNE visualization of raw vs. VBM images encoded by DenseNet trained on age prediction with $N_{train} = 9253$. We distinguished images from internal test (coming from already-seen sites) and external test. Here $\Delta\text{MAE} = |\text{MAE(external test)} - \text{MAE(internal test)}|$ where $\text{MAE}(x)$ corresponds to the age prediction MAE (Mean Absolute Error) for the test set x. It can thus be seen as a proxy to measure the domain gap between internal and external test sets. Distinct regions for the same age range (blue/red and yellow/cyan) can be observed when encoding raw images. However, these regions clearly overlap for VBM encoded images. It suggests a higher over-fitting effect related to site on raw images than on VBM.

D Grid-search on DL architectures

In this study, we intended to select the most representative convolutional architectures in the neuroimaging field that integrate the most recent advances made in computer vision (e.g. skip-connection[64], dropout[125], batch normalization[126], features re-using[65]). We acknowledge that there is no universal architecture appropriate for all tasks. Nevertheless, we argue that the 3 selected architectures (AlexNet, DenseNet, ResNet) i) provide SOTA results on age prediction, sex prediction, bipolar and ASD detection; ii) are able to outperform SML given enough data (age) or the right initialization point (bipolar disorder and ASD). As such, it suggests that the selected DL architectures have enough expressive power for the range of neuroimaging applications we are tackling. While it not feasible to perform a grid-search on all possible configurations of DL architectures, we have compared the performance of DL models retained in our study with various other CNNs by varying depth and number of convolutional layers. Additionally, since the attention mechanism has shown strong performances on image recognition tasks[127], we included a Transformer architecture in this comparison. We chose a small architecture that is especially suited for image recognition (ViT-Small[128]) and we split our 3D volumes of shape $128 \times 128 \times 128$ into $16 \times 16 \times 16$ patches to which we added a fixed sin-cosine positional embedding[129].

Implementation details for ViT-Small. We followed the current practice[129] for training: we used AdamW as optimizer and we cross-validated the learning rate $\alpha \in \{10^{-3}, 10^{-4}, 10^{-5}\}$ for age regression as well as the weight decay $wd \in \{10^{-2}, 10^{-3}, 10^{-4}\}$ and $\gamma \in \{0.2, 0.4, 0.8\}$ in *LRStep* scheduler for psychiatric disorders classification tasks.

In Table 12, we report the results on age prediction with BHB-10K, the largest dataset available in this study ($N = 10^4$). It demonstrates that i) AlexNet is the best performing network for this task; ii) all 3 selected CNN give among the best results on the external test; and iii) Transformer architecture under-performs compared to CNN. This last observation is somewhat expected as Transformer is known to require a very large amount of (pre-)training data (typically 10 to 300 million images[127]) to give strong results. In our case, we deal with less than 10^4 volumes and a large domain gap between natural and medical images and for which standard data augmentation is not adapted. We further compare CNN to Transformer models in Table 13 for the 3 clinical tasks. This is especially challenging for the Transformer's optimization as we reduce even more the sample

Architecture	Backbone	Internal Test			External Test		
		MAE	RMSE	R^2	MAE	RMSE	R^2
CNN	Conv(3)-FC(5)	$2.53_{\pm 0.02}$	$3.51_{\pm 0.04}$	$93.88_{\pm 0.19}$	$3.68_{\pm 0.03}$	$4.91_{\pm 0.10}$	$90.13_{\pm 0.32}$
	Conv(4)-FC(4)	$2.61_{\pm 0.06}$	$3.65_{\pm 0.08}$	$93.37_{\pm 0.29}$	$3.62_{\pm 0.08}$	$4.92_{\pm 0.07}$	$90.62_{\pm 0.54}$
	Conv(5)-FC(3)	$2.55_{\pm 0.17}$	$3.35_{\pm 0.14}$	$94.77_{\pm 0.18}$	$3.47_{\pm 0.10}$	$4.67_{\pm 0.15}$	$91.35_{\pm 0.42}$
	Conv(6)-FC(2)	$2.46_{\pm 0.04}$	$3.40_{\pm 0.03}$	$94.26_{\pm 0.16}$	$3.48_{\pm 0.05}$	$4.77_{\pm 0.19}$	$91.74_{\pm 0.11}$
Transformer	ViT-Small/16	$2.97_{\pm 0.21}$	$4.17_{\pm 0.27}$	$91.49_{\pm 1.14}$	$3.95_{\pm 0.17}$	$5.30_{\pm 0.19}$	$89.26_{\pm 1.19}$
CNN	AlexNet	$\mathbf{2.36_{\pm 0.04}}$	$3.39_{\pm 0.04}$	$94.42_{\pm 0.01}$	$\mathbf{3.43_{\pm 0.02}}$	$4.80_{\pm 0.08}$	$91.86_{\pm 0.21}$
	DenseNet	$2.58_{\pm 0.09}$	$3.56_{\pm 0.13}$	$93.81_{\pm 0.30}$	$3.53_{\pm 0.07}$	$4.72_{\pm 0.14}$	$91.87_{\pm 0.15}$
	ResNet18	$2.49_{\pm 0.08}$	$3.46_{\pm 0.07}$	$93.93_{\pm 0.28}$	$3.49_{\pm 0.08}$	$4.72_{\pm 0.12}$	$91.64_{\pm 0.49}$

Table 12. Age prediction performance on BHB-10K ($N = 10^4$, VBM data) for CNN and Transformer architectures. For CNN, we report performance as the number of convolutional blocks C and fully-connected (FC) layers F vary, keeping $F + C = 8$ constant. Each convolutional block contains Conv-BatchNorm-ReLU and we keep the number of hidden neurons to 128 in each FC. For Transformer, we select the small Vision Transformer (ViT-Small[129]) with patch size $16 \times 16 \times 16$, using a fixed 3D sin-cosine positional embedding, following the design in [129] that we adapted to 3D images. We compare the performance with the three main CNN backbones used throughout this study: AlexNet, ResNet and DenseNet. AlexNet is best performing on this task but all 3 main families give close results with the other DL architectures, validating the choice of CNN in our study.

Task	Architecture	Internal Test	External Test
SCZ vs. HC	CNN (DenseNet)	$85.27_{\pm 1.60}$	$75.52_{\pm 0.12}$
	Transformer (ViT-Small/16)	$77.85_{\pm 2.31}$	$71.74_{\pm 3.35}$
BD vs. HC	CNN (DenseNet)	$76.49_{\pm 2.16}$	$68.57_{\pm 4.72}$
	Transformer (ViT-Small/16)	$64.92_{\pm 0.78}$	$58.19_{\pm 0.31}$
ASD vs. HC	CNN (DenseNet)	$65.74_{\pm 1.47}$	$62.93_{\pm 2.40}$
	Transformer (ViT-Small/16)	$57.26_{\pm 1.17}$	$55.01_{\pm 3.89}$

Table 13. Psychiatric disorders classification performance (% AUC) for CNN and Transformer architectures on VBM data. We select a small architecture for Transformer to avoid strong over-fitting issues during optimization on these relatively small-scale datasets.

size (by $\times 10$), requiring more cross-validation of hyper-parameters (see "Implementation details for ViT-Small" above). We consistently observe lower performances for ViT-Small compare to CNN on all tasks, confirming previous results on age prediction.. Overall, these results validate our choice of DL architectures retained in our study.

E Dimensionality reduction hurts performance for SML models

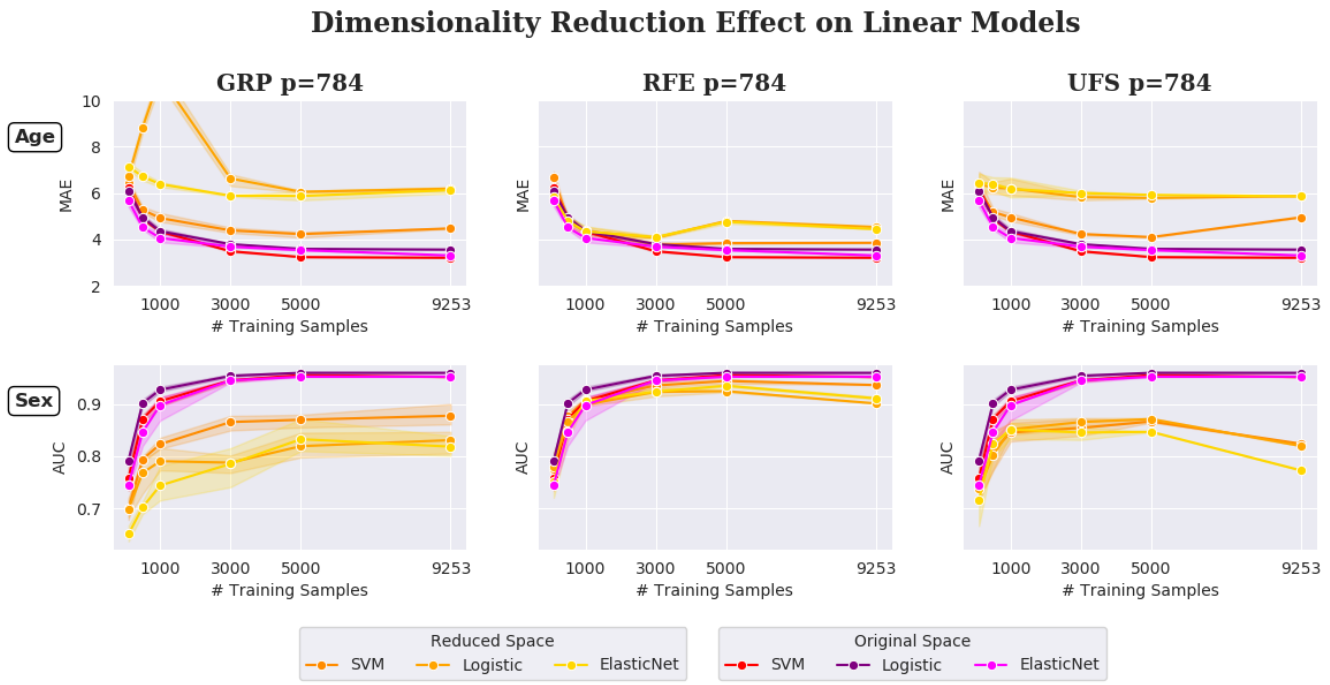

Figure 8. Three dimensionality reduction methods are evaluated on 3D anatomical VBM images from BHB-10K, namely Gaussian Random Projection (GRP), Recursive Feature Elimination (RFE) and Univariate Feature Selection (UFS). We reproduce the same experimental setting as in previous studies[13,14] by setting the number of reduced dimensions to $p = 784$ (from 300K gray matter voxels). Performance of SML (including rbf-SVM and penalized linear models) on reduced data are reported on both phenotype prediction ($N_{train} \in \{100, 500, 1000, 3000, 5000, 9253\}$) and diagnosis classification tasks (see Fig. 10). We use the same training/validation/testing splits as previously. The performance in the original space is also reported for comparison purpose. All models are tested on both the internal test (shown here) and the external one (see Fig. 11), with similar conclusions. In all cases, dimensionality reduction provides no improvement for SML and it significantly decreases performance when using UFS or GRP.

Previous studies[13,14] argued that dimensionality reduction was a necessary step for SML models to limit over-fitting and work properly (especially considering the very high-dimensionality of 3D MRI, $n_{voxels} > 300$K). We carefully reproduce the experimental design from these studies (same feature space dimension $p = 784$ and reduction methods), and we observe that i) SML performs very well in the original space (as DL do) and ii) dimensionality reduction hurts SML performance by removing discriminative features.

Specifically, we use three different feature selection methods, Gaussian Random Projection (GRP), Random Feature Elimination (RFE), and Univariate Feature Selection (UFS) similarly to Abrol et al.[13] and Schulz et al.[14]. As opposed to RFE and UFS, GRP is an unsupervised feature selection method that applies a random matrix to the data and preserves the euclidean distance between points, up to an error ε depending on the number of features selected. In Fig. 8, we plot the performance of SML in the reduced space with $p = 784$ features (similarly to Abrol et al.[13] and Schulz et al.[14]) as compared to the original space. We also performed the same experiments on the three clinical datasets, and we reported the results Fig. 10. We observe a strong degradation in performance for all models tested, especially with GRP (drop by 12% AUC for sex prediction, +2.7 MAE for age regression, >10% AUC for all binary classification tasks on clinical datasets with SML models, and the maximum number of training samples). This is somewhat expected since GRP is fully unsupervised, *i.e.* it does not rely on the target variable to preserve relevant features (and thus can focus on general non-biological variability, e.g., acquisition site). RFE seems to be the best performing method while still hurting the performance compared to preserving the original data (−5% AUC for sex classification and +0.97 MAE for age prediction with $N_{train} = 9253$). This suggests that the non-redundancy and sparsity hypothesis in the final solution has been violated on these tasks[60]. Similar results have also been obtained using the external test set (see Fig. 11). In these experiments, the number of selected components p is a critical hyper-parameter, and it was not discussed in previous studies[13,14]. For completeness, we also perform additional experiments with $p = 10$k (see Fig. 12), showing that we can reach similar performances than in the original space on age and sex prediction by reducing the

input size by 30 (the gray matter mask containing about 300k voxels), with RFE (ΔMAE$=0.03$ and ΔAUC$=0.32\%$ with SML models for age and sex prediction).

Overall, these experiments show that dimensionality reduction is not necessary for SML in our experiments and it can decrease performance without careful model selection. Regularized SML models can also learn from very high-dimensional data.

E.1 Replication on UKBioBank

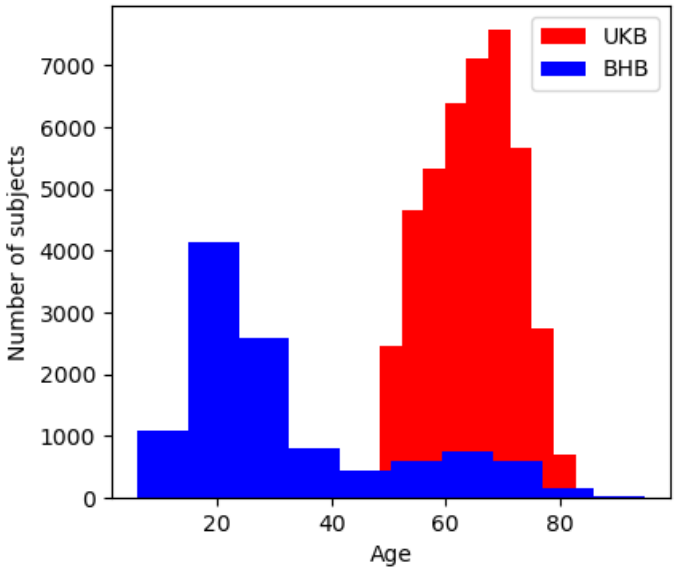

Figure 9. Age histogram for BHB-10K ($n = 11210$) and UKBioBank dataset ($n = 42923$).

We have previously shown that dimensionality reduction strongly degrades performance for SML on BHB-10K. This results might seem surprising in light of previous studies on this topic[13,14] that obtained better performance, especially with GRP technique. We attribute this discrepancy to two main factors related to BHB-10K itself:

- BHB-10K is diverse in terms of acquisition protocols and MRI machines (>70 acquisition sites), which is not the case for the other benchmarking data used in previous studies (UKBioBank for age and sex prediction);
- BHB-10K includes a much younger population than UKBioBank (age=32 ± 18 for BHB-10K vs 64 ± 8 for UKBioBank, see Fig. 9) however the biological brain ageing process differs between adolescents, young adults and the elderly population[130].

To test this hypothesis, we have reproduced the same experimental design as prior works for SML models by using the same dataset (UKBioBank) and we compare the results with what we found on BHB-10K. Importantly, we used our own code implementation and we applied the same pre-processing as for BHB-10K (VBM pre-processing). We also used the same number of training samples as in BHB-10K ($n_{train} = 9253$) and we have performed a 20-fold CV repetitions as in[13] using $n_{test} = 1000$ test samples. We report the results with GRP as dimensionality reduction (extracting p features) and the original VBM data in Table 14, from which we can draw several observations:

1. **We successfully reproduce the results on UKB with our code and data pre-processing:** using GRP reduction with $p = 8k$ features, we retrieve the same performance as in[13] for the linear models (which was among the best results for SML). We acknowledge that we required more features than prior works to achieve similar results but we attribute this discrepancy to tiny difference in VBM pre-processing pipeline and a smaller number of training samples to be comparable with BHB-10K dataset.
2. **We improve previous results without dimensionality reduction:** by further removing the dimensionality reduction step, we achieve significantly better performance ($p_{age} = 2.15 \times 10^{-8}$ and $p_{sex} = 3.59 \times 10^{-6}$, two-tailed paired sampled t-test) than previous works[13,14] on the same data for both age and sex prediction, using less training samples.

Dataset	Input features	# features	Task	
			Age (MAE)	Sex (Acc %)
UKB	Feature Selection (ours)	$p = 784$	4.05 ± 0.08	89.54 ± 0.89
	Feature Selection (ours)	$p = 4$k	3.51 ± 0.07	95.34 ± 0.74
	Feature Selection (ours)	$p = 8$k	3.35 ± 0.07	96.56 ± 0.51
	Whole-brain (ours)	$p = 331$k	$\mathbf{3.12 \pm 0.08}$	$\mathbf{98.10 \pm 0.49}$
	Feature Selection (from [13])	$p = 784$	3.36 ± 0.08	96.68 ± 0.69
	Whole-brain (from [22])	(Unknown)	3.3 ± 0.7	88.4 ± 1.0
BHB-10K	GRP (ours)	$p = 784$	6.19 ± 0.05	75.94 ± 0.02
	GRP (ours)	$p = 10$k	4.64 ± 0.09	86.15 ± 1.40
	Whole-brain (ours)	$p = 331$k	3.31 ± 0.001	90.68 ± 0.18

Table 14. We reproduce the same experiment as[13] on UKB using our implementation of SML and VBM pre-processing. We set $p \in \{784, 4000, 8000\}$ features for GRP and we use $n_{train} = 9253$ samples as training size to be comparable with BHB-10K. We report the results on $n_{test} = 1000$ samples with a 20-fold Monte-Carlo CV. We indicate the results obtained by[13] with $n_{train} = 10000$ samples as reference with Logistic Regression for sex prediction and ElasticNet for age prediction. We also report the results from[22] using linear models for these two tasks and comparable sample size as well as the results on BHB-10K (internal test) from our manuscript.

3. **We achieve better performance on UKB than on BHB-10K:** we observe a performance gap between models trained on BHB-10K and UKB dataset for the same tasks. This is true both for GRP or VBM features. Nonetheless, as for BHB-10K, the performance is better without dimensionality reduction.

These observations give further credit to our original findings on BHB-10K regarding dimensionality reduction with SML: it hurts performance for SML and simple linear models produce better results with higher-dimensional input. As we mentioned in the manuscript, BHB-10K is more diverse and challenging dataset than UKB as it is highly multi-centric, but it is also more representative of the challenges emerging from clinical neuroimaging data.

E.2 Complementary experiments on dimensionality reduction for SML

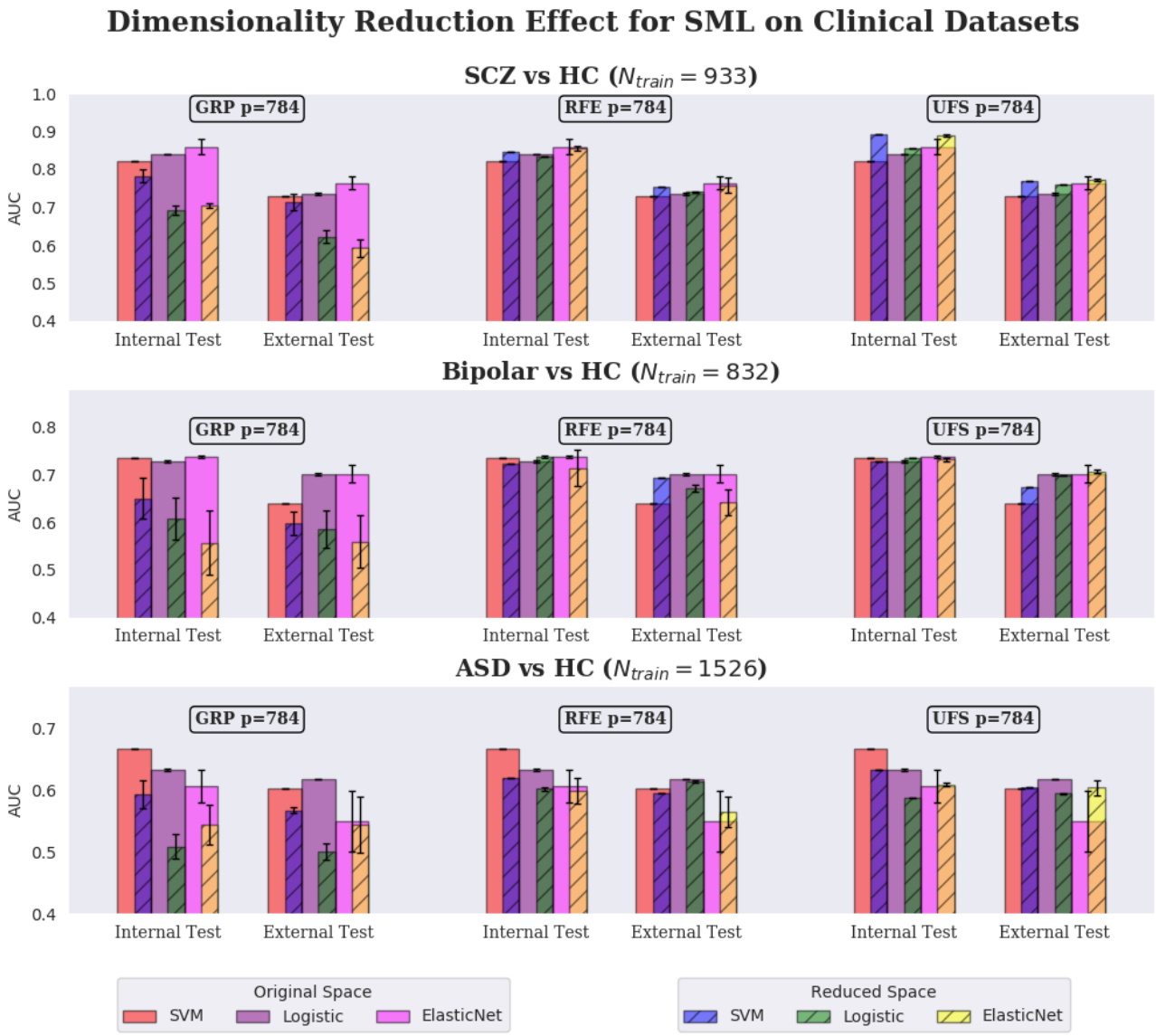

Figure 10. SML performance after dimensionality reduction (GRP, RFE, UFS) on all 3 clinical datasets (schizophrenia detection, bipolar and ASD classification). The performance in the original space is also reported for comparison purposes. Overall, SML methods perform well in the original space without the need of additional feature extraction.

F Replication of SML results on UKBioBank for age regression

In order to replicate our SML analysis pipeline on another dataset and to directly compare our results with the literature, we conduct the same experiment as in[22] with linear models on UKBioBank data pre-processed in the same way we did for BHB-10K with VBM. In details, we perform a 20-fold Monte-Carlo CV sub-sampling on UKB with $N_{train} \in \{100, 9253\}$ training samples and $N_{test} = 1000$ test samples. Hyper-parameters for ElasticNet are cross-validated with a nested 5-fold CV on the same range as detailed in Section 2.5. We report the results in Table 15. We obtain far better results than prior work for SML both in the small-scale and large-scale regime, suggesting a poor estimation of hyper-parameters in Peng et al. These results on UKB are in line with our original results on BHB-10K (internal test).

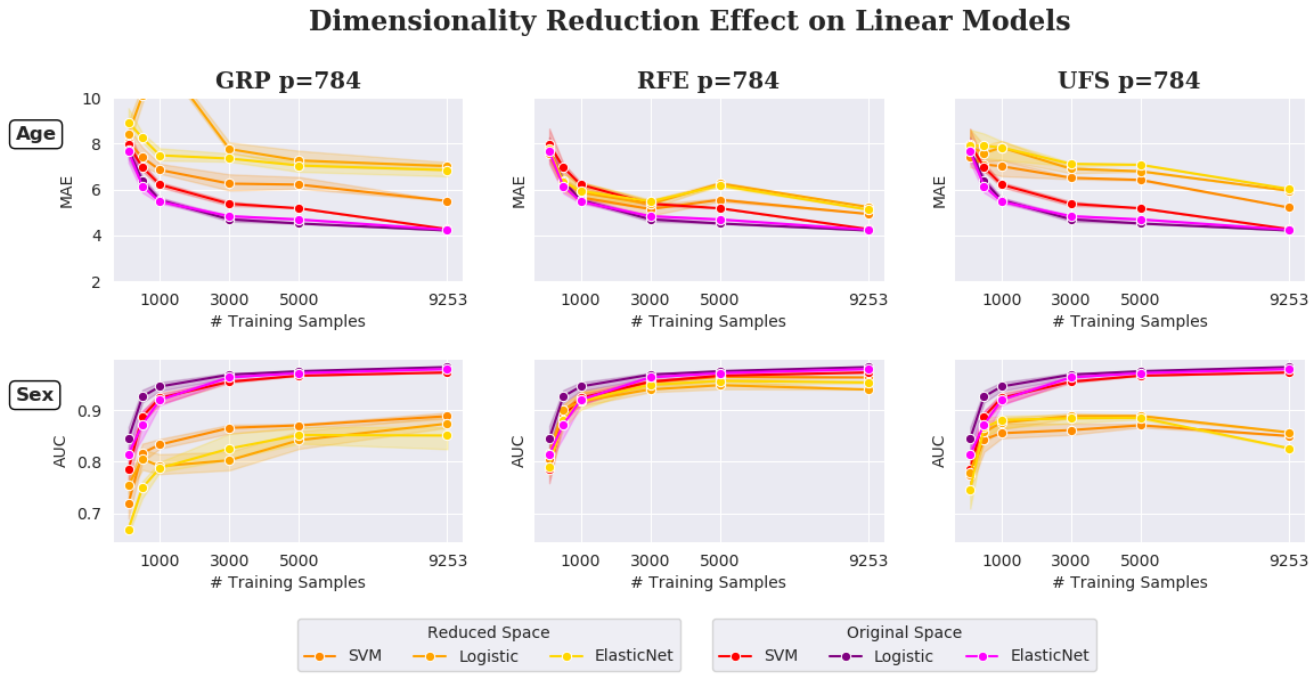

Figure 11. SML model performance after dimensionality reduction on the external test set. The same performance trends can be observed both on internal and external test: feature reduction hurts the performance and is not necessary for SML.

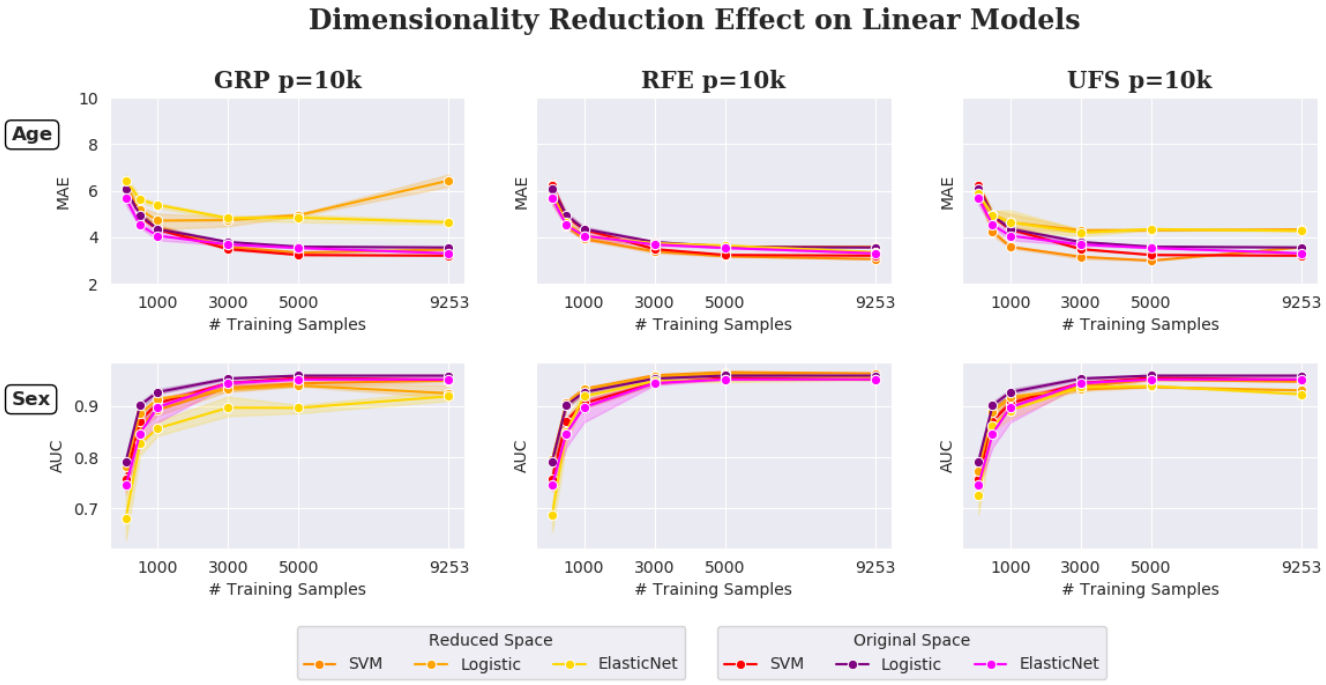

Figure 12. SML model performance after dimensionality reduction on the internal test set when the final reduced dimension space is $p = 10^4$. There is no particular gain in performing feature extraction, no matter the sample size.

Dataset	Model	Age regression (MAE)	
		$n_{train} = 100$	$n_{train} = 9253$
UKB	Linear (ours)	4.53 ± 0.21	3.12 ± 0.08
	Linear (from [22])	5.4 ± 0.8	3.3 ± 0.7
	SFCN (from [22])	4.6 ± 0.8	2.2 ± 0.05
BHB-10K (internal)	Linear (ours)	5.64 ± 0.43	3.31 ± 0.001
	AlexNet (ours)	5.12 ± 0.58	2.36 ± 0.04
BHB-10K (external)	Linear (ours)	7.64 ± 0.66	4.25 ± 0.001
	AlexNet (ours)	8.00 ± 1.26	3.43 ± 0.02

Table 15. We reproduce the same experiment as Peng et al.[22] on UKB using our implementation of SML (ElasticNet here). We study small-scale data regime with $n_{train} = 100$ samples and large-scale data regime $n_{train} = 9253$ to be comparable with BHB-10K. We perform a 20-fold Monte-Carlo CV with $n_{test} = 1000$ and we cross-validate the hyper-parameters with a 5-fold nested CV. We indicate the results obtained by Peng et al. as reference on UKB and we also report the results on BHB-10K from our manuscript.