Supervised Phenotype Discovery from Multimodal Brain Imaging

Weikang Gong[a], Song Bai[b], Ying-Qiu Zheng[a], Stephen M. Smith[a], Christian F. Beckmann[a,c,d]

[a] *Centre for Functional MRI of the Brain (FMRIB), Nuffield Department of Clinical Neurosciences, Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK.*
[b] *Department of Engineering Science, University of Oxford, Oxford, UK.*
[c] *Radboud University Medical Centre, Department of Cognitive Neuroscience, Nijmegen, Netherlands.*
[d] *Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, Nijmegen, Netherlands.*

Abstract

Data-driven discovery of image-derived phenotypes (IDPs) from large-scale multimodal brain imaging data has enormous potential for neuroscientific and clinical research by linking IDPs to subjects' demographic, behavioural, clinical and cognitive measures (i.e., non-imaging derived phenotypes or nIDPs). However, current approaches are primarily based on unsupervised approaches, without use of information in nIDPs. In this paper, we proposed Supervised BigFLICA (SuperBigFLICA), a semi-supervised, multimodal, and multi-task fusion approach for IDP discovery, which simultaneously integrates information from multiple imaging modalities as well as multiple nIDPs. SuperBigFLICA is computationally efficient and largely bypasses the need for parameter tuning. Using the UK Biobank brain imaging dataset with around 40,000 subjects and 47 modalities, along with more than 17,000 nIDPs, we showed that SuperBigFLICA enhances the prediction power of nIDPs, benchmarked against IDPs derived by conventional expert-knowledge and unsupervised-learning approaches (with average nIDP prediction accuracy improvements of up to 46%). It also enables learning of generic imaging features that can predict new nIDPs. Further empirical analysis of the SuperBigFLICA algorithm demonstrates its robustness in different prediction tasks and the ability to derive biologically meaningful IDPs in predicting health outcomes and cognitive nIDPs, such as fluid intelligence and hypertension scores.

Keywords: Multimodality, Brain imaging, UK Biobank, Imaging-derived phenotypes, Non-imaging derived phenotypes.

1. Introduction

Large-scale population neuroimaging datasets, such as the data from UK Biobank, provide high-quality multimodal magnetic resonance imaging (MRI) data, with the potential for generating markers of psychiatric and neurodegenerative diseases and uncovering the neural basis of cognition through linking across imaging features to behavioural or genetic data (Miller et al., 2016). However, such massive high-dimensional data make statistical modelling challenging due to their multimodal nature and cohort size. Therefore, instead of working directly from voxel-level spatial maps, it is becoming popular to reduce these maps into summary measures, sometimes referred to as 'imaging-derived phenotypes (IDPs)'(Gong et al., 2021; Dadi et al., 2020; Elliott et al., 2018). IDPs can be spatial summary statistics such as global and regionally averaged tissue volumes, while other IDPs can be measures of functional and structural connectivity or tissue biology. Building statistical models from an informative set of IDPs can significantly reduce the computational burden and, compared to working from voxel-wise data, has a similar or even improved signal-to-noise ratio for use in associations with non-imaging variables and predictive analysis linking to, e.g., behaviour and genetics.

Methods for deriving IDPs can be divided into two categories, one expert-knowledge-based and the other data-driven. The former approaches are typically concerned with extracting summary signals from pre-defined anatomy or functional brain atlases (Eickhoff et al., 2018). Although simple and efficient, this approach has a few limitations. First, the atlases may not be equally valid across different areas of the brain. For example, existing atlases typically provide fine-grained delineations across sensory cortices and less detailed across multimodal association cortices. These differences may result in increased inter-individual differences across different brain areas, potentially masking the signal of interest. Second, these regional characterisations are often derived from underlying features that may not appropriately map onto different data modalities. For example, atlases based on cytoarchitectonic features may differentially be suitable for IDPs reflecting regional cortical thickness but may be less suitable for summarising measures of functional connectivity. Furthermore, with

[*] Correspondence should be addressed to: Weikang Gong. E-mail: weikang.gong@ndcn.ox.ac.uk
[**] SS and CB are co-senior authors.

multimodal data, expert-knowledge-based approaches typically ignore cross-modal relationships and thus have limited ability to capture continuous modes of variations shared by different modalities. Data-driven approaches for identifying IDPs, e.g., variants of unsupervised spatial dimensionality reduction techniques, may overcome the aforementioned limitations of expert-knowledge-based approaches. For example, independent component analysis (ICA) and dictionary learning (DicL) have been widely used to define "soft" brain parcellations in resting-state functional MRI analysis (Beckmann and Smith, 2004; Varoquaux et al., 2011). They are based on arguably objective criteria such as maximising non-Gaussianity or minimising data reconstruction errors and can, in theory, be applied to a wide variety of different modalities. In a multimodal setting, FMRIB's Linked ICA (FLICA) (Groves et al., 2011) is one approach for identifying continuous spatial modes of individual variations that are related to a range of behavioural phenotypes and diseases (e.g., lifespan development (Douaud et al., 2014) and attention deficit hyperactivity disorder (Ball et al., 2019)). In our previous work, we developed BigFLICA, extending the original computationally expensive FLICA to handle larger datasets such as UK Biobank (Gong et al., 2021). These data-driven approaches have the advantages of being objective and considering cross-modal relationships, thereby revealing patterns that are ignored by expert-knowledge-based approaches (Calhoun and Sui, 2016; Uludağ and Roebroeck, 2014).

One of the primary applications of extracting imaging features as IDPs is predicting non-imaging derived phenotypes (nIDPs), including demographic, behavioural, clinical and cognitive measures from individuals. While the approaches listed above are designed to capture spatial modes of variation from the imaging data faithfully, they are not explicitly optimised for the latter prediction task. Incorporating the "target" nIDP information into IDP discovery, therefore, may benefit IDP extraction. Various studies have proposed (semi-)supervised approaches for IDP discovery, e.g., Qi et al. (2017) developed a multimodal fusion with reference approach and applied it to find multimodal modes related to schizophrenia (Sui et al., 2018) and major depressive disorder (Qi et al., 2018). Another line of research focused on complex nonlinear approaches, such as multiple kernel learning (Zhang et al., 2012a; Zhou et al., 2020; Liu et al., 2020), graph-based transductive learning (Wang et al., 2017) and neural networks such as multilayer perceptrons (Lu et al., 2018; Lee et al., 2019), which proved successful in predicting neurological disorders such as Alzheimer's disease. However, two caveats still exist in the above approaches. First, most of them do not scale well to big datasets due to expensive computational loads and high memory requirements. Second, nonlinear approaches heavily rely on parameter tuning and therefore require additional (cross-)validation for setting appropriate values. Furthermore, it is often difficult to make meaningful interpretations of the"black-box" nonlinear approaches, as explanations for deep neural networks produced by existing methods largely remain elusive and are yet to be standardised (Adadi and Berrada, 2018; Gilpin et al., 2018). As a result, it remains difficult to interpret the neural system that each feature (or spatial summary statistic) represents.

In this paper, to address these issues, we developed 'Supervised BigFLICA' (SuperBigFLICA), a semi-supervised, multimodal, and multi-task fusion approach for IDP discovery, which simultaneously integrates information from multiple imaging modalities as well as from multiple nIDPs. By incorporating nIDPs in the modelling, one can hope to achieve better nIDP prediction than by training on the imaging data alone, as this exploits the covariance structure inherent in the nIDP space in addition to the predictive power of the imaging data. In the model, we use one or more target nIDPs to help the model learn spatial features that are biologically important in that they are generically useful in prediction, rather than only taking the route of classical unsupervised approaches of simply focusing on learning features for representing/reconstructing the image data with minimal loss. Further, using multiple nIDPs in training - a technique known as multi-task learning (Zhang and Yang, 2017) - one can hope to refine the learned latent space better than when using single nIDPs, which are often noisy descriptions of the phenotype of interest (e.g., fluid intelligence). Compared with learning to predict each of the response variables individually, training across a range of noisy but related tasks simultaneously guides the model to characterise feature space shared across tasks, potentially leading to improved predictive power of the derived IDPs (Zhang and Yang, 2017; Rahim et al., 2017; Marquand et al., 2014). Additionally, the multi-task learning frameworks may still be useful even if one is interested in predicting unseen nIDPs (new tasks), because the latent space learned via a multi-task setting generically is more transferable and thus has higher predictive power.

SuperBigFLICA decomposes the imaging data into common 'subject modes' across modalities, which characterise the inter-individual variation of a given underlying spatial component, along with modality-specific sparse spatial loadings and weightings (Groves et al., 2011; Gong et al., 2021). It minimises a composite loss function, consisting of both reconstruction errors of the imaging data ('unsupervised learning') and the prediction errors of nIDPs ('supervised learning'), while additionally having constraints pushing for spatially sparse representations. Further, being built on a Bayesian framework, SuperBigFLICA can automatically balance the weights of different modalities and nIDPs, thereby aiming to largely bypass the need for parameter tuning. Optimised by a mini-batch stochastic gradient descent algorithm, SuperBigFLICA is computationally efficient and scalable to large datasets. In this study, we evaluate the performance of SuperBigFLICA across 39,770 UKB subjects, using 47 imaging modalities and 17,485 nIDPs. We show that SuperBigFLICA enhances the predictive power of the derived IDPs, benchmarked against those discovered by the conventional expert-knowledge

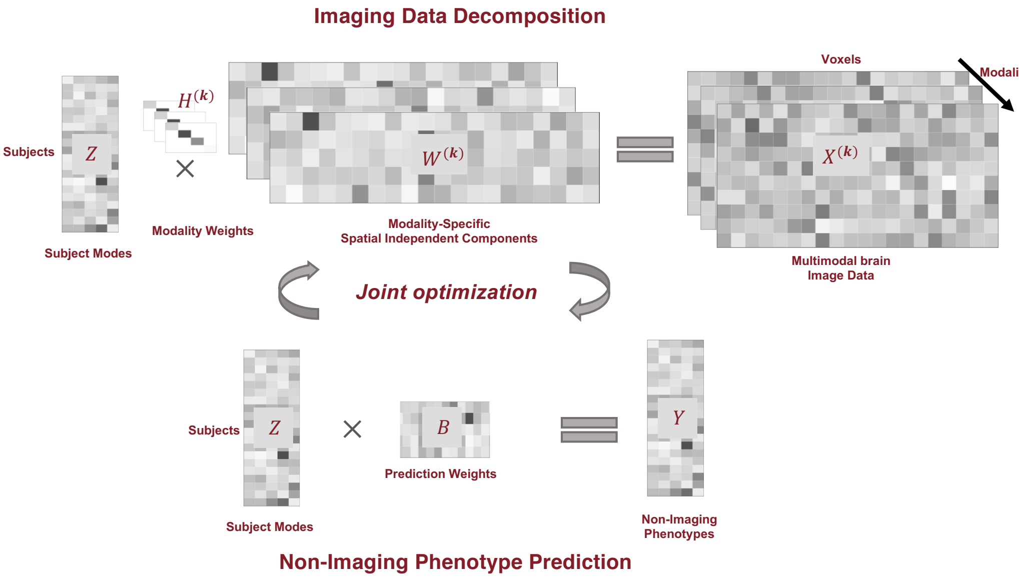

Figure 1: Overview of the proposed SuperBigFLICA approach for supervised multimodal fusion and phenotype discovery.

and unsupervised-learning approaches. Finally, we provide a comprehensive empirical analysis of the SuperBigFLICA algorithm and demonstrate its potential for predicting health outcomes and cognitive nIDPs.

2. Methods

2.1. *SuperBigFLICA from an optimisation perspective.*

We assume our data is being derived from a group of N subjects with multiple imaging modalities. Each of these modalities has been processed to produce one or more voxel-wise maps (or network matrices). For example, a task fMRI scan may produce several task contrast maps through statistical parametric mapping (Penny et al., 2011), and a diffusion MRI analysis can produce maps such as fractional anisotropy (FA) and mean diffusivity (MD) per subject. We assume that we have a total of K modality maps per subject, and each modality k is represented by a matrix $\mathbf{X}^{(\mathbf{k})}$ of size $N \times P_k$, where P_k is the number of feature values (e.g., voxels, tracts, areas, edges or vertices). We also assume that there are Q nIDPs per subject, summarised in a matrix $\mathbf{Y}$ of size $N \times Q$. We want to find an L-dimensional latent space across modalities, optimally predicting multiple nIDPs of interests in unseen subjects and representing the original imaging data. This latent space corresponds to the weights of continuous spatial modes representing inter-individual variations.

Formalising this in terms of a generative model, we will assume each modality map is generated as the product of the shared latent space, modality-specific spatial loadings and weights plus some Gaussian residual noise:

$$\mathbf{X}^{(\mathbf{k})} = ZH^{(k)}W^{(k)} + E^{(k)},\ k = 1, \dots, K \tag{1}$$

Meanwhile, the nIDPs of interest are generated by the product of shared latent space and the prediction weights plus some Gaussian residual noise:

$$\mathbf{Y} = ZB + E. \tag{2}$$

In the above two equations, $W^{(k)}_{(L\times P_k)}$ are the spatial loadings of the k-th modality map, which models the importance of each voxels to each latent dimension; $H^{(k)}_{(L\times L)}$ is a positive and diagonal modality weighting matrix (with $\sum_{k=1}^{K} H^{(k)}_{ll} = 1$), which reflects, for each component, the overall contribution of each modality; and $Z_{(N\times L)}$ is the latent features, i.e., the subject course shared across modalities; B is an $L \times Q$ matrix of prediction weights, which reflects the contribution of each latent dimension for predicting each nIDP; and finally, $E^{(k)}$ and E are independent Gaussian random error terms. **Fig.** 1 shows an overview of the proposed approach.

We have three assumptions for the model. First, the subject loading $Z_{(N\times L)}$ is generated by:

$$Z = \frac{1}{K}\sum_{k=1}^{K} Z^{(k)} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{X}^{(\mathbf{k})}(H^{(k)}W^{(k)})', \tag{3}$$

This shared subject loading is a weighted average of subject loadings per modality, in analogy to the original FLICA model (Groves et al., 2011). This quantity is useful when we want to estimate the contribution of different modalities to the final prediction.

Second, the spatial loadings $W^{(k)}$ are approximately row-wise uncorrelated. We used a reconstruction loss to achieve this, which has been used in reconstruction independent component analysis (Le et al., 2011):

$$\min_{W,H} \sum_{k=1}^{K} \|\mathbf{X^{(k)}} - Z(H^{(k)}W^{(k)})\|_2^2 \tag{4}$$

Note that the transpose of $H^{(k)}W^{(k)}$ in Eq. (3) - due the soft orthogonality constraint for spatial loadings - will approximate the matrix inverse (Le et al., 2011). For example, when we have a single modality, the loss becomes $\min_{W,H} \|X - XWHH'W'\|_2^2$, which means $WHH'W'$ will approximate the identity matrix. A similar property holds when we have K modalities.

Third, we assume sparsity in both spatial loadings and prediction weights - this we enforce through L_1 regularisation. The orthogonal and sparsity constraints on spatial loadings will drive the model to find spatial sparse and non-Gaussian sources, similar to independent component analysis.

We combine the above model assumptions by means of the following objective function:

$$\min_{W,H,B} \underbrace{\sum_{k=1}^{K} (\lambda_1^{(k)} \|\mathbf{X^{(k)}} - Z(H^{(k)}W^{(k)})\|_2^2 + \lambda_2^{(k)}|W^{(k)}|)}_{\text{data reconstruction loss}} + \underbrace{\sum_{i=1}^{Q} (\lambda_3^{(i)} \|\mathbf{Y_i} - ZB_i\|_2^2 + \lambda_4^{(i)}|B_i|)}_{\text{nIDP prediction loss}} \tag{5}$$

Where Y_i is the i-th column of Y, and B_i is the i-th column of B.

2.2. *SuperBigFLICA in a Bayesian framework.*

Identification of the relative weighting parameters between the reconstruction loss, the prediction loss and the sparsity loss (i.e., the λ parameters) through cross-validation is prohibitively expensive. Instead, we take a Bayesian perspective to tune the parameters. We assume normally distributed residual errors for the reconstruction and prediction terms and place Laplacian priors on the spatial loadings and prediction weights. Consequently, the λ parameters above will automatically become parameters in the distributions that can be jointly optimised with other model parameters.

For each imaging modality $\mathbf{X^{(k)}}$, the probabilistic model for the data reconstruction part is:

$$P(\mathbf{X^{(k)}}|Z, W^{(k)}, H^{(k)}, \sigma^{(k)}) = N(ZH^{(k)}W^{(k)}, (\sigma^{(k)})^2 I) \tag{6}$$

Where $\sigma^{(k)}$ is the modality-specific noise term (where we have assumed the noise is shared across voxels as in FLICA(Groves et al., 2011)). We place a Laplacian prior on each element of spatial loadings (Park and Casella, 2008):

$$P(W^{(k)}|b^{(k)}) = \frac{1}{2b^{(k)}} \exp\left(-\frac{|W^{(k)}|}{b^{(k)}}\right) \tag{7}$$

Further, we place a Gamma prior on each $x = (\sigma^{(k)})^2$ as $P(x|\alpha_1, \beta_1) = \frac{\alpha_1^{\beta_1} x^{\alpha_1 - 1} e^{-\beta_1 x}}{\Gamma(\alpha_1)}$, and a non-informative scale-invariant marginal prior on $x = (b^{(k)})^2$ as $P(x) = 1/x$.

The probabilistic model for the prediction part is:

$$P(Y_i|Z, B_i, \gamma_i^2) = \mathcal{N}(ZB_i, \gamma_i^2 I), \tag{8}$$

where γ_i is a task-specific noise term. We also place a Laplacian prior on prediction weights B_i (Park and Casella, 2008):

$$P(B_i|c_i) = \frac{1}{2c_i} \exp\left(-\frac{|B_i|}{c_i}\right), \tag{9}$$

and place a Gamma prior on each $x = (\gamma^{(k)})^2$ as $P(x|\alpha_2, \beta_2) = \frac{\alpha_2^{\beta_2} x^{\alpha_2 - 1} e^{-\beta_2 x}}{\Gamma(\alpha_2)}$, together with a non-informative scale-invariant marginal prior on $x = c_i^2$ as $P(x) = 1/x$.

The posterior distribution of model parameters $\theta = (W^{(k)}, H^{(k)}, B_i, \sigma^{(k)}, b^{(k)}, \gamma_i, c_i), k = 1, \ldots, K, i = 1, \ldots, Q$, given the data $D = (\mathbf{X^{(1)}}, \ldots, \mathbf{X^{(K)}}, \mathbf{Y})$ then becomes:

$$\log P(\theta|D) \propto \log P(D|\theta)P(\theta) = \log P(\mathbf{X^{(1)}}, \ldots, \mathbf{X^{(K)}}|\theta)P(\mathbf{Y}|Z, \theta)P(\theta) \tag{10}$$

Note that the “auto” weights among imaging and nIDPs are proportional to the inverse of the residual prediction variance. But these two types of data usually have completely different properties, so that we don’t want to treat them equally. We therefore add one tuning parameter $\lambda \in [0,1]$ to balance the weights between reconstruction and prediction losses. Tuning this λ is shown to be useful in different kinds of prediction tasks in our experiments later. Thus, we get a modified posterior to be maximized:

$$\begin{aligned}
&\lambda \sum_{k=1}^{K} \left(\log P(\mathbf{X^{(k)}}|\theta) + \log P(\sigma^{(k)}, W^{(k)}, b^{(k)})\right) + (1-\lambda) \sum_{i=1}^{Q} \left(\log P(\mathbf{Y_i}|Z,\theta) + \log P(\gamma_i, B_i, c_i)\right) \\
&= \lambda \underbrace{\sum_{k=1}^{K} \left[\left(-\frac{1}{2(\sigma^{(k)})^2}\|\mathbf{X^{(k)}} - ZH^{(k)}W^{(k)}\|_2^2 + (2\alpha_1 - 3)\log(\sigma^{(k)}) - \beta_1(\sigma^{(k)})^2\right) + \left(-2\log(b^{(k)}) - \frac{1}{b^{(k)}}|W^{(k)}|\right)\right]}_{\text{data reconstruction loss}} \qquad (11) \\
&+ (1-\lambda) \underbrace{\sum_{i=1}^{Q} \left[\left(-\frac{1}{2(\gamma_i)^2}\|\mathbf{Y_i} - ZB_i\|_2^2 + (2\alpha_2 - 3)\log(\gamma_i) - \beta_2(\gamma_i)^2\right) + \left(-2\log(c_i) - \frac{1}{c_i}|B_i|\right)\right]}_{\text{nIDP prediction loss}} + \text{const}
\end{aligned}$$

We can appreciate from Eq. (11) that, the λs in Eq. (5) have been replaced by learnable parameters in the prior distribution of spatial weights and prediction weights (e.g., Gaussian and Laplacian priors), and these learnable parameters have their priors (i.e., Gamma or non-informative priors). The weights among modalities and nIDPs are proportional to the inverse of the residual prediction variance. This is analogous to a Bayesian linear regression with unknown residual variance (Bishop, 2006). The motivation is that the tasks (e.g., prediction or reconstruction) with larger error/uncertainty will be given lower weights (Kendall et al., 2018). We can also replace the Laplacian prior with other sparsity priors to achieve an equivalent sparsity effect. Such alternative priors include the automatic relevance determination (ARD) prior(Wipf and Nagarajan, 2008), the spike-and-slab prior (Mitchell and Beauchamp, 1988), and the Gaussian mixture model prior (used in our original FLICA work (Groves et al., 2011)).

We may later be interested in the contribution of a modality within a latent component to prediction of a specific nIDP. This can be estimated by the correlation between a column of $Z^{(k)}$ in Eq. (3) and an nIDP. This is because $Z^{(k)}$ is the subject course generated by the k-th modality, which has been used to predict an nIDP linearly.

2.3. *Model parameter optimization.*

The SuperBigFLICA model is implemented using Pytorch which can be easily run on a CPU and can be adapted to GPU usage for a more efficient model training. We obtain the maximum-a-posterior (MAP) solution of all parameters using a standard mini-batch stochastic gradient-descent (SGD) algorithms. Here, we have used the Adam optimizer (Kingma and Ba, 2014) for parameters $W^{(k)}, H^{(k)}, B_i$, and the RMSprop optimizer (Tieleman and Hinton, 2012) for parameters $\sigma^{(k)}, b^{(k)}, \gamma_i, c_i$, owing to empirical performance. The first order gradient-based algorithms were used because of their computational efficiency and low memory requirement suitable for optimising high-dimensional parameter space. Below, we evaluate the proposed combined optimisers with other standard first-order methods such as SGD with momentum (Sutskever et al., 2013), Adam or RMSprop, and a quasi-Newton methods L-BFGS (Liu and Nocedal, 1989). We fixed the mini-batch size to 512 subjects and chose the optimal learning rate from 0.0001, 0.001, 0.01, and the tuning parameters λ from $1E-5$ to 0.99999. Dropout regularization with $p = 0.2$ is used on input modalities $X^{(k)}$ and subject loading Z to decrease the chance of overfitting (Srivastava et al., 2014). Batch normalisation is used on Z in the training stage (Ioffe and Szegedy, 2015). The total number of epochs (number of times the full data passes through the model) is 50, the learning rate decreases by 1/2 every ten epochs. The model weights are initialised by Gaussian-distributed random numbers of mean 0 and variance 1.

3. Experiments

3.1. *Brain imaging data.*

Voxel-wise neuroimaging data of 47 modalities from 39,770 subjects were used in this paper, including: (1) 25 "modalities" from the resting-state fMRI ICA dual-regression spatial maps after Z-score normalisation (Miller et al., 2016) ; (2) 6 modalities from the emotion task fMRI experiment: 3 contrast (shapes, faces, faces>shapes) of Z-statistics and 3 *contrasts of parameter estimate* maps (Miller et al., 2016) that reflect %BOLD signal change; (3) 10 diffusion MRI derived modalities (9 TBSS features, including FA, MD, MO, L1, L2, L3, OD, ICVF, ISOVF (Smith et al., 2006; Zhang et al., 2012b) and a summed tractography map of 27 tracts from AutoPtx in FSL (De Groot et al., 2013)); (4) 4 T1-MRI derived modalities (grey matter volume and Jacobian deformation map (which shows expansion/contraction generated by the nonlinear warp to standard space, and hence reflects local volume) in the volumetric space, and cortical area and thickness in the Freesurfer's fsaverage surface space; (5) 1 susceptibility-weighted MRI map (T2-star); (6) 1 T2-FLAIR MRI derived modality (white matter hyperintensity map estimated by BIANCA (Griffanti et al., 2016)). The UK Biobank imaging data were mainly

Table 1: Non-imaging derived phenotypes used in this study.

Non-imaging derived phenotypes
Fluid intelligence
Hypertension
Handedness
Treatment/medication code (1140884600 - metformin)
Diabetes diagnosed by doctor
Non-cancer illness code, self-reported (1261 - multiple sclerosis)
Overall health rating
Age started wearing glasses or contact lenses
Number of treatments/medications taken
Mean time to correctly identify matches
Maximum digits remembered correctly
Number of self-reported non-cancer illnesses

preprocessed by FSL (Smith et al., 2004; Jenkinson et al., 2012) and FreeSurfer (Fischl, 2012) following an optimized processing pipeline(Alfaro-Almagro et al., 2018) (`https://www.fmrib.ox.ac.uk/ukbiobank/`). From the voxel-wise modality maps, we generate 3,913 IDPs, including global and local features from the six imaging modalities (T1, T2-FLAIR, swMRI, task fMRI, resting-state fMRI, and diffusion MRI) (Smith et al., 2020) further specified in the **supplementary file**.

*3.2. **Non-imaging derived phenotypes.***

A total of 17,485 non-imaging derived phenotypes (nIDPs) were used in this paper. We mainly analyzed 12 of them in the 'physical', 'cognition', and 'health outcome' domains, summarised in **Table 1**. These nIDPs were selected to allow for a direct comparison to our previous work, as they were the best predicted nIDPs in cognition and health outcome domains by our baseline approach BigFLICA (Gong et al., 2021). The direct comparison of performance between SuperBigFLICA and BigFLICA approaches enables us to study the benefits of including the supervised loss terms.

*3.3. **Confounding variables and missing values.***

Before carrying out nIDP prediction, a total of 597 confounding variables were regressed out from both voxelwise imaging data and nIDPs, using linear regression (Alfaro-Almagro et al., 2021). Missing modality data for a subject were imputed by the mean map of all other subjects. We did not impute missing nIDPs, and therefore, in the SuperBigFLICA model, only data-reconstruction-related losses play a role for subjects with missing nIDP data.

*3.4. **Imaging data pre-reduction using dictionary learning.***

There are tens of thousands of voxels per modality, so a direct fitting of SuperBigFLICA using voxel-level data is computationally expensive and memory-consuming. Although we can perform mini-batch optimisation on the subjects, we need to keep K big voxel-by-component spatial maps as learnable parameters in memory. One possible solution is to use sparse dictionary learning for voxel space dimension reduction before running SuperBigFLICA. As shown in our previous work (Gong et al., 2021), for BigFLICA, sparse dictionary learning will reduce the computation load of the optimisation and may increase the modality-specific signal-to-noise ratio. This is because the overall model is linear, so that a pre-dimension reduction using a (linear) dictionary learning should not interfere with important information that can be captured by BigFLICA, but will potentially have the de-noising effect. Owing to the similar property of the models employed in BigFLICA and SuperBigFLICA, we expect sparse dictionary learning to perform similarly well. Here, we evaluate the effect of data reduction on the final prediction across the voxel-domain between 100 and 2,000 dictionary features per modality before running any subsequent algorithm, e.g., BigFLICA or SuperBigFLICA. Note that this pre-reduction can also be performed with nIDP information included. We therefore also tested applying a "single-modality" SuperBigFLICA to each single modality map (which is just a special case with $K = 1$) with all 17,485 nIDPs as supervision before using SuperBigFLICA for multimodal analysis.

*3.5. **Baseline: nIDP prediciton using hand-curated IDPs and modes of BigFLICA.***

In real data, we rely on the performance of predicting nIDPs as a surrogate criterion to evaluate different methods, given that "ground-truth" IDPs do not in general exist. As a baseline approach, we compare hand-curated IDPs and modes of BigFLICA. The pipeline for prediction follows our previous work (Gong et al., 2021). In brief, 3,913 IDPs and 1,000 modes of BigFLICA are extracted from UK Biobank imaging data. BigFLICA used a 3,500-dimensional MIGP approach (Gong et al., 2021) and a 2,000-dimensional dictionary learning approach as data preprocessing (Gong et al., 2021) before running the core FLICA variational Bayesian optimization.

Here, a high dimensional decomposition was chosen as in our previous work on BigFLICA, which achieved the best prediction accuracy for most nIDPs (Gong et al., 2021). Further, elastic-net regression, from the `glmnet` package (Zou and Hastie, 2005), was used to predict each of the 12 nIDPs separately (known as single-task learning) using IDPs or FLICA subject modes as model regressors (features). This approach is widely used and has been shown to achieve robust and state-of-the-art performance in many neuroimaging studies(Cui and Gong, 2018; Jollans et al., 2019).

We randomly selected a subset of 25,000 subjects for model training, while the validation set contains 5,000 different subjects, and the test set was formed from the remaining 9,770 subjects. All methods' comparisons are using the same train-test split. For single-task learning, the prediction accuracy was quantified as the Pearson correlation between predicted and the true values of each nIDP in the test sets. For multi-task learning, the prediction accuracy was quantified as the sum of correlations with nIDPs larger than 0.1.

3.6. *Predicting nIDPs using single-task and multi-task SuperBigFLICA.*

In order to demonstrate how SuperBigFLICA with one or more target nIDPs in training can boost the performance compared to hand-curated IDPs and unsupervised BigFLICA, first, we trained single-task SuperBigFLICA with each of the 12 nIDPs as a supervision target. We then trained multi-task SuperBigFLICA by including each target and the top 24 or 99 most highly correlated (with the target) nIDPs from all 17,485 available nIDPs from the UK Biobank dataset, in the training stage. In previous work, it was already established that the inclusion of correlated tasks as targets could boost the performance of single-task learning (Zhang and Yang, 2017; Rahim et al., 2017). Therefore, we performed 25- or 100-dimensional multi-task learning (separately) for each of the 12 nIDPs and evaluated the prediction performance. Note that while an additional 24 (or 99) nIDPs (that are correlated to the target nIDP) are used to help in the training, they are not used in test data to help the prediction - only the imaging data from test subjects is used in predictions of nIDPs in test subjects. Finally, we trained SuperBigFLICA with all 17,485 nIDPs as supervision targets. For the tuning parameters, the number of latent components was chosen to be 25, 100, 250, 500, or 1,000, and the λ parameter is chosen from $1E-5$ to 0.99999. We evaluated the influence of λ, and different random model parameter initialisations on the final prediction performance.

3.7. *Evaluation of the generalisability of SuperBigFLICA modes on unseen nIDPs.*

One of the fundamental goals of phenotype discovery is to learn a low-dimensional latent space that is generalisable in that it can predict "unseen" nIDPs. In the above analyses, the nIDP to be predicted was always included in the training stage. Here, we evaluated whether SuperBigFLICA can generate a good representation for entirely new tasks, wherein in the training stage, we only use nIDPs that are not our targets. To do this, for a given nIDP that we want to predict (e.g., fluid intelligence), we first compute the correlations between this target nIDP and all other 17,485 nIDPs. We selectively include the 16,485 least correlated nIDPs for training SuperBigFLICA, ignoring the 1,000 most strongly covarying nIDPs in training. This means that, for example, the mean correlation of 16,485 nIDPs with the target variable fluid intelligence is 0.007, and nIDPs with correlation > 0.032 with fluid intelligence are removed in training. This experiment simulates a "bad" situation where almost no nIDPs are related to our target when generating the latent space. We then used elastic-net regression to predict our target nIDP using the resulting latent space, in order to evaluate the degree to which the inclusion of unrelated nIDPs constraints the learning towards IDP features that are more generally useful for prediction across a wide range of nIDPs, and thereby ultimately also improves prediction for specific nIDPs of interest. This may also relate to the concept of transfer learning, where we can use multiple nIDPs to learn a space that generically has high predictive power.

4. Results

4.1. *Comparing SuperBigFLICA with hand-curated IDPs and modes of unsupervised BigFLICA.*

We first compared SuperBigFLICA with hand-curated IDPs and BigFLICA, and then compared different variants of SuperBigFLICA, in terms of prediction accuracy of nIDPs (Method sections 3.5, 3.6 and 3.7). Each subfigure of **Fig.2** shows the overall prediction accuracy of different experimental approaches for 12 nIDPs. We report the results that use the dictionary dimension of 250 for each modality (detailed in the next section). The test accuracy is obtained using the best tuning parameters (λ and number of latent components) selected in the validation set.

The first two columns are the accuracy of elastic-net regression with hand-curated IDPs and modes of unsupervised BigFLICA as input features, while the third column is the accuracy of single-task SuperBigFLICA trained end-to-end. We can see that in almost all cases, the accuracy of semi-supervised training outperforms hand-curated IDPs and unsupervised BigFLICA. The average percent improvement of single-task SuperBigFLICA compared with hand-curated IDPs and BigFLICA are 46% and 25%.

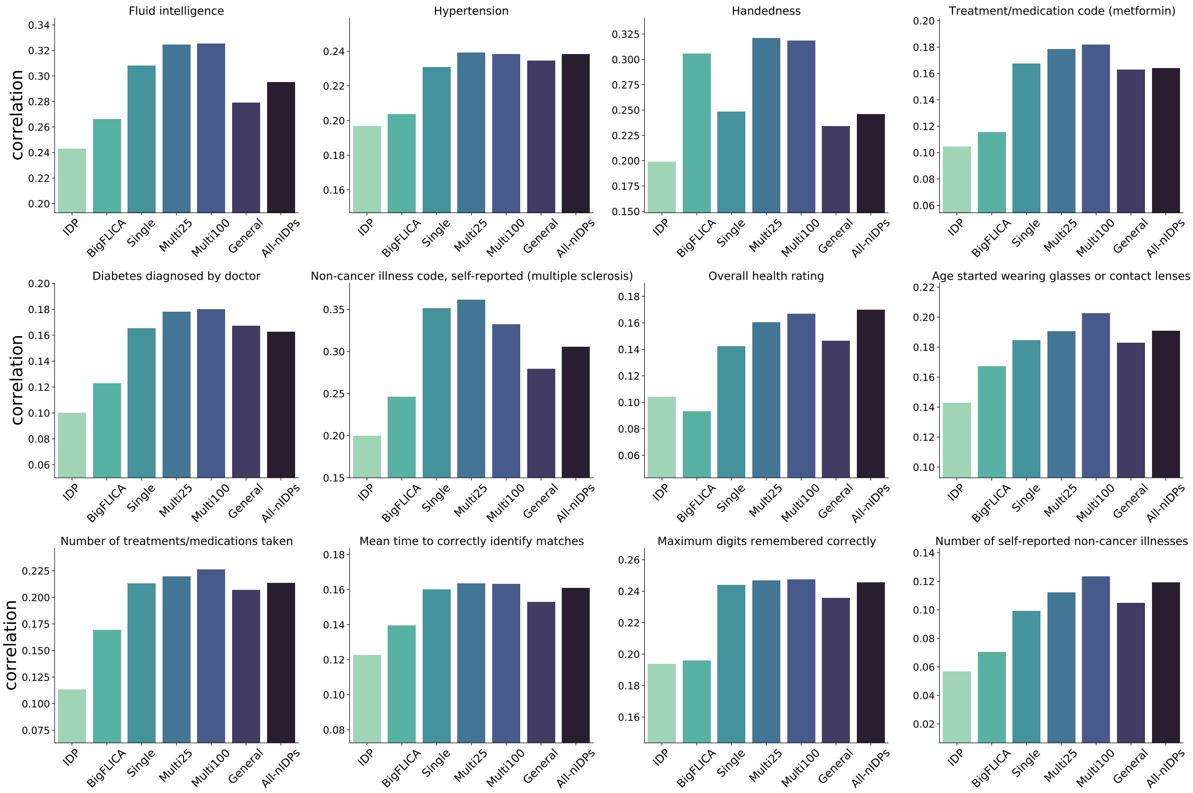

Figure 2: Comparison of SuperBigFLICA with hand-curated IDPs and modes of unsupervised BigFLICA for the 12 nIDPs listed in Table 1. Each figure shows - for a different nIDP - the predictive correlation of different approaches and different parameter settings. The first and second column shows the 'baseline' prediction accuracy of IDPs and BigFLICA. The third column shows the accuracy of single-task SuperBigFLICA trained end-to-end. The fourth and fifth columns show prediction accuracy of multi-task SuperBigFLICA, with 24 and 99 most correlated nIDPs as auxiliary tasks for supporting the training of the main nIDP of interest. The sixth column shows the generalisability accuracy of SuperBigFLICA modes on unseen nIDPs, where the main nIDP of interest is not included in the learning of latent space but only 16485 nIDPs that are least correlated with it. The seventh column shows the prediction accuracy when we fuse all 47 modalities and use all 17,485 nIDPs to train a single model.

The fourth and fifth columns of each subfigure of **Fig.2** show prediction accuracy of multi-task SuperBigFLICA, with 24 and 99 most correlated nIDPs as auxiliary tasks for supporting the training of the main nIDP of interests. We can see that multi-task learning usually further improves the prediction accuracy compared with single-task SuperBigFLICA. The average per cent improvement of two multi-task BigFLICA compared with single-task SuperBigFLICA are 9% and 7%.

The sixth column of each subfigure of **Fig.2** shows the generalisability of SuperBigFLICA modes on unseen nIDPs, where the main nIDP of interests is not included in learning the latent space (only nIDPs that are at best weakly correlated with the main nIDP of interests are involved). We then used elastic-net regression to predict the main nIDP based on the learned latent space as regressors. Overall, this is expected to perform worse than single-task and multi-task SuperBigFLICA because the target nIDP is not involved in the training. However, it still outperforms unsupervised BigFLICA plus elastic-net by 21%, and is slightly worse than single-task SuperBigFLICA. This experiment demonstrated that the inclusion of even irrelevant tasks in the "supervised" training could boost the predictive performance of the generated latent space.

The seventh column of each subfigure of **Fig.2** shows the prediction accuracy when we fuse all 47 modalities and 17,485 nIDPs to train one multi-task SuperBigFLICA model. We can see that the prediction accuracy is similar to single-task SuperBigFLICA for most of the nIDPs.

4.2. *Analysis of SuperBigFLICA algorithm.*

4.2.1. *Relationship between prediction accuracy and hyper-parameters.*

We evaluated the influence of the relative weighting between reconstruction loss and prediction loss, i.e., λ, and the number of latent components, on the final prediction performance. **Fig.3A** shows the mean prediction correlation of 12 nIDPs listed in **Table** 1 across different λ and latent components using single-task SuperBigFLICA. When the latent dimension is small, we need to choose a small λ (i.e., $\lambda < 0.5$, which means that the prediction loss has a higher weight than the reconstruction loss) to achieve optimal prediction. When

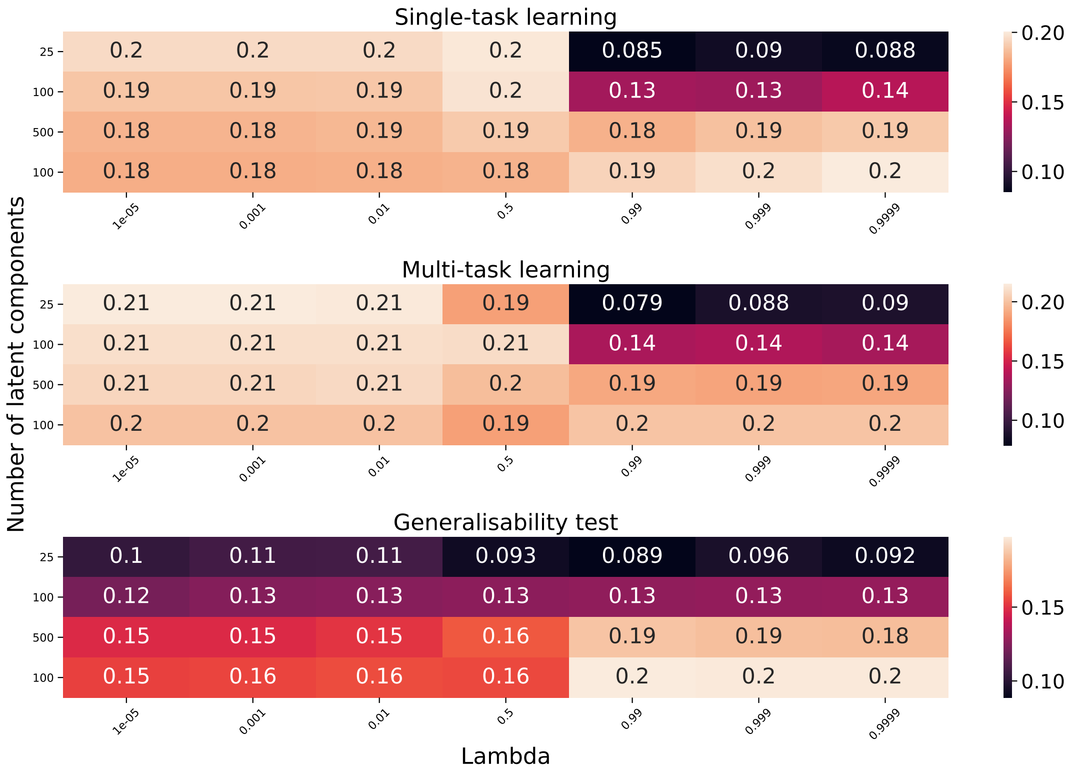

Figure 3: The relationship between the number of latent dimensions (y-axis) and weights between reconstruction and prediction losses (x-axis) with the mean prediction accuracy across 12 nIDPs listed in Table 1. A. Single-task learning setting. **B.** Multi-task learning setting with 24 auxiliary tasks. **C.** generalisability test.

the latent dimension is large, although a smaller λ also achieves the best prediction accuracy, the influence of λ becomes much smaller than using a smaller latent dimension. We can draw a similar conclusion from the results shown in **Fig.3B**, which is the case of multi-task SuperBigFLICA with 24 most correlated nIDPs as auxiliary tasks. Conversely, for the generalisability test, **Fig.3C** shows that the prediction accuracy is low when the latent dimension is small. When the latent dimension is large, the prediction accuracy is highest when $\lambda > 0.5$, i.e., the reconstruction loss has higher weight than the prediction loss. This analysis guides how to select hyperparameters in different circumstances and demonstrates the usefulness of including both data reconstruction and prediction losses in the objective function.

4.2.2. Influence of imaging space dimension reduction on prediction accuracy.

We first tested whether the imaging space pre-dimension reduction with dictionary learning influenced the final prediction accuracy of nIDPs (Method section 3.4). **Fig.4A** shows that, for different nIDPs, the average accuracy of single-task SuperBigFLICA is similar ($0.18 < r < 0.22$) across different dictionary dimensions. The 250-dimensional dictionary learning achieves slightly better performance.

We further evaluated the use of SuperBigFLICA to perform imaging space dimension "pre-reduction" (250-dimension for each modality, the same with dictionary learning), with all 17,485 nIDPs as supervised targets. We did not see an improvement relative to using dictionary learning, e.g., for fluid intelligence, the best prediction correlation is only around $r = 0.20$, much lower than the result achieved by dictionary learning $r = 0.33$. A lower prediction correlation was also observed for other nIDPs and other SuperBigFLICA dimensions. A possible reason is that using all nIDPs in the pre-dimension reduction stage discards information in the imaging data related to the target nIDP. Also, we find that using SuperBigFLICA in this situation is more memory intensive because we need to keep a huge voxel-by-component spatial weight matrix in memory. In contrast, in dictionary learning, we only need to keep a smaller subject-by-component matrix in memory because it performs a mini-batch optimisation on the voxel dimension. Finally, another disadvantage of the two-stage supervised learning strategy is the need for two nested cross-validation loops for parameter selection, significantly increasing the computation cost.

4.2.3. Influence of parameter initialisation on the prediction accuracy.

We evaluated the influence of random model parameter initialisations on the final prediction accuracy. We tested whether two different random initialisations will result in different accuracy. Therefore, we performed

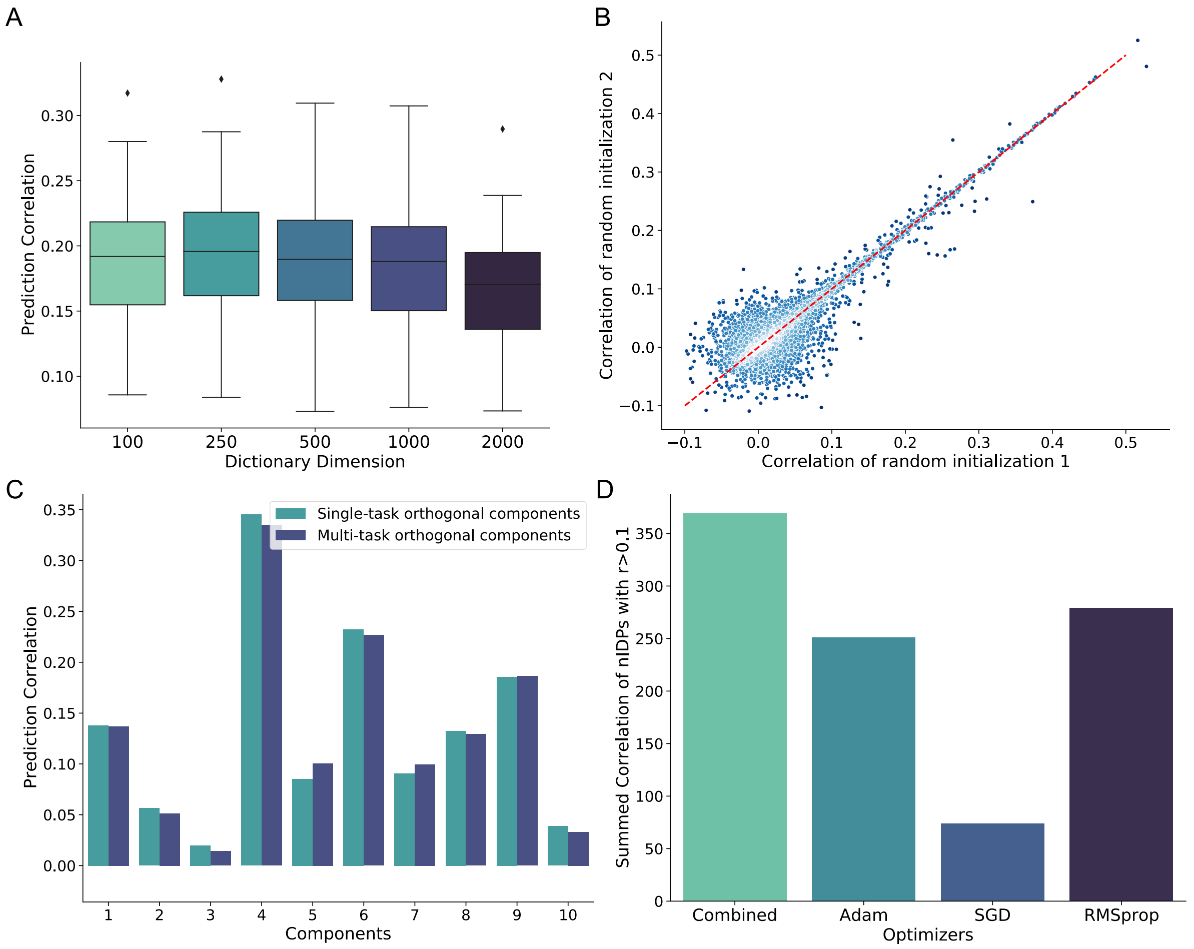

Figure 4: Evaluations of SuperBigFLICA model. A. The average prediction accuracy of 12 nIDPs using different dictionary learning dimensions in the image data pre-reduction stage. **B**. The prediction accuracy of 17,485 nIDPs as a result of different random parameter initializations. **C**. Comparing the prediction accuracy of single-task SuperBigFLICA vs. multi-task SuperBigFLICA for predicting top 10 orthogonal principal components derived from all nIDPs. **D**. Comparing the prediction accuracy of all 17,485 nIDPs of SuperBigFLICA optimized by different numerical optimizers.

two multi-task SuperBigFLICA experiments with all 17,485 nIDPs as targets. The only difference between these two experiments is the different seeds for parameter initialisation. **Fig.4B** shows the scatter plots of prediction correlations of the different nIDPs from two different initialisations. We can see that nIDP correlations < 0.1 result in a roughly spherical point cloud, while correlations > 0.1 are highly correlated, i.e., different initialisations lead to similar nIDP predictions.

4.2.4. Influence of nIDP task covariance structure on the prediction accuracy.

We further evaluated whether the increases in prediction accuracy of multi-task learning compared with single-task learning result from incorporating information about the task covariance structure into the estimation. We tested this hypothesis by predicting uncorrelated (orthogonal) nIDPs either using separate single-task SuperBigFLICA or jointly using multi-task SuperBigFLICA. To obtain these orthogonal nIDPs, we extracted the top 10 principal components from the subject-by-nIDP data matrix. We compared the accuracy when the model predicted them separately as single-task learning and jointly as multi-task learning. The result shows that the performance is almost the same for each of the 10 “orthogonal“ tasks (**Fig.4C**). This experiment shows that the covariance structure between the different nIDPs enables the multi-task model to improve over and above the single-task model.

4.2.5. Influence of optimisation algorithms on the prediction accuracy.

We finally evaluated the choice of numerical optimisation algorithms on the prediction accuracy. We took an example where we used SuperBigFLICA to fuse all 47 modalities and 17,485 nIDPs to discover a 1,000-dimensional latent space. The Adam, SGD, and RMSprop optimisers all perform worse than the combined optimisation algorithm in terms of the overall prediction accuracy of the 17,485 IDPs (**Fig.4D**). The overall accuracy is estimated by the sum of correlation of nIDPs larger than 0.1.

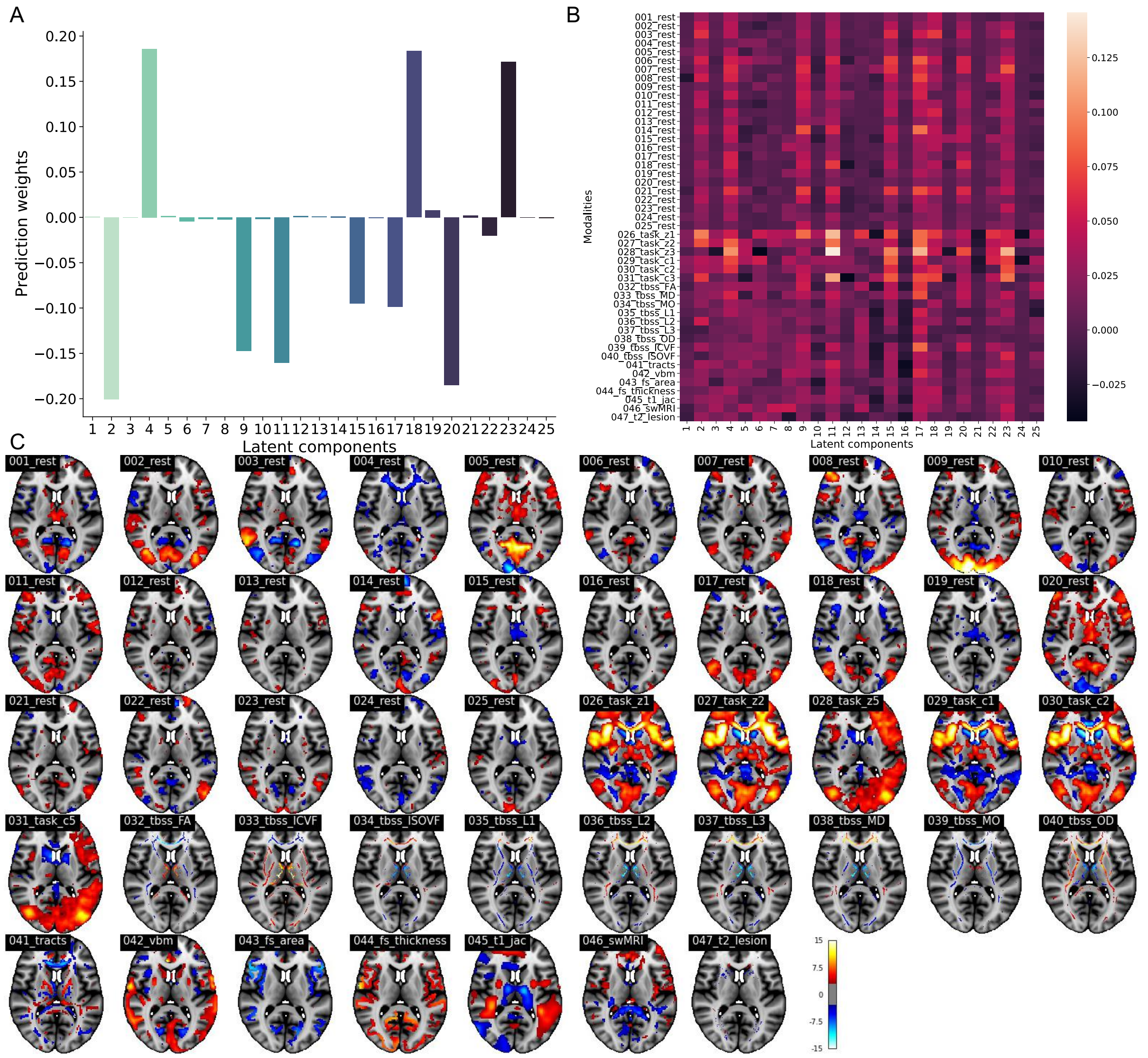

Figure 5: Example results of predicting fluid intelligence with a 25-dimentional SuperBigFLICA. A. The prediction weights of fluid intelligence in a 25-dimensional SuperBigFLICA analysis. **B.** The contribution of different modalities within each of 25 SuperBigFLICA components for predicting fluid intelligence. **C.** The Z-score normalized 47 spatial maps of modes with strongest contribution to the prediction of fluid intelligence in a 25-dimensional SuperBigFLICA decomposition, with fluid intelligence as the supervision target (component 2, MNI152 coordinate z=10).

4.3. ***Real data further qualitative analysis.***

4.3.1. *Fluid intelligence prediction using a low-dimensional SuperBigFLICA.*

We first applied a 25-dimensional SuperBigFLICA to predict fluid intelligence scores using data from 47 modalities. **Fig.5A** shows the weights of each of the 25 latent components on predicting fluid intelligence scores (B in Eq. (2)). Components 2, 4, 9, 11, 15, 17, 18, 20, 23 are selected to contribute to the prediction of fluid intelligence, while the remaining components were switched off by the Laplacian prior to only contribute to the reconstruction of imaging data. The overall prediction correlation is 0.30.

Fig.5B shows the contribution of each component and each modality to the prediction of fluid intelligence (Method section 2.2)). The components selected for fluid intelligence prediction have higher overall contributions, which demonstrated the validity of the proposed way to estimate modality contribution. Across the different modalities, the components from task contrast maps have the highest contribution to the prediction of fluid intelligence, while the resting-state dual-regression spatial maps have a slightly lower contribution. The dMRI and sMRI derived modalities contribute least to the prediction.

Fig. 5C shows the z-score spatial maps of component 2 for each modality, which was generated by regressing latent components back onto the original voxel-wise data. This component contributes most to the prediction of fluid intelligence scores. The task contrast maps and resting-state map 5 have the highest voxel-wise activations. The regions that show the highest contribution to the prediction of fluid intelligence are mainly located in the precuneus cortex, posterior cingulate gyrus, lateral occipital cortex, insular cortex, inferior frontal gyrus, and frontal pole (Amodio and Frith, 2006). Among them, the insular cortex, inferior frontal gyrus, and frontal pole were found significant in task modalities in our previous BigFLICA approach (Gong et al., 2021), but with a much higher 750-dimensional decomposition. This reflects the fact that adding supervision to the model can help the model learn task-specific patterns easier. In addition, here, with the SuperBigFLICA approach, we can observe a more comprehensive 'multimodal' effect, such as the changes in cortical surface area and thickness in these regions, and changes of tracts and white matter microstructures that connect these regions, which are also reported in literature (Menary et al., 2013; Chen et al., 2020). Moreover, we also observe the precuneus cortex and posterior cingulate gyrus in several resting-state maps, as part of the default mode network, involved in fluid intelligence prediction (Santarnecchi et al., 2017).

4.3.2. Phenotype discovery using a high-dimensional SuperBigFLICA.

We finally applied SuperBigFLICA to perform a 1,000-dimensional decomposition with all 17,485 nIDPs as supervision targets. This 1,000-dimensional latent space can serve as a set of new data-driven IDPs.

Fig. A.6 shows the Z-score normalised spatial maps of the component that most strongly contributes to the prediction of *hypertension* scores. The prediction correlation is 0.24. The regions that show the highest contribution to the prediction of hypertension are mainly located in the precuneous cortex, visual cortex, middle temporal gyrus, central opercular cortex, Heschel's gyrus, inferior frontal gyrus and insular cortex, and also external capsule tracts. Again, the modes are more 'multimodal', and several consistent findings have been reported in the literature (Li et al., 2015; Den Heijer et al., 2005; Hannawi et al., 2018).

Likewise, **Fig. A.7** shows the Z-score normalised spatial maps of the mode that contributes most strongly to the prediction of *age started wearing glasses or contact lenses*. The prediction correlation is 0.19. The regions that show the highest contribution to the prediction of hypertension are mainly located in visual areas, especially for resting-state dual regression spatial maps 5, which represents the visual network.

5. Discussion

In this paper, we propose SuperBigFLICA, a semi-supervised multimodal data fusion approach that simultaneously reconstructs the original voxel-wise imaging data and best predicts non-imaging derived phenotypes. The approach is scalable to extreme high-dimensional data sets, e.g., UK Biobank scale neuroimaging datasets. SuperBigFLICA inherits the Bayesian framework from the previous FLICA model (Groves et al., 2011; Gong et al., 2021). Additionally, it incorporates an additional prediction term to enable supervised learning of the target variable of interests (i.e., multiple nIDPs). SuperBigFLICA can discover spatially sparse and orthogonal modes that can serve as generic data-driven IDPs for future prediction of new nIDPs. The weighting of different modalities and nIDPs can be automatically inferred from the data, avoiding manual specification.

Compared to previous linear approaches (e.g., (Qi et al., 2017)), the scalability of our approach to huge data sets is improved through the use of advanced stochastic optimisation algorithms. Our model can use multiple nIDPs as supervision targets and can predict unseen nIDPs. Compared to nonlinear approaches (e.g., (Zhang et al., 2012a; Zhou et al., 2020; Liu et al., 2020)), our approach can explicitly discover a low-dimensional linear latent space as new image-derived phenotypes. We performed a comprehensive comparison of SuperBigFLICA with the hand-curated IDPs currently being created by our group on behalf of UK Biobank, and modes of unsupervised BigFLICA, and found a significantly improved performance on predicting nIDPs. We also showed that by using the multi-task learning paradigm, SuperBigFLICA showed a further improvement than its single-task setting. We demonstrated SuperBigFLICA's performance in learning a generalisable latent space by applying it to predict unseen nIDPs. These tests were performed using the largest neuroimaging dataset to date (UK Biobank), with 47 different modalities, 39,770 subjects, and 17,485 nIDPs, which illustrates the ability of SuperBigFLICA for analysing large-scale datasets. In real data examples, we demonstrated that SuperBigFLICA finds interpretable modes predictive of health outcome and cognitive nIDPs.

There are multiple future directions for improving the current approach. First, we could further explore the possibility of improving prediction of unseen nIDPs by using advanced techniques in transfer learning (Pan and Yang, 2009). Second, a deeper understanding of the latent space, including the interpretation of spatial maps and the influence of dimensionality of latent space with prediction power, could be interesting. Third, another straightforward extension would be adding nonlinearity to SuperBigFLICA, which enables it to extract more complex nonlinear patterns from brain imaging data, with or without the supervision of nIDPs. Many options exist to achieve this by using either deep neural networks (Goodfellow et al., 2016) or traditional machine learning approaches such as Gaussian process latent variable model (Lawrence, 2005) and multiple kernel learning (Gönen and Alpaydın, 2011). Nonlinear approaches such as deep convolutional neural networks have shown excellent age and sex prediction accuracy using structural MRI data (Peng et al., 2019) and in Alzheimer disease progression (Nguyen et al., 2020), but the usefulness of nonlinear models for neuroimaging data is still under debate (Schulz et al., 2020; He et al., 2020; Abrol et al., 2020) due to the increased complexity of evaluating and interpreting their performance. Therefore, besides developing a nonlinear model for improving predictive performance, deriving interpretable nonlinear features is also an important task.

In summary, all of the above will be explored in future improvements to our analysis approach. An easy-to-use version of this software will be integrated into an upcoming version of the FSL software package (Smith et al., 2004; Jenkinson et al., 2012). Results from applying SuperBigFLICA on UK Biobank will also be released via the UKB database as new data-driven IDPs (image features), further contributing to the richness of the sample and enabling neuroscientific research.

Acknowledgments.

We are grateful to UK Biobank and its participants (access application 8107). We would like to thank Dr. Alberto Llera for providing the code of the original FLICA algorithm. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. Financial support was provided by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z.

Competing interests.

The authors declare that they have no competing financial interests.

Data availability.

For UK Biobank, all source data (including raw and processed brain imaging data, derived IDPs, and non-imaging measures) is available from UK Biobank via their standard data access procedure (see http://www.ukbiobank.ac.uk/register-apply).

Code availability.

SuperBigFLICA code is available at https://github.com/weikanggong/SuperBigFLICA, and will also be released as part of an upcoming version of FSL.

References

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., and Calhoun, V. (2020). Hype versus hope: Deep learning encodes more predictive and robust brain imaging representations than standard machine learning. *bioRxiv*.

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160.

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

Alfaro-Almagro, F., McCarthy, P., Afyouni, S., Andersson, J. L., Bastiani, M., Miller, K. L., Nichols, T. E., and Smith, S. M. (2021). Confound modelling in uk biobank brain imaging. *NeuroImage*, 224:117002.

Amodio, D. M. and Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature reviews neuroscience*, 7(4):268–277.

Ball, G., Malpas, C. B., Genc, S., Efron, D., Sciberras, E., Anderson, V., Nicholson, J. M., and Silk, T. J. (2019). Multimodal structural neuroimaging markers of brain development and adhd symptoms. *American Journal of Psychiatry*, 176(1):57–66.

Beckmann, C. F. and Smith, S. M. (2004). Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE transactions on medical imaging*, 23(2):137–152.

Bishop, C. M. (2006). Pattern recognition and machine learning. *Journal of Electronic Imaging*, 16:049901.

Calhoun, V. D. and Sui, J. (2016). Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness. *Biological psychiatry: cognitive neuroscience and neuroimaging*, 1(3):230–244.

Chen, P.-Y., Chen, C.-L., Hsu, Y.-C., Tseng, W.-Y. I., et al. (2020). Fluid intelligence is associated with cortical volume and white matter tract integrity within multiple-demand system across adult lifespan. *NeuroImage*, 212:116576.

Cui, Z. and Gong, G. (2018). The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *Neuroimage*, 178:622–637.

Dadi, K., Varoquaux, G., Machlouzarides-Shalit, A., Gorgolewski, K. J., Wassermann, D., Thirion, B., and Mensch, A. (2020). Fine-grain atlases of functional modes for fmri analysis. *arXiv preprint arXiv:2003.05405*.

De Groot, M., Vernooij, M. W., Klein, S., Ikram, M. A., Vos, F. M., Smith, S. M., Niessen, W. J., and Andersson, J. L. (2013). Improving alignment in tract-based spatial statistics: evaluation and optimization of image registration. *Neuroimage*, 76:400–411.

Den Heijer, T., Launer, L., Prins, N., Van Dijk, E., Vermeer, S., Hofman, A., Koudstaal, P., and Breteler, M. (2005). Association between blood pressure, white matter lesions, and atrophy of the medial temporal lobe. *Neurology*, 64(2):263–267.

Douaud, G., Groves, A. R., Tamnes, C. K., Westlye, L. T., Duff, E. P., Engvig, A., Walhovd, K. B., James, A., Gass, A., Monsch, A. U., et al. (2014). A common brain network links development, aging, and vulnerability to disease. *Proceedings of the National Academy of Sciences*, 111(49):17648–17653.

Eickhoff, S. B., Yeo, B. T., and Genon, S. (2018). Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, 19(11):672–686.

Elliott, L. T., Sharp, K., Alfaro-Almagro, F., Shi, S., Miller, K. L., Douaud, G., Marchini, J., and Smith, S. M. (2018). Genome-wide association studies of brain imaging phenotypes in UK Biobank. *Nature*, 562(7726):210.

Fischl, B. (2012). Freesurfer. *Neuroimage*, 62(2):774–781.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Gönen, M. and Alpaydın, E. (2011). Multiple kernel learning algorithms. *The Journal of Machine Learning Research*, 12:2211–2268.

Gong, W., Beckmann, C. F., and Smith, S. M. (2021). Phenotype discovery from population brain imaging. *Medical Image Analysis*, page 102050.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U. G., Kuker, W., Battaglini, M., Rothwell, P. M., et al. (2016). BIANCA (brain intensity abnormality classification algorithm): a new tool for automated segmentation of white matter hyperintensities. *NeuroImage*, 141:191–205.

Groves, A. R., Beckmann, C. F., Smith, S. M., and Woolrich, M. W. (2011). Linked independent component analysis for multimodal data fusion. *Neuroimage*, 54(3):2198–2217.

Hannawi, Y., Yanek, L. R., Kral, B. G., Vaidya, D., Becker, L. C., Becker, D. M., and Nyquist, P. A. (2018). Hypertension is associated with white matter disruption in apparently healthy middle-aged individuals. *American Journal of Neuroradiology*, 39(12):2243–2248.

He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., and Yeo, B. T. (2020). Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. *NeuroImage*, 206:116276.

Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.

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

Jollans, L., Boyle, R., Artiges, E., Banaschewski, T., Desrivières, S., Grigis, A., Martinot, J.-L., Paus, T., Smolka, M. N., Walter, H., et al. (2019). Quantifying performance of machine learning methods for neuroimaging data. *NeuroImage*.

Kendall, A., Gal, Y., and Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

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

Lawrence, N. (2005). Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of machine learning research*, 6(Nov):1783–1816.

Le, Q. V., Karpenko, A., Ngiam, J., and Ng, A. Y. (2011). Ica with reconstruction cost for efficient overcomplete feature learning. In *Advances in neural information processing systems*, pages 1017–1025.

Lee, G., Nho, K., Kang, B., Sohn, K.-A., and Kim, D. (2019). Predicting alzheimer's disease progression using multi-modal deep learning approach. *Scientific reports*, 9(1):1–12.

Li, X., Liang, Y., Chen, Y., Zhang, J., Wei, D., Chen, K., Shu, N., Reiman, E. M., and Zhang, Z. (2015). Disrupted frontoparietal network mediates white matter structure dysfunction associated with cognitive decline in hypertension patients. *Journal of Neuroscience*, 35(27):10015–10024.

Liu, D. C. and Nocedal, J. (1989). On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528.

Liu, Y., Fan, L., Zhang, C., Zhou, T., Xiao, Z., Geng, L., and Shen, D. (2020). Incomplete multi-modal representation learning for alzheimer's disease diagnosis. *Medical Image Analysis*, page 101953.

Lu, D., Popuri, K., Ding, G. W., Balachandar, R., Beg, M. F., Initiative, A. D. N., et al. (2018). Multiscale deep neural network based analysis of fdg-pet images for the early diagnosis of alzheimer's disease. *Medical image analysis*, 46:26–34.

Marquand, A. F., Brammer, M., Williams, S. C., and Doyle, O. M. (2014). Bayesian multi-task learning for decoding multi-subject neuroimaging data. *NeuroImage*, 92:298–311.

Menary, K., Collins, P. F., Porter, J. N., Muetzel, R., Olson, E. A., Kumar, V., Steinbach, M., Lim, K. O., and Luciana, M. (2013). Associations between cortical thickness and general intelligence in children, adolescents and young adults. *Intelligence*, 41(5):597–606.

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

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Nguyen, M., He, T., An, L., Alexander, D. C., Feng, J., Yeo, B. T., Initiative, A. D. N., et al. (2020). Predicting alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, 222:117203.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A., and Smith, S. M. (2019). Accurate brain age prediction with lightweight deep neural networks. *BioRxiv*.

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., and Nichols, T. E. (2011). *Statistical parametric mapping: the analysis of functional brain images*. Elsevier.

Qi, S., Calhoun, V. D., van Erp, T. G., Bustillo, J., Damaraju, E., Turner, J. A., Du, Y., Yang, J., Chen, J., Yu, Q., et al. (2017). Multimodal fusion with reference: searching for joint neuromarkers of working memory deficits in schizophrenia. *IEEE transactions on medical imaging*, 37(1):93–105.

Qi, S., Yang, X., Zhao, L., Calhoun, V. D., Perrone-Bizzozero, N., Liu, S., Jiang, R., Jiang, T., Sui, J., and Ma, X. (2018). Microrna132 associated multimodal neuroimaging patterns in unmedicated major depressive disorder. *Brain*, 141(3):916–926.

Rahim, M., Thirion, B., Bzdok, D., Buvat, I., and Varoquaux, G. (2017). Joint prediction of multiple scores captures better individual traits from brain images. *Neuroimage*, 158:145–154.

Santarnecchi, E., Emmendorfer, A., Tadayon, S., Rossi, S., Rossi, A., and Pascual-Leone, A. (2017). Network connectivity correlates of variability in fluid intelligence performance. *Intelligence*, 65:35–47.

Schulz, M.-A., Yeo, B. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., Richards, B., and Bzdok, D. (2020). Different scaling of linear models and deep learning in ukbiobank brain images versus machine-learning datasets. *Nature communications*, 11(1):1–15.

Smith, S. M., Elliott, L. T., Alfaro-Almagro, F., McCarthy, P., Nichols, T. E., Douaud, G., and Miller, K. L. (2020). Brain aging comprises many modes of structural and functional change with distinct genetic and biophysical associations. *Elife*, 9:e52677.

Smith, S. M., Jenkinson, M., Johansen-Berg, H., Rueckert, D., Nichols, T. E., Mackay, C. E., Watkins, K. E., Ciccarelli, O., Cader, M. Z., Matthews, P. M., et al. (2006). Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage*, 31(4):1487–1505.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., et al. (2004). Advances in functional and structural mr image analysis and implementation as FSL. *Neuroimage*, 23:S208–S219.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Sui, J., Qi, S., van Erp, T. G., Bustillo, J., Jiang, R., Lin, D., Turner, J. A., Damaraju, E., Mayer, A. R., Cui, Y., et al. (2018). Multimodal neuromarkers in schizophrenia via cognition-guided mri fusion. *Nature communications*, 9(1):1–14.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31.

Uludağ, K. and Roebroeck, A. (2014). General overview on the merits of multimodal neuroimaging data fusion. *Neuroimage*, 102:3–10.

Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., and Thirion, B. (2011). Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Biennial International Conference on information processing in medical imaging*, pages 562–573. Springer.

Wang, Z., Zhu, X., Adeli, E., Zhu, Y., Nie, F., Munsell, B., Wu, G., et al. (2017). Multi-modal classification of neurodegenerative disease by progressive graph-based transductive learning. *Medical image analysis*, 39:218–230.

Wipf, D. P. and Nagarajan, S. S. (2008). A new view of automatic relevance determination. In *Advances in neural information processing systems*, pages 1625–1632.

Zhang, D., Shen, D., Initiative, A. D. N., et al. (2012a). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *NeuroImage*, 59(2):895–907.

Zhang, H., Schneider, T., Wheeler-Kingshott, C. A., and Alexander, D. C. (2012b). NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage*, 61(4):1000–1016.

Zhang, Y. and Yang, Q. (2017). A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

Zhou, T., Thung, K.-H., Liu, M., Shi, F., Zhang, C., and Shen, D. (2020). Multi-modal latent space inducing ensemble svm classifier for early dementia diagnosis with neuroimaging data. *Medical Image Analysis*, 60:101630.

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

Appendix A.

Table A.2: A description of 47 Modalities of UKB dataset used in this paper.

Abbreviation	full description
rest k (k=1-25)	Dual regression between IC k of 25 dimensional decomposition of rsfMRI and the whole brain
task z1	Z-statistics of emotion task contrast "shapes"
task z2	Z-statistics of emotion task contrast "face"
task z5	Z-statistics of emotion task contrast "faces>shapes"
task c1	Contrasts of parameter estimate of emotion task contrast "shapes"
task c2	Contrasts of parameter estimate of emotion task contrast "face"
task c5	Contrasts of parameter estimate of emotion task contrast "faces>shapes"
TBSS-FA	Tract-Based Spatial Statistics - fractional anisotropy
TBSS-MD	Tract-Based Spatial Statistics - mean diffusivity
TBSS-MO	Tract-Based Spatial Statistics - tensor mode
TBSS-L1	Tract-Based Spatial Statistics - amount of diffusion along the principal directions 1
TBSS-L2	Tract-Based Spatial Statistics - amount of diffusion along the principal directions 2
TBSS-L3	Tract-Based Spatial Statistics - amount of diffusion along the principal directions 3
TBSS-OD	Tract-Based Spatial Statistics - orientation dispersion index
TBSS-ICVF	Tract-Based Spatial Statistics - intra-cellular volume fraction
TBSS-ISOVF	Tract-Based Spatial Statistics - isotropic or free water volume fraction
tracts	summed tractography map of 27 tracts from AutoPtx in FSL
VBM	voxel-based morphometry
Area	Cortical surface area from Freesurfer
Thickness	Cortical surface thickness from Freesurfer
Jacobian	Jacobian map of nonlinear registration of T1 image to MNI152 standard space
swMRI	T2* image derived from swMRI
T2 lesion	White matter hyperintensity map estimated by BIANCA

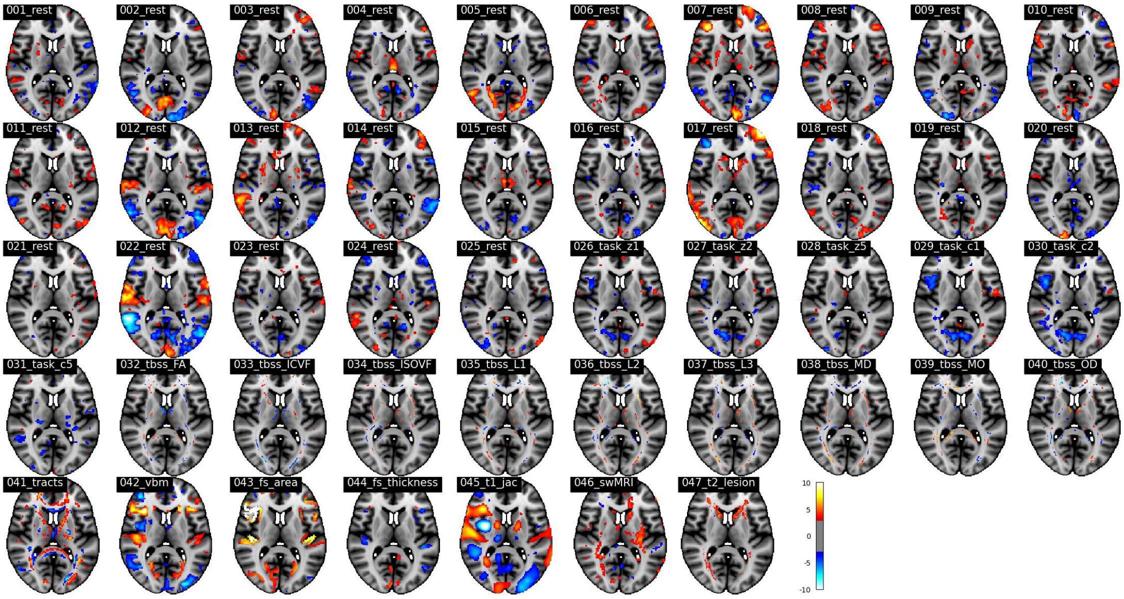

Figure A.6: The Z-score normalized spatial maps of the strongest modes that contributing to the prediction of *hypertension* in a 1,000-dimensional SuperBigFLICA, with all 17,485 nIDPs as supervision targets (MNI152 coordinate z=10).

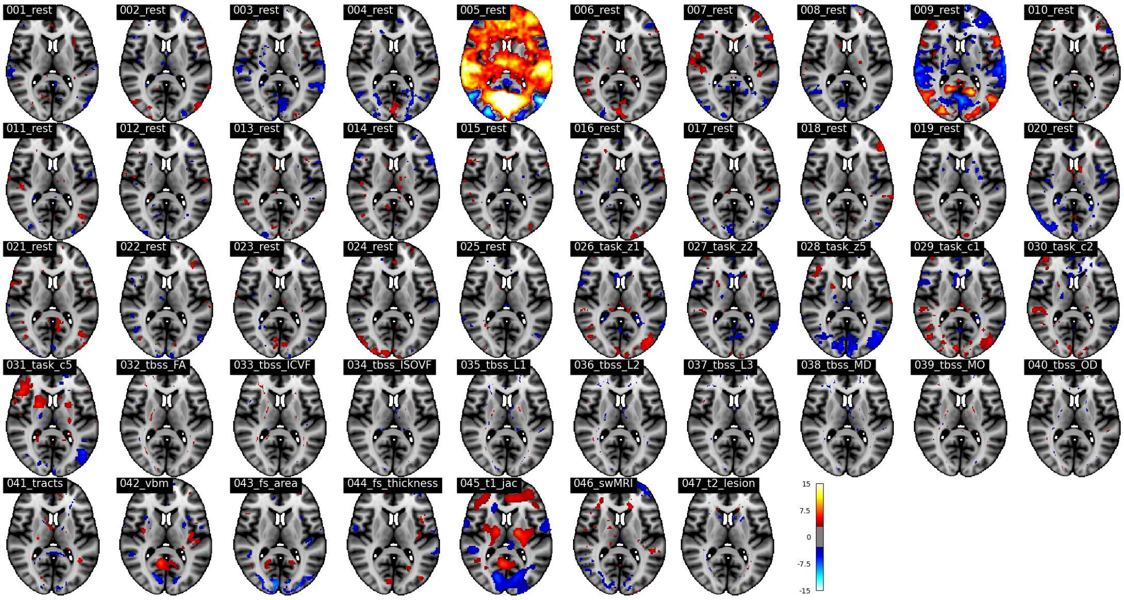

Figure A.7: The Z-score normalized spatial maps of the strongest modes that contributing to the prediction of *age started wearing glasses or contact lenses* in a 1,000-dimensional SuperBigFLICA, with all 17,485 nIDPs as supervision targets (MNI152 coordinate z=10).