Non-linearity matters: a deep learning solution to generalization of hidden brain patterns across population cohorts

Mariam Zabihi[1,2], Seyed Mostafa Kia[1,2], Thomas Wolfers[1,2,3], Richard Dinga[1,2], Alberto Llera Arenas[1], Danilo Bzdok[4,5], Christian F. Beckmann[1,2,6], Andre Marquand[1,2,7]

[1]Donders Institute for Brain, Cognition and Behavior, Radboud University Nijmegen, Nijmegen, the Netherlands
[2]Department for Cognitive Neuroscience, Radboud University Medical Center Nijmegen, Nijmegen, the Netherlands
[3]NORMENT, KG Jebsen Centre for Psychosis Research, Division of Mental Health and Addiction, Oslo University Hospital & Institute of Clinical Medicine, University of Oslo, Oslo, Norway
[4]Multimodal Imaging and Connectome Analysis Lab, McConnell Brain Imaging Centre, Montreal Neurological Institute and Hospital, McGill University, Montreal, Quebec, Canada
[5]Mila - Quebec Artificial Intelligence Institute, Montreal, Quebec, Canada
[6]Centre for Functional MRI of the Brain, University of Oxford, Oxford, United Kingdom
[7]Department of Neuroimaging, Institute of Psychiatry, Psychology, & Neuroscience, King's College London, London, United Kingdom

Abstract

The increasing number of neuroimaging scans in recent years has facilitated the use of complex nonlinear approaches to analyzing such data. More specifically, deep learning, which has been previously hindered by the curse of dimensionality is now feasible. However, it remains challenging to use these techniques develop reliable biomarkers and find an optimal representation of data that explains the biological underpinnings of the mental disorders Here, we employed a 3-dimensional autoencoder with an architecture designed from the ground up for task-fMRI data. Our study presented a coherent strategy for optimizing model parameters and architecture and a method for visualizing, exploring and interpreting the latent space representation. We trained our model with multi-task fMRI data derived from the Human Connectome Project (HCP) that provides whole-brain coverage across a range of cognitive tasks. Next, in a transfer learning setting, we tested the generalization of our latent space on UK Biobank data as an independent dataset. We showed that the model did not only learn salient features such as age but also high-level behavioral characteristics and that this representation was highly generic and generalizable to an independent dataset. Furthermore, we demonstrated that the projection of latent space back into the original space is meaningful and interpretable. Finally, our results show that with careful implementation, nonlinear features can provide complementary information that accessible to purely linear methods. Our results provide an important step toward learning interpretable and generalizable latent representations that link cognition with underlying brain systems.

Introduction

The application of machine learning methods and, more specifically, deep-learning methods have shown remarkable returns on many medical imaging problems. [1]. In particular, the applications of deep learning to neuroimaging data are rapidly increasing, mostly the use of supervised learning approaches to solve, for example, classification problems [2–9]. However, linear approaches often [10–12], but not always [13], produce equal or better performance compared to complex deep neural network models. This is not because the mapping between external covariates and brain correlates is necessarily linear but may instead be due to the lack of data or the choice of input features (e.g., regional averages). While in neuroimaging studies, dealing with a limited number of high-dimensional data had hindered employing such approaches for a time due to the curse of dimensionality [14], the recent increase in the availability of neuroimaging data has provided a great opportunity to move toward employing complex nonlinear methods [12,13,15–19].

One ultimate challenge in the application of machine learning to neuroimaging is to find reliable biomarkers that explain the biological underpinnings of healthy and disordered mental states. Model interpretability is therefore an important consideration when determining which model to use.[1,20–23]. Many deep learning studies in neuroimaging use hand-crafted features e.g., regions of interest (ROIs) or image-derived phenotypes (IDPs), which are potentially suboptimal for prediction because (i) hand-crafted features may not accurately capture complex structural or functional brain characteristics e.g. overlapping latent representations encoded in the brain, nor their intricate relationships with behavior and (ii) they do not benefit from the strength of deep neural networks in automatically learning the optimal representation from the data (for example using convolutional filters). Particularly in task fMRI studies, which are designed to study mappings from brain activations to cognition and behavior, there are many challenges in understanding the underlying mechanisms, including the extensive heterogeneity across subjects, finding an optimal representation, and a reliable reference to compare the activations [24–30]. Consequently, using hand-crafted features potentially leads to losing crucial information relevant, for example, for understanding inter-subject variability [23,31]. In these scenarios, learning an optimal representation of high-dimensional neuroimaging data rather than – for example – using pre-defined ROIs may enable us to better understand individual variation and more accurately predict clinical and cognitive variables. This representation, also called a latent representation, allows us

to reduce the data dimensionality and extract only the essential features from the data. In other words, a latent representation model maps complex and high-dimensional data into a reduced and low-dimensional space[32]. There are two steps to assess the latent representations here: first, whether the derived latent representation shows a stronger association with clinically relevant covariates compared to data in the original space (e.g., mapping from brain to behavioral scores) and further, whether the latent space can be generalized to new data (new brain scans/ new participants) which may have a partially different distribution. In the event that this is proven applicable, then, the knowledge learned from one (big) dataset can be transferred to modeling smaller datasets in a transfer learning paradigm [33].

Most applications of deep learning in neuroscience focus on learning a latent representation that is optimized for a single supervised learning problem, such as predicting age/ sex (e.g. [13] [10–12]). However, this may reduce the generalizability of the learned latent representation to other problems. Therefore, we sought to learn a general-purpose latent space that is not bound to a particular task, and instead aims to learn features from the data that are predictive of many different cognitive scores. There have been a number of efforts to that end, e.g. to generate synthetic neuroimaging data [34–37]. However, most of these studies evaluate the data representation on the basis of specific measures like reconstruction error. However, this does not necessarily suggest that the latent space presents relevant features, and what is more important is how accurately such representations can predict behavioral or clinical variables. Although linear data-driven transformations like Principal Component Analysis (PCA) and Independent Component Analysis (ICA) [38–42] are widely used for feature representation and dimensionality reduction in neuroimaging, these methods often fail to extract complex nonlinear relationships in data. [43,44]. In addition, we show how we can control the latent representation learned by the autoencoder by adding a supervised learning term to the reconstruction.

In this paper, we propose to explore the value of learning a general purpose nonlinear latent space representation of task-fMRI images using a 3-dimensional autoencoder. Autoencoder neural networks provide a powerful tool in various applications in neuroimaging studies, from image segmentation to abnormality detection and latent representation [4,15,16,45–48]. Complementary to these approaches, we are interested in automatically learning contextual features using an autoencoder. Briefly, an autoencoder is a deep neural network architecture that consists of two parts an encoder

and a decoder. The encoder projects the inputs to a lower-dimensional latent space using a non-linear transformation. The decoder translates back the latent space to the original space by reconstructing the inputs[49]. In contrast to many previous approaches, this does not require the prior specification of nodes or regions of interest, can learn overlapping representations, can use the full range of spatial patterns in the fMRI signal and takes advantage of the strengths of deep learning, for example by learning convolutional filters that capture low-level features of the images.

More specifically, in a fully data-driven approach, we showed that there is useful information about the data in the nonlinear latent space that is not fully captured by a linear data representation and that such information can be extracted using a hierarchical non-linear autoencoder architecture. Here, we employed an autoencoder with an architecture designed from the ground up for task-fMRI data and provide a method for visualizing, exploring and interpreting the learned representation. We trained our model with multi-task fMRI data derived from the Human Connectome Project (HCP)[24] that provides whole-brain coverage across a range of cognitive tasks. Next, in a transfer learning setting, we tested the generalization of our latent space on a UK Biobank dataset[28] after fine-tuning. Our experimental results show that our nonlinear data representation provides a strong foundation for subsequent analysis of brain-behaviour mappings and results in more accurate brain-behavior mappings than a commonly used linear approach.

Methods

An overview of our approach is shown in Figure 1. Here, we used two different data sets. The first data set is task-based fMRI data from HCP [24] S500 release. The second tfMRI data is from the 2020 UK Biobank imaging release[50].

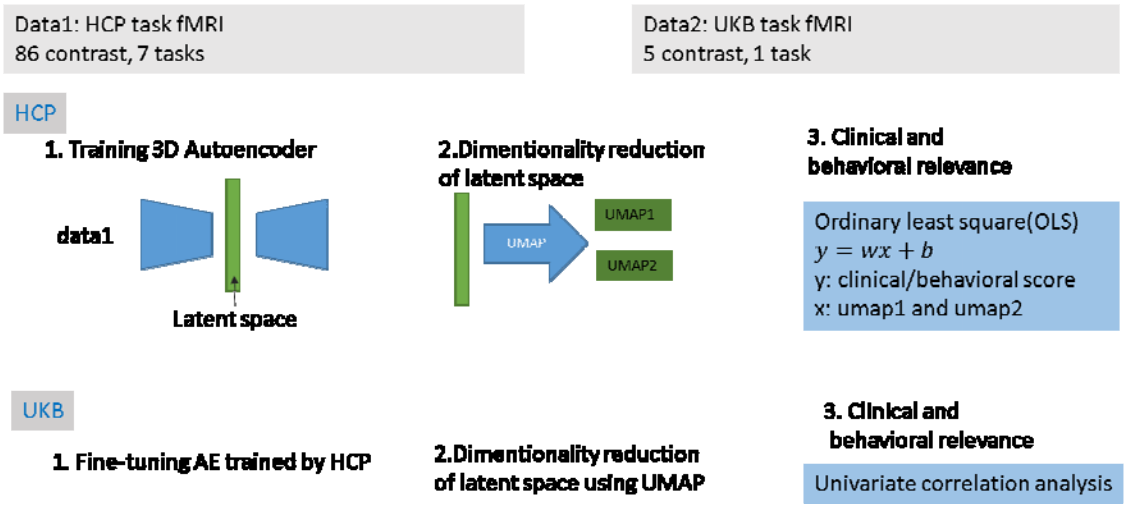


Figure 1:Method's overview

HCP data

Imaging data: We used tfMRI contrast data from 468 participants in total (187 males and 281 females, Age= 29.2±3.5) from seven different tasks (emotion processing, gambling, language, relational processing, social cognition, motor, working memory) across 86 contrasts which served as the basis in previous brain-imaging work [51,52]. This yields a total of N≈40K task-fMRI contrasts. The HCP dataset is well suited for this purpose because the task battery covers a wide range of cognitive domains and the neuronal activations associated with the task provide good coverage of the entire brain [25]. The number of participants may vary from task to task; not all the participants have data in all the tasks. While HCP has a large number of samples, the number of participants is relatively small. Therefore, we split data into 5 subsets in a 5-fold cross-validation scheme. The splits are made at the subject level so that each fold contains all the contrasts for a specified set of subjects in order to prevent overly optimistic estimates of generalizability due to the correlations between different contrasts from the same subject. More specifically, in each fold, about 95 participants (20% of the data) were reserved for the test set (N=8K brain scans) and the rest for the training (N= 32K brain scans, 373 participants). For each fold, we trained a separate model.

Non-imaging data: the HCP dataset contains various sets of clinical, behavioral, and cognitive tasks, which we use to assess the quality of the latent representation (i.e. how well it predicts behavior). We grouped the scores into 19 categories e.g. sleep quality contains scores relevant to Pittsburgh Sleep Quality Index (PSQI) scores [53] (see supplementary information for the full list of categories). We only included the measures

that their scores are available more than half of participants. Moreover, in line with previous studies [54,55] the measures that had same value for more than 80% of the participants were excluded from further analysis.

UK Biobank

Imaging data: we used UK Biobank task-fMRI contrast data from 20781 participants and 5 contrasts, in total N≈104K scans (9,860 males, 10,921 females, Age=54.6 ±7.4). The fMRI data derived from UK Biobank uses the same paradigm as the emotion task from the HCP with only minor modifications (e.g. to accommodate shorter run length) [28,50]. Since UK Biobank provides a larger number of participants than HCP, we trained separate model for each contrast. We randomly selected N=15585 of participants for the train set and 5196 for the test set. All the contrast-models employ the same dataset configuration (the test and train sets).

Non-imaging data: The UK Biobank study provides an extensive number of clinical, behavioral, lifestyle and cognitive scores, which categorized to eight groups e.g., cognitive phenotypes, lifestyle, and mental health (see supplementary information). The exclusion criteria are the same as HCP measures.

Preprocessing image data and Model architecture selection:

For both datasets we used the volumetrically preprocessed images in standard reference space provided by the respective consortia [56,57] (for HCP using the 'minimally processed' pipeline [56]). Subsequently the scans images were downsampled from 2mm to 3mm voxel resolution to reduce the computational burden then cropped tightly to the whole brain such that the dimension of the image decreased to 56×64×56. The model was trained on the whole-brain contrast images.

We first conducted a pilot study to determine the optimal autoencoder architecture across a wide range of architectures, optimizers, feature scaling approaches with various numbers of hyper-parameters. To do so, we trained the network using half of the HCP training data and tested the performance of the model on reserved test data containing 30 participants (N≈2580 scans). Note that to ensure accurate estimates of generalizability, the reserved test data was used exclusively during pilot study (i.e. we did not re-use these data in subsequent analyses). The details of the models evaluated can be found in the supplementary Table 3. To preserve the details in the image, we choose 3×3×3 kernels for the convolutional layers. More specifically, here, we tested different image normalization (sample-wise and feature-wise of Min-Max and

standardization), the number of latent variates (nodes) in the fully connected layer, the number of filters in each layer in addition to different model's optimizer. The reconstruction error of the model reported using mean squared error (MSE) as performance metric of interest.

Training the model

The final instance of our autoencoder with the optimal architecture from the pilot study was trained using HCP training data with 200 epochs using early-stopping approach to prevent model from overfitting. Next, we used the trained model and transferred the learned parameters from HCP to UK Biobank. We fine-tuned the model with the same hyper-parameters using UK Biobank data. We tested if the convolutional kernels are learning relevant features through the fine-tuning and demonstrated several random kernels' weights in the supplementary information. This is important to ensure that the autoencoder avoids learning the trivial solution (i.e. a semi-identity function via down-sampling in the encoding phase and up-sampling in the decoding phase). To compare nonlinear transformation to linear, we trained a PCA model with the same number of components as the number of nodes in the autoencoder's fully-connected layer using HCP data. PCA is a natural choice of baseline reference model because it is equivalent (up to a rotation) to a single fully-connected layer autoencoder with a linear activation function [49]. We applied the same PCA model to UK Biobank to compare the results after transfer learning.

Latent space representation using UMAP

To further evaluate our model, we visualized the latent space using a Uniform Manifold Approximation and Projection (UMAP) approach [58] with two components. UMAP is a manifold learning technique similar to t-distributed stochastic neighbor embedding (t-SNE) [59] that preserves the local structure of high dimensional data in a nonlinear space. UMAP is superior to tSNE since it better preserves the global structure of data (in addition to its local structure). Furthermore, it is more stable under perturbation or resampling of the data. Moreover, UMAP is relatively fast which is beneficial when tuning the hyperparameters.

Here, we used UMAP to visualize the learned latent space without any preprocessing steps [58] (see supplementary information for more details).

To assist the interpretation of the latent space, we use a simple method to project back the latent spaces in input (i.e., brain) space. To achieve this, we take advantage of the

fact that the UMAP algorithm finds clearly separated clusters for the different fMRI contrasts (see results below). Then, for each contrast, we calculated the center of its cluster (i.e., the centroid of K-means clustering) in 2- dimensional UMAP space. We transformed these centroid points to the latent space (using the inverse UMAP transformation) and used the decoder component of the autoencoder to reconstruct the images corresponding to these cluster centers.

Behavioral Associations

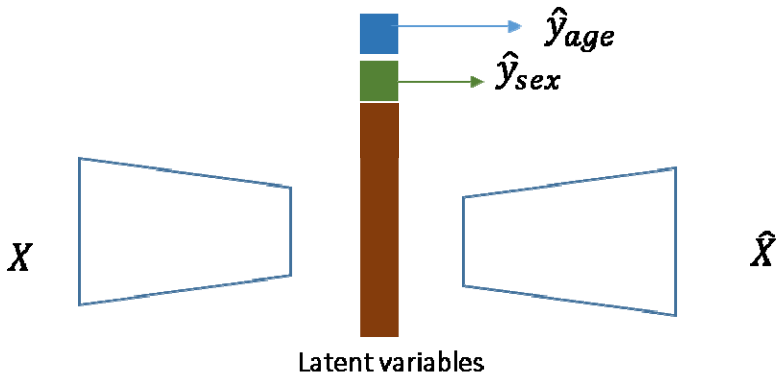
Furthermore, to assess the biological validity of our latent space, we calculated the linear association between clinical and behavioral measures and the reduced latent space for HCP and UK Biobank data.

We used two different approaches because of the different resampling schemes employed in each dataset. For HCP (trained under cross-validation), we regressed non-imaging scores onto the first two UMAP-components using ordinary least squares (OLS) regression in each contrast individually. We used OLS because: (i) the number of test participants in each fold was relatively small in HCP, and (ii) each split resulted in a different latent space, which would make a straight correlation across splits uninterpretable. In the end, we reported the average association (adjusted-r) across separate regression models estimated for each of the five cross-validation splits. In contrast, a single training/test split was used for UK Biobank, so we calculated a straight Pearson correlation between each UMAP component and non-imaging score individually. We repeated the same procedure for behavioral association analysis on the first two PCA components separately.

Semi-Supervised Autoencoder

An important feature of our approach is the ability to control the latent representation by incorporating supervised components into the autoencoder. We illustrate by adding age and sex as supervised components. There are several ways to do this: a straightforward approach would be to add read-out heads for these variables to the existing latent representation, but in this work, we augment the latent representation with additional latent variables that we use to capture age- and sex-related variation. This is similar to the architecture proposed by Pinaya et.al[15]. (Figure 2), although in contrast with this prior model[15] , we used age as a continuous variable rather than a one-hot encoded matrix (i.e. which would effectively treat the regression as a classification problem). This enables us to generalize beyond the age range used in the

training dataset, which is important for transfer learning because of potential differences between cohorts.



$$Loss = Image\ reconstruction\ loss + age\ prediction\ loss + sex\ prection\ loss + Cross - Covariance\ loss$$

Figure 2: Semi- unsupervised Autoencoder.Cross-covariance cost is used to separate out the confounding variables (i.e., age and sex) from the other latent features of the data during training[15].

Results

Model selection: autoencoder architecture and hyper-parameter tuning

The optimal model architecture derived from the pilot study is shown in Figure 3. In more detail, the pilot study (Table 3 in supplementary information) suggested that the relatively larger number of filters in the first layers and last layers of the autoencoder results in lower reconstruction error. Moreover, among different image normalization strategies, robust feature-wise normalization was more effective. The number of nodes in the fully-connected layer was another important hyper-parameter, and our results show that the reconstruction performance improves significantly from 10 nodes to 100 nodes while from 50 to 100 nodes, the improvement is small ((Table 3 in supplementary information)).

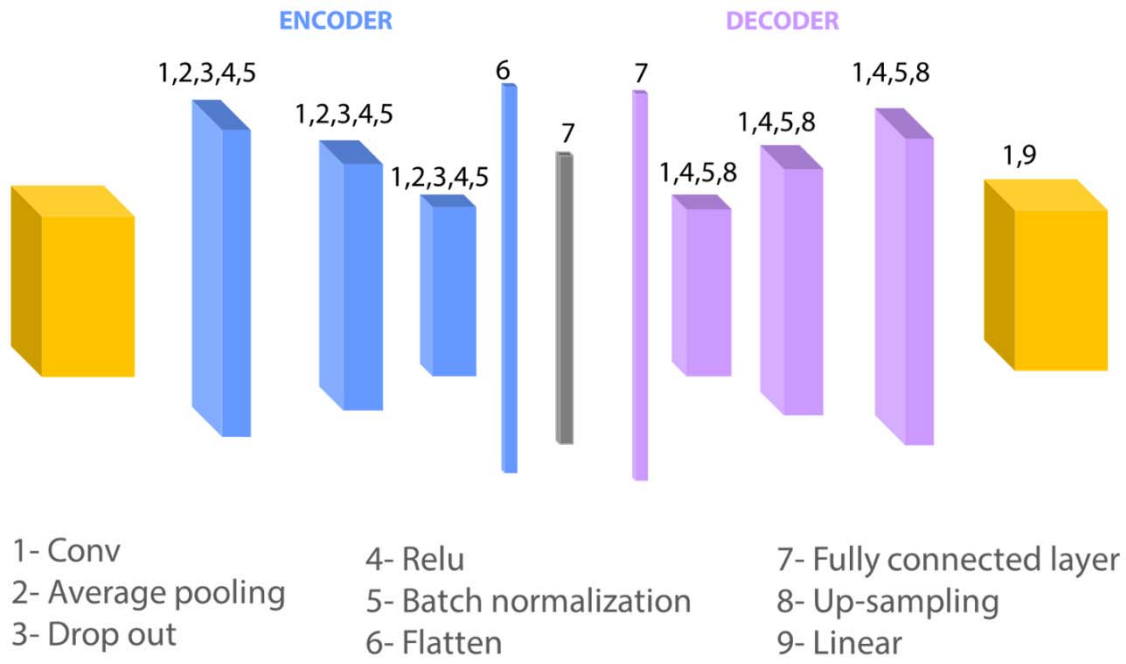


Figure 3: Autoencoder architecture

Autoencoder outperforms PCA in separating tasks and subtasks

The out-of-sample MSE of PCA, shown in Table 1, is comparable with the autoencoder (but for practical purposes equivalent) performance, however as noted above, the association with behavior is of greater importance for the purposes of assessing the quality of the representation.

Table 1: Mean square for PCA model and Autoencoder on HCP data

	MSE	**EXPLAINED VARIANCE**
PCA(100)	0.706 ±0.009	0.331 ± 0.002
AE(100)	0.741 ±0.012	0.301 ± 0.007

The scatterplot of the first two components of PCA and two UMAP components of the autoencoder's latent variables for selected contrasts [25] is shown in Figure 4 (See supplementary information for more tasks). As noted above, the UMAP projections show a good separation between tasks.

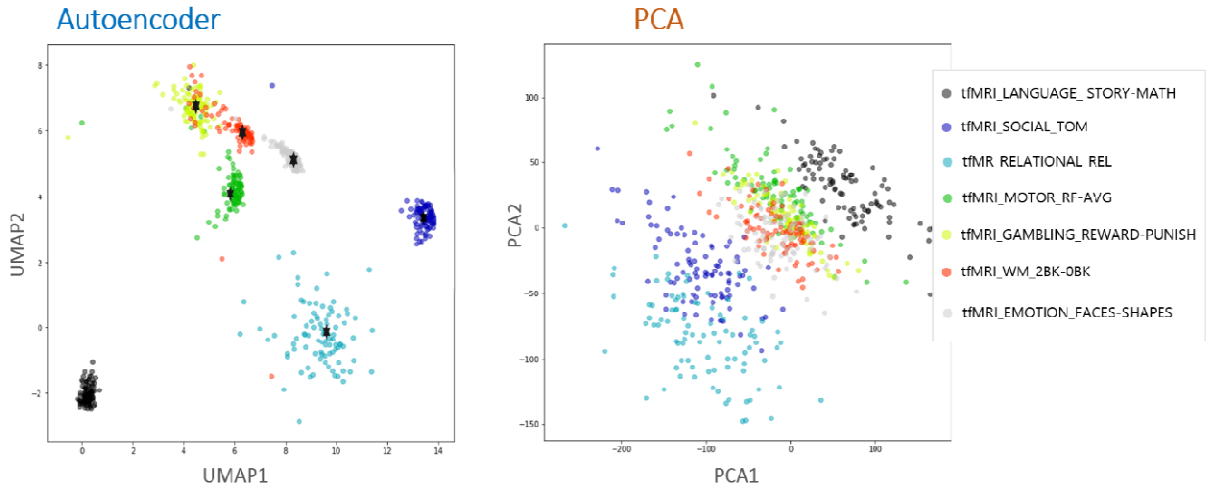


*Figure 4: UMAP of the latent space versus two PCA components for selected contrasts according to Barch 2013. *s are pointing to the centroid of clusters.*

Projection the latent representations to brain images

Figure 5 shows the centroids of contrasts that are back-projected from the UMAP latent space to the original brain space. The patterns of activations for these contrasts show an excellent correspondence with the expected task activations as shown in with previous studies (e.g. [25]). For instance, for language task, our projection of latent space to original image space shows the left lateralization which is accord with previous findings[25].

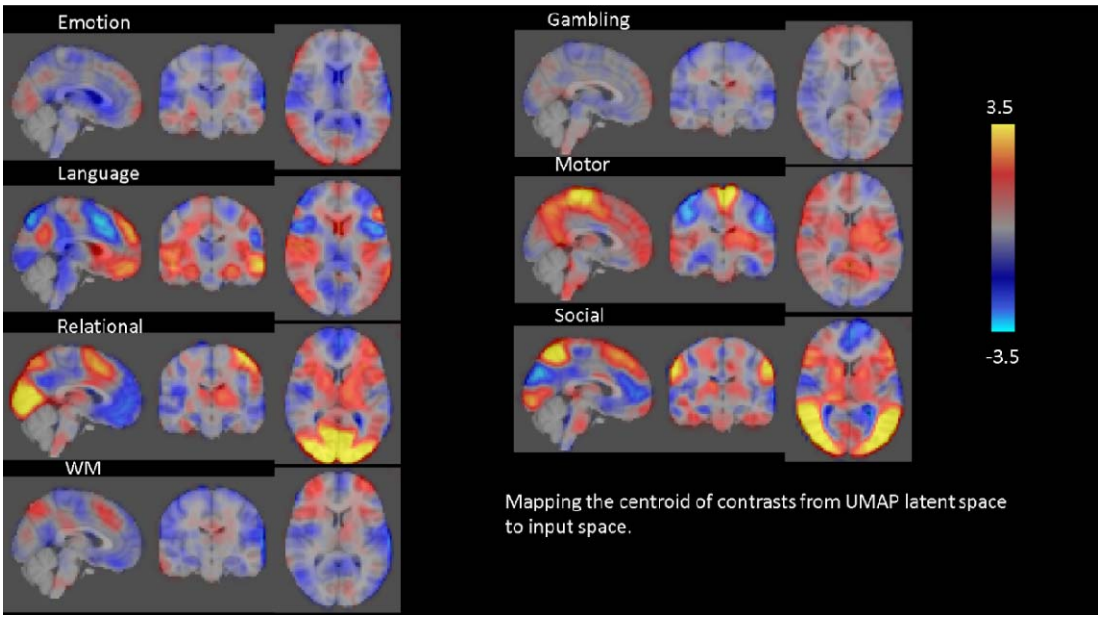


Figure 5: The projection of centroid of UMAP in the latent space to the input brain space. The centers of UMAP of latent space were calculated using K-means clustering across the test data. The centroids corresponding to each contrast were passed to encoder of autoencoder to map to input original space.

Association between latent variables and non-imaging covariates

The radar plots in Figure 6 shows the overall association across contrasts per each task in HCP data. These plots indicate that the autoencoder's latent variables have stronger associations with non-imaging variables in comparison to PCA.

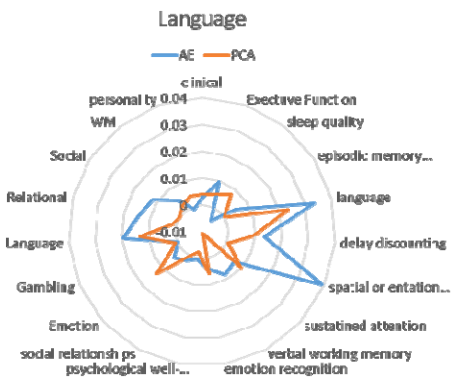
Figure 7 shows the same association for the most important contrasts from social, language, gambling and emotion tasks selected [25] (See supplementary information more extensive results). This shows that the autoencoder consistently produces a stronger association with behavior than the PCA (i.e. produces higher correlation in nearly all contrasts).

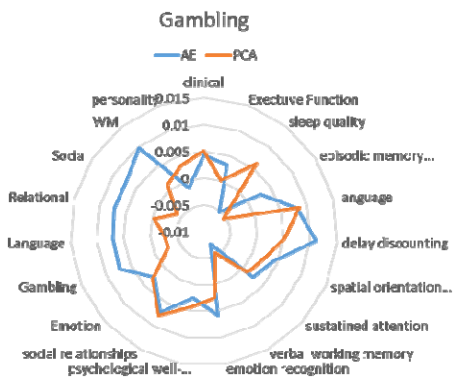
Parameter transfer from HCP data to UK Biobank The performance of PCA and fine-tuned model on UK Biobank data (Emotion, face-Shape contrast) are shown in (Table 2) indicating that the autoencoder performed equivalently well to PCA in terms of the reconstruction error and explained variance (or slightly better).

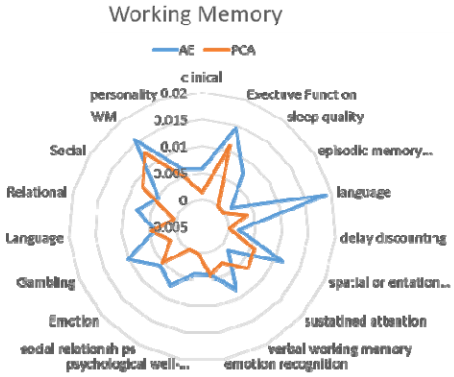
Table 2: Mean square for PCA model and Autoencoder on UK Biobank data

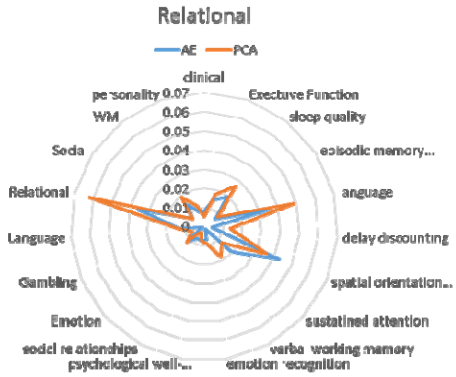
	MSE	EXPLAINED VARIANCE
HCP-PCA(100)	0.529	0.33
Fine-tuned AE(100)	0.458	0.38

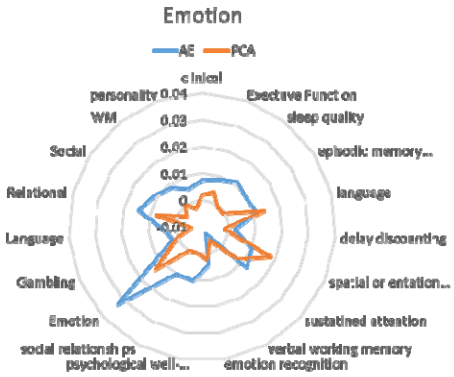
Figure 8 shows the Manhattan plot of p-value of univariate correlation between non-imaging measures and brain measures; UMAP components of latent space in addition to mean square error. Overall, the autoencoder detected 19 significant Bonferroni-corrected associations with behavior relative to 3e for PCA, demonstrating that after transfer learning to a second independent dataset, the autoencoder also provides a latent space that links more strongly to behavior. Here, PCA principally detects age while autoencoder links to more high-level features e.g., prospective memory, fluid intelligence, length of working weak or mother's age family history. The Manhattan plots of all of contrasts are available in supplementary information.

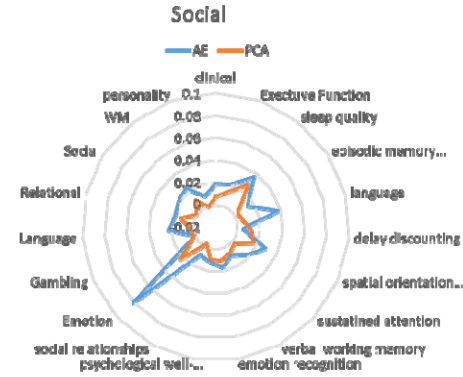

Language
AE
PCA
clinical
Executive Function
sleep quality
episodic memory...
language
delay discounting
spatial orientation...
sustained attention
verbal working memory
emotion recognition
psychological well-...
social relationships
Emotion
Gambling
Language
Relational
Social
WM
personality
0.04
0.03
0.02
0.01
0
-0.01


Gambling
AE
PCA
clinical
Executive Function
sleep quality
episodic memory...
language
delay discounting
spatial orientation...
sustained attention
verbal working memory
emotion recognition
psychological well-...
social relationships
Emotion
Gambling
Language
Relational
Social
WM
personality
0.015
0.01
0.005
0
-0.005
-0.01


Working Memory
AE
PCA
clinical
Executive Function
sleep quality
episodic memory...
language
delay discounting
spatial orientation...
sustained attention
verbal working memory
emotion recognition
psychological well-...
social relationships
Emotion
Gambling
Language
Relational
Social
WM
personality
0.02
0.015
0.01
0.005
0
-0.005


Relational
AE
PCA
clinical
Executive Function
sleep quality
episodic memory...
language
delay discounting
spatial orientation...
sustained attention
verbal working memory
emotion recognition
psychological well-...
social relationships
Emotion
Gambling
Language
Relational
Social
WM
personality
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0


Emotion
AE
PCA
clinical
Executive Function
sleep quality
episodic memory...
language
delay discounting
spatial orientation...
sustained attention
verbal working memory
emotion recognition
psychological well-...
social relationships
Emotion
Gambling
Language
Relational
Social
WM
personality
0.04
0.03
0.02
0.01
0
-0.01


Social
AE
PCA
clinical
Executive Function
sleep quality
episodic memory...
language
delay discounting
spatial orientation...
sustained attention
verbal working memory
emotion recognition
psychological well-...
social relationships
Emotion
Gambling
Language
Relational
Social
WM
personality
0.1
0.08
0.06
0.04
0.02
0
-0.02

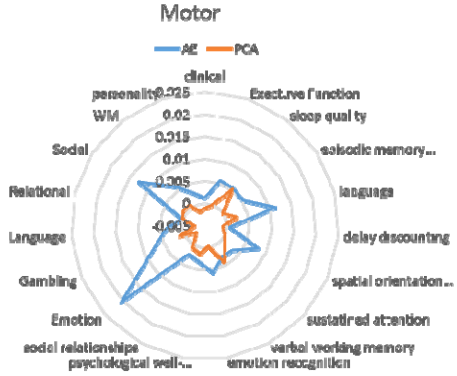

Motor
AE
PCA
clinical
Executive Function
sleep quality
episodic memory...
language
delay discounting
spatial orientation...
sustained attention
verbal working memory
emotion recognition
psychological well-...
social relationships
Emotion
Gambling
Language
Relational
Social
WM
personality
0.025
0.02
0.015
0.01
0.005
0
-0.005

Figure 6: The summary of association of latent variables and non-imaging measures across different behavioral categories in HCP data. Note that this is the average of association (OLS adjusted-r) over five separated models. Autoencoder indicates stronger associations in most of targets compare to PCA.

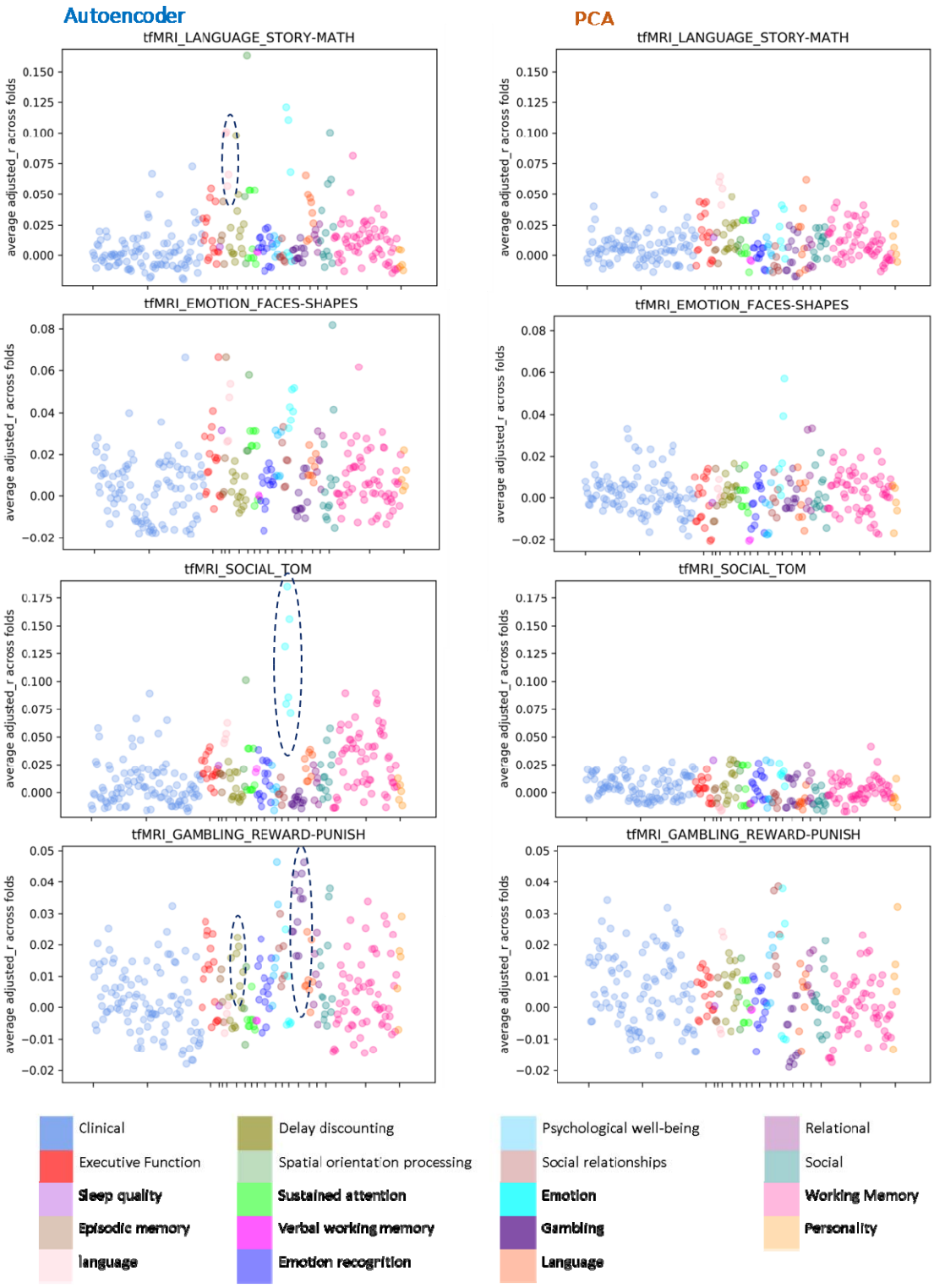


Figure 7: The association of UMAP of latent variables and non-brain imaging variables in HCP contrasts. Showing contrasts were selected according to 25. The association (adjusted-r score of OLS) of each contrast were averaged across five-fold cross-validation. The dotted circles show the higher associations for some behavioral measures in the relevant task in the autoencoder model compared to PCA.

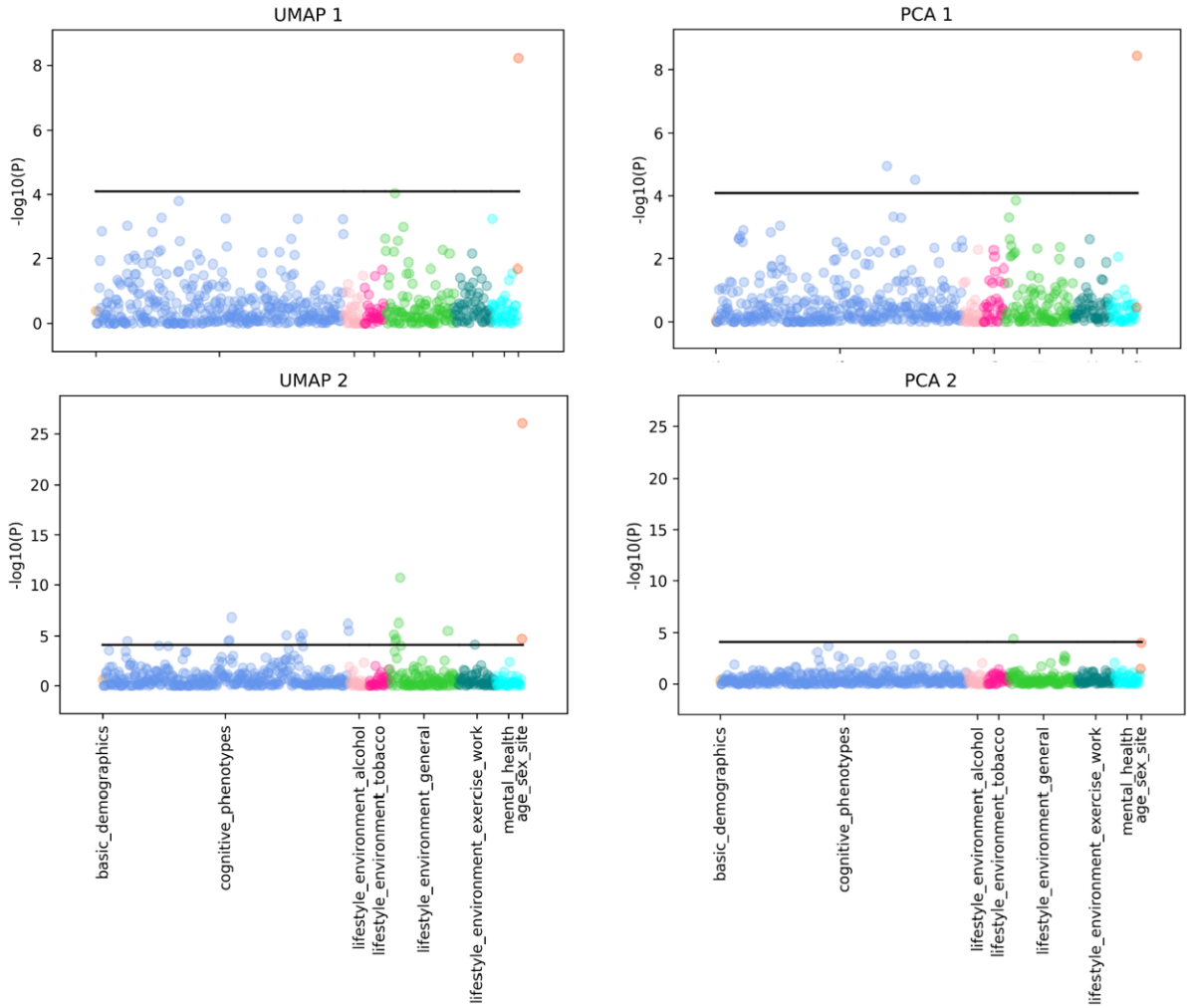


Figure 8: Manhattan plot of p-value of univariate correlation of non-imaging measures with UMAP of latent variables (left) and first two components of PCA (right) for faces-Shapes subtask. The black line is Bonferroni-corrected p-value threshold

Performance of Semi-unsupervised architecture

Results of the semi-supervised autoencoder for age and sex prediction are presented in Table 3. Note that the range of ages for HCP and UKB is very different.

Table 3: Semi-supervised autoencoder prediction performance

	BALANCED ACCURACY	**SEX**	**AGE MAE**
HCP	83% ± 3%		3.03 ± 0.10
UKB	89% ± 4%		4.44 ± 0.25

Figure 9 shows the UMAP of latent space of Emotion task (supervised and unsupervised nodes) is differentiating females from males and generating an age continuum.

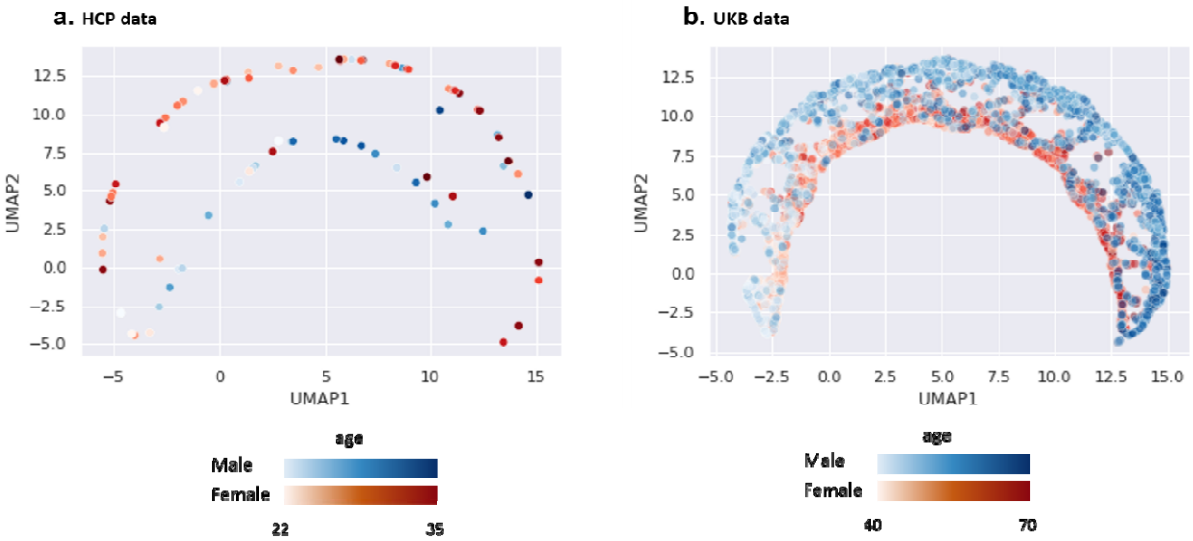


Figure 9: UMAP of latent space of EMTION- FACE-SHAPE task using semi-unsupervised autoencoder a) HCP b)UKB

Discussion

In this study, we present an optimal 3-D convolutional autoencoder architecture for non-linear transformation of fMRI data to a low-dimensional yet informative latent space. We present a coherent strategy for optimizing parameters and architecture of the model and a method to visualize and interpret the learned latent space representation. We showed that our model learned not only salient features that capture age and other sources of population stratification but also high-level behavioral features and that this representation was highly generic and generalized to the UK Biobank population cohort as an independent dataset.

Task-fMRI data

Regarding the dataset, HCP task-fMRI data enabled us to estimate a generic latent space representation with diverse cognitive tasks and behavioral and clinical scores [24,25]. Notably, the HCP data provides good whole brain coverage across all the tasks [25]. During the training, this comprehensive mapping allows the autoencoder to learn the various activation patterns across the brain instead of learning specific task-related

effects that may be localized to particular brain regions. To validate the generalizability of latent representation of HCP, we used UKB . Complementarily, UK-biobank contains the Hariri faces-shapes emotion task [60], which is very similar to the emotion task of HCP (effectively a shorter version). The common contrasts provide a great opportunity for further validation of the model and test the across-cohort generalization of the latent space.

Latent space

The HCP-derived UMAP representation illustrates that the autoencoder can better differentiate between tasks and contrasts compared to a linear model like PCA. As shown in Figure 4, contrasts are well-separated in the latent space in comparison to the first two components of PCA. This observation demonstrates that the resulting latent space indicated meaningful features in the data. In contrast, the PCA representation does not separate the different task contrasts as cleanly, potentially leading to a mixing or averaging of different signals. Linear methods like PCA tend to identify prominent information in the data, e.g. age, site, sex. Indeed, these variables were the only features that were associated with the latent representation derived from PCA in UK Biobank. These features have a substantial effect on data distribution and are easily identified by linear models, in contrast complex behavioral associations are often missed [21]. Here, using a nonlinear transformation, we aim to identify more complex features. The fact that our nonlinear mapping is more sensitive to the different tasks compared to PCA indicates that the nonlinear latent space can retrieve a different and more informative dimensions in data.

Autoencoders reduce a complex and multidimensional input space into a non-linear combination of latent factors. In contrast with linear decomposition methods such as PCA or supervised learning methods, which are linked to a specific decision problem, the work presented here, provides and effective tool to distill a non-linear representation of images and a powerful and more accurate pipeline to detect anomalies in brain images [5,15,61].

Projection of latent representation to original space:

The back-projection of the latent space shows that we may be able to translate the learned features to a clinical biomarker. For the majority of the contrasts and particularly language (story-math), social (theory of mind) and relational (relational-baseline), the projection of the center of K-means of latent space to the original scan image space

were in line with findings in [25]. It is an increasingly challenging problem to figure out how to adapt machine learning algorithms to the discourses we carry on in neuroimaging[62,63]. In this context, the meaningful projection of the latent space can be viewed as an example of explainable AI in complex models.

By adding age and sex to the model, we provide a method to explore the functional and anatomical manifolds of brain states by controlling what the autoencoder represents. While unsupervised training of the model yields interpretable representations of different tasks, using semi-supervised autoencoder, our representation was able to be tailored to focus on specific differences. We illustrate this by training an autoencoder that simultaneously reconstructs the data, whilst also predicting age and sex. Importantly, this results in an interpretable latent manifold that clearly reflects individual differences related to the representation of demographic variables in the underlying imaging data.

Association of the latent representation with non-imaging variables

Overall, the latent variables extracted by autoencoder shows stronger association with non-imaging variables compared with PCA. In the HCP dataset, the autoencoder has shown stronger associations with the non-imaging variables in six of the seven tasks, excepting the relational task where PCA showed a stronger association with behavior. Note that the radar plot is the average correlation of the subtasks with each non-imaging variable within the set of clinical and cognitive measures [51]. The detailed correlation map shown in Figure 6 and Figure 7, indicates that PCA and autoencoder have a comparable performance for captured brain patterns related to categories like clinical and working memory. Importantly, many interesting measures are not associated with the PCA components but are strongly associated with the autoencoder's latent variables. For instance, many social relationships category measures in the Theory of Mind (TOM) contrast show strong correlations with latent variables from the autoencoder but not in the PCA model (dotted circle in Figure 7).

Another example of relevant-behaviors association with latent variables is the reward-punish contrast; while the overall performance in terms of correlation strength sounds similar in two models, the gambling category measures are showing relatively high correlation with latent variables in the autoencoder but not with PCA components.

These results suggest strongly that our autoencoder can be adjusted to identify and predict a different set of non-imaging measures and therefore contains potentially

complementary information to linear models. Note that the HCP autoencoder was trained using all the contrasts- in contrast with UK Biobank which was trained only on single Emotion task- hence, the network may only learn the features relevant to all the contrasts and lose the information on contrasts variability. In spite of that, we still can find relevant non-imaging measures associated with the latent variables, showing that the nonlinear manifold may contain interesting information about the data, which can be simply identified through correlation-based correspondence. The association plots show that the derived latent space on UK Biobank data is associated with otherwise challenging-to-predict measures in the lifestyle and the cognitive-phenotype category alongside sex and age. Whilst we recognize the possibility of many confounding effects in large cohorts like UK Biobank [64], our goal was to show that autoencoders produce a meaningful representation for finding brain-behavior associations. Therefore, we consider that a detailed examination of the potential effects of confounding variables was out of the scope of this paper.

Network architecture

The architectural hyper-parameters of the autoencoder were chosen during the pilot study, solely based on how the models performed in terms of the reconstruction error and no other readouts i.e., non-imaging measures were used for evaluations of the models.. Some decisions about the network structure have been made before estimating the model. For example, to preserve the morphology of the images and hence better interpretability , we decided to use a 3-D convolutional network [34–36,45]. In order to control order of latent space, we used dense layer in the bottleneck of the autoencoder [49].

We emphasize that we designed our autoencoder with the specific nature of our high-dimensional neuroimaging data in mind and therefore, a number of constraints were imposed on the model beforehand. For example, the networks evaluated were not particularly deep, also to reduce the memory usage and computational complexity, we took advantage of the weight sharing of convolutional layers. Here, we are in search of low-level features that may be translation invariant, but a more important benefit is that the weight sharing enables the networks to be scaled to whole-brain data [65]. The kernel size was set to be $3 \times 3 \times 3$ to keep the details of the downsampled image scans. Average pooling layers were positioned right after each convolutional layer to ignore the sharp features, reduce the number of parameters and consequently, minimize the

chance of overfitting. We relied on the pilot study to select the rest of the model's parameters, such as the number of filters.

While the new trend in neuroimaging is to use linear transformation instead of nonlinear models [12,66,67], here, we show that employing nonlinear mappings can add value. We showed that if the nonlinear model's hyperparameters are selected with careful consideration, then we can uncover particular features in the data that are not often revealed by linear models. As a matter of fact, having a solid methodological pipeline is not specific to complex nonlinear models, but since the chance of over-engineering and hence overfitting is higher than these methods, it is necessary to carefully develop the architectural model design.

Here, we employed a model with a simple reconstruction loss function while the network's primary goal was finding the relevant features to the non-imaging variables. Accordingly, we under-trained our network to preserve the essential aspects of the data, which could be assumed noise from the loss function perspective. It should be noted that our vanilla network is entirely unsupervised, although we also show how additional information e.g., age can fed to the loss. This results in a semi-supervised setting that allows the latent space to partially encode particular features of the data [15]. Another interesting future direction is to train an autoencoder to predict different data (e.g., a follow-up timepoint in longitudinal studies). This would serve to sensitize the latent space to changes relevant to ageing or pathology.

The increased number of neuroimaging scans provides a unique opportunity to transcend the linear mapping, but it is also necessary to acknowledge some limitations. The traditional image processing techniques are not completely applicable here. For example, while data augmentation using image mirroring, flipping, skewing, or segmenting is a straightforward approach to increase the number of samples and has been applied before in neuroimaging applications [13], we did not consider it to be appropriate here because such augmentation strategies do not faithfully preserve invariances known to occur in the brain, for example the lateralization of brain functions e.g. the association of left lateralization in language processing [68]. Another limitation is computational complexity. Training an autoencoder on large neuroimaging data is computationally more demanding comparing with similar linear models.

Conclusion

This study aimed to explore nonlinear manifold structures embedded in brain imaging using 3-dimensional autoencoders. We presented a coherent strategy for optimizing parameters and architecture of the model and a method to visualize and interpret the learned latent space representation. We showed that our model learned not only salient features like age but also high-level behavioral features and that this representation was interpretable, highly generic and generalizable to an independent dataset. Finally, our results show that with a careful implementation, nonlinear features can provide complementary information that is not accessible to linear methods.

References

1. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal*. 2017;42(December 2012):60-88. doi:10.1016/j.media.2017.07.005

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

3. Durstewitz D, Koppe G, Meyer-Lindenberg A. Deep neural networks in psychiatry. *Mol Psychiatry*. February 2019:1. doi:10.1038/s41380-019-0365-9

4. Suk H Il, Lee SW, Shen D. Deep ensemble learning of sparse regression models for brain disease diagnosis. *Med Image Anal*. 2017;37:101-113. doi:10.1016/j.media.2017.01.008

5. Vieira S, Pinaya WHL, Mechelli A. Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neurosci Biobehav Rev*. 2017;74:58-75. doi:10.1016/J.NEUBIOREV.2017.01.002

6. Wolfers T, Buitelaar JK, Beckmann CF, Franke B, Marquand AF. From estimating activation locality to predicting disorder: A review of pattern recognition for neuroimaging-based psychiatric diagnostics. *Neurosci Biobehav Rev*. 2015;57:328-349. doi:10.1016/j.neubiorev.2015.08.001

7. Mensch A, Mairal J, Bzdok D, Thirion B, Varoquaux G. Learning Neural Representations of Human Cognition across Many fMRI Studies. 2017;(Nips).

8. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *Neuroimage*. 2017;145:137-165. doi:10.1016/j.neuroimage.2016.02.079

9. Wolfers T, Floris DL, Dinga R, et al. From pattern classification to stratification: towards conceptualizing the heterogeneity of Autism Spectrum Disorder. *Neurosci Biobehav Rev*. 2019;104:240-254. doi:10.1016/j.neubiorev.2019.07.010

10. Davatzikos C. Machine learning in neuroimaging: Progress and challenges. *Neuroimage*. 2019;197:652.

11. He T, Kong R, Holmes AJ, et al. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *Neuroimage*. 2020;206. doi:10.1016/j.neuroimage.2019.116276

12. Schulz MA, Yeo BTT, Vogelstein JT, et al. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nat Commun*. 2020;11(1). doi:10.1038/s41467-020-18037-z

13. Peng H, Gong W, Beckmann CF, Vedaldi A, Smith SM. Accurate brain age prediction with lightweight deep neural networks. *Med Image Anal*. 2021;68:101871. doi:10.1016/j.media.2020.101871

14. Bellman RE. *Adaptive Control Processes: A Guided Tour*. Vol 2045. Princeton university press; 2015.

15. Pinaya WHL, Mechelli A, Sato JR. Using deep autoencoders to identify abnormal brain structural patterns in neuropsychiatric disorders: A large-scale multi-sample study. *Hum Brain Mapp*. 2019;40(3):944-954. doi:10.1002/hbm.24423

16. Pinaya WHL, Scarpazza C, Garcia-Dias R, et al. Normative modelling using deep autoencoders: a multi-cohort study on mild cognitive impairment and Alzheimer's disease. *bioRxiv*. 2020.

17. Cole JH, Poudel RPK, Tsagkrasoulis D, et al. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *Neuroimage*. 2017;163:115-124.

18. Dinsdale NK, Bluemke E, Smith SM, et al. Learning patterns of the ageing brain in MRI using deep convolutional networks. *Neuroimage*. 2021;224:117401. doi:10.1016/j.neuroimage.2020.117401

19. Kiesow H, Spreng RN, Holmes AJ, et al. Hidden population modes in social brain morphology: Its parts are more than its sum. *bioRxiv*. August 2020:2020.08.07.241497. doi:10.1101/2020.08.07.241497

20. Poldrack RA, Baker CI, Durnez J, et al. Scanning the horizon: Towards transparent and reproducible neuroimaging research. *Nat Rev Neurosci*. 2017;18(2):115-126. doi:10.1038/nrn.2016.167

21. Mwangi B, Tian TS, Soares JC. A review of feature reduction techniques in Neuroimaging. *Neuroinformatics*. 2014;12(2):229-244. doi:10.1007/s12021-013-9204-3

22. Korolev S, Safiullin A, Belyaev M, Dodonova Y. Residual and plain convolutional neural networks for 3D brain MRI classification. *Proc - Int Symp Biomed Imaging*. 2017:835-838. doi:10.1109/ISBI.2017.7950647

23. Gong W, Beckmann CF, Smith SM. Phenotype Discovery from Population Brain Imaging. *bioRxiv*. 2020. doi:10.1101/2020.03.05.973172

24. Van Essen DC, Smith SM, Barch DM, Behrens TEJ, Yacoub E, Ugurbil K. The WU-Minn Human Connectome Project: An overview. *Neuroimage*. 2013;80:62-79. doi:10.1016/j.neuroimage.2013.05.041

25. Barch DM, Burgess GC, Harms MP, et al. Function in the human connectome: Task-fMRI and individual differences in behavior. *Neuroimage*. 2013;80:169-189. doi:10.1016/j.neuroimage.2013.05.033

26. Smith SM, Nichols TE, Vidaurre D, et al. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nat Neurosci*. 2015;18(11):1565-1567. doi:10.1038/nn.4125

27. Finn ES, Shen X, Scheinost D, et al. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat Neurosci*. 2015;18(October):1-11. doi:10.1038/nn.4135

28. Miller KL, Alfaro-Almagro F, Bangerter NK, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*. 2016;19(11):1523-1536. doi:10.1038/nn.4393

29. Gupta L, Besseling RMH, Overvliet GM, et al. Spatial heterogeneity analysis of brain activation in fMRI. *NeuroImage Clin*. 2014;5:266-276. doi:10.1016/j.nicl.2014.06.013

30. Burgess GC, Gray JR, Conway ARA, Braver TS. Neural mechanisms of interference control underlie the relationship between fluid intelligence and working memory span. *J Exp Psychol Gen*. 2011. doi:10.1037/a0024695

31. Bzdok D. Classical statistics and statistical learning in imaging neuroscience. *Front Neurosci*. 2017;11(OCT):543. doi:10.3389/fnins.2017.00543

32. Suk H Il, Lee SW, Shen D. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Struct Funct*. 2015;220(2):841-859. doi:10.1007/s00429-013-0687-3

33. Bengio Y. Deep learning of representations for unsupervised and transfer learning. In: *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*. ; 2012:17-36.

34. Tudosiu P-D, Varsavsky T, Shaw R, et al. Neuromorphologicaly-preserving Volumetric data encoding using VQ-VAE. *arXiv*. February 2020:1-13. http://arxiv.org/abs/2002.05692. Accessed January 6, 2021.

35. Kwon G, Han C, Kim D. Generation of 3D Brain MRI Using Auto-Encoding Generative Adversarial Networks. *Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*. 2019;11766 LNCS:118-126. http://arxiv.org/abs/1908.02498. Accessed January 6, 2021.

36. Choi H, Kang H, Lee DS. Predicting aging of brain metabolic topography using variational autoencoder. *Front Aging Neurosci*. 2018;10(JUL):212. doi:10.3389/fnagi.2018.00212

37. Huang H, Hu X, Zhao Y, et al. Modeling Task fMRI Data Via Deep Convolutional Autoencoder. *IEEE Trans Med Imaging*. 2018;37(7):1551-1561. doi:10.1109/TMI.2017.2715285

38. Bunea F, She Y, Ombao H, Gongvatana A, Devlin K, Cohen R. Penalized least squares regression methods and applications to neuroimaging. *Neuroimage*. 2011;55(4):1519-1527. doi:10.1016/j.neuroimage.2010.12.028

39. Sidhu G, Asgarian N, Greiner R, Brown MRG. Kernel principal component analysis for dimensionality reduction in fMRI-based diagnosis of ADHD. *Front Syst Neurosci*. 2012;6(October):1-17. doi:10.3389/fnsys.2012.00074

40. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. *Neuroimage*. 2017;145:166-179. doi:10.1016/j.neuroimage.2016.10.038

41. Calhoun VD, Liu J, Adali T. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage*. 2009;45(1 Suppl). doi:10.1016/j.neuroimage.2008.10.057

42. Thirion B, Faugeras O. Dynamical components analysis of fMRI data through kernel PCA. *Neuroimage*. 2003;20(1):34-49. doi:10.1016/S1053-8119(03)00316-1

43. Bzdok D, Yeo BTT. Inference in the age of big data: Future perspectives on neuroscience. *Neuroimage*. 2017;155:549-564. doi:10.1016/j.neuroimage.2017.04.061

44. Smith SM, Nichols TE. Statistical Challenges in "Big Data" Human Neuroimaging. *Neuron*. 2018;97(2):263-268. doi:10.1016/j.neuron.2017.12.018

45. Payan A, Montana G. Predicting Alzheimer ' s disease: a neuroimaging study with 3D convolutional neural networks. *arXiv Prepr arXiv150202506*. 2015:1-9.

46. Savage JE, Jansen PR, Stringer S, et al. Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet*. 2018;50(7):912-919. doi:10.1038/s41588-018-0152-6

47. Huang H, Hu X, Zhao Y, et al. Modeling Task fMRI Data Via Deep Convolutional Autoencoder. *IEEE Trans Med Imaging*. 2018;37(7):1551-1561. doi:10.1109/TMI.2017.2715285

48. Myronenko A. 3D MRI brain tumor segmentation using autoencoder regularization. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 11384 LNCS. Springer Verlag; 2019:311-320. doi:10.1007/978-3-030-11726-9_28

49. Goodfellow I, Bengio Y, Courville A, Bengio Y. *Deep Learning*. Vol 1. MIT press Cambridge; 2016.

50. Littlejohns TJ, Holliday J, Gibson LM, et al. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. *Nat Commun*. 2020;11(1):2624. doi:10.1038/s41467-020-15948-9

51. Bzdok D, Varoquaux G, Grisel O, Eickenberg M, Poupon C, Thirion B. Formal Models of the Network Co-occurrence Underlying Mental Operations. Bassett DS, ed. *PLOS Comput Biol*. 2016;12(6):e1004994. doi:10.1371/journal.pcbi.1004994

52. Bzdok D, Eickenberg M, Grisel O, Thirion B, Varoquaux Semi G, Varoquaux G. *Semi-Supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data*.; 2015. https://hal.archives-ouvertes.fr/hal-01211248. Accessed February 17, 2021.

53. Buysse DJ, Reynolds CF, Monk TH, Berman SR, Kupfer DJ. The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Res*. 1989;28(2):193-213. doi:10.1016/0165-1781(89)90047-4

54. Smith SM, Nichols TE, Vidaurre D, et al. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nat Neurosci*. 2015;18(11):1565-1567. doi:10.1038/nn.4125

55. Marquand AF, Haak K V., Beckmann CF. Functional corticostriatal connection topographies predict goal-directed behaviour in humans. *Nat Hum Behav*. 2017;1(8):146. doi:10.1038/s41562-017-0146

56. Glasser MF, Sotiropoulos SN, Wilson JA, et al. The minimal preprocessing pipelines for the Human Connectome Project. *Neuroimage*. 2013;80:105-124. doi:10.1016/j.neuroimage.2013.04.127

57. Alfaro-Almagro F, Jenkinson M, Bangerter NK, et al. Image processing and Quality Control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage*. 2018;166:400-424. doi:https://doi.org/10.1016/j.neuroimage.2017.10.034

58. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv*. February 2018. http://arxiv.org/abs/1802.03426. Accessed January 6, 2021.

59. der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res*. 2008;9(11).

60. Hariri AR, Tessitore A, Mattay VS, Fera F, Weinberger DR. The amygdala response to emotional stimuli: A comparison of faces and scenes. *Neuroimage*. 2002;17(1):317-323. doi:10.1006/nimg.2002.1179

61. Perera P, Patel VM. Learning deep features for one-class classification. *IEEE Trans Image Process*. 2019;28(11):5450-5463.

62. Heinrichs B, Eickhoff SB. Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Hum Brain Mapp*. 2020;41(6):1435-1444. doi:10.1002/hbm.24886

63. Bzdok D, Ioannidis JPA. Exploration, inference, and prediction in neuroscience and biomedicine.

Trends Neurosci. 2019;42(4):251-262.

64. Alfaro-Almagro F, McCarthy P, Afyouni S, et al. Confound modelling in UK Biobank brain imaging. *Neuroimage*. 2021;224:117002.

65. Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539

66. D'amour A, Heller K, Moldovan D, et al. Underspecification in Machine Learning Underspecification Presents Challenges for Credibility in Modern Machine Learning. *arXiv Prepr arXiv201103395*. 2020.

67. Saxe A, Nelli S, Summerfield C. If deep learning is the answer, what is the question? *Nat Rev Neurosci*. 2020. doi:10.1038/s41583-020-00395-8

68. Frost JA, Binder JR, Springer JA, et al. Language processing is strongly left lateralized in both sexes. Evidence from functional MRI. *Brain*. 1999;122(2):199-208. doi:10.1093/brain/122.2.199