# Modelling Multimodal Population Brain Imaging Data




Weikang Gong

St Peter's College

University of Oxford

A thesis submitted for the degree of

*Doctor of Philosophy*

Trinity 2021

# Acknowledgements

Firstly, I would express my sincere gratitude to my supervisors Prof. Stephen Smith and Prof. Christian Beckmann for their support, patience and guidance during my DPhil study. I have learnt so much from them. It's impossible for me to complete my thesis without them.

I would also thank all my colleagues at FMRIB, WIN and NDCN. I feel very fortunate to have the opportunity to work and play with them. Everyone is so talented and inspire me much during my study. Finally, I would thank my parents who always support and help me during my study; it would not be possible for me without you.

# Abstract

Neuroimaging allows for the non-invasive study of the brain in rich detail. Data-driven discovery of patterns of population variability in the brain has the potential to be extremely valuable for early disease diagnosis and understanding the brain. The resulting patterns can be used as imaging-derived phenotypes (IDPs), and may complement existing expert-curated IDPs. However, population datasets, comprising many different structural and functional imaging modalities from thousands of subjects, provide a computational challenge not previously addressed. Here, we proposed two approaches, termed BigFLICA and SuperBigFLICA, that are scalable for data fusion of voxel-level neuroimaging data in the full UK Biobank (UKB) dataset, that will soon reach 100,000 imaged subjects. The BigFLICA is an unsupervised multimodal independent component analysis approach, and the SuperBigFLICA improves BigFLICA by using the information from non-imaging phenotypes in a semi-supervised fashion. Both of them can estimate modes of population variability that enhance the ability to predict thousands of phenotypic and behavioural variables compared to previous approaches. Many interpretable associations with non-imaging phenotypes were identified, including multimodal spatial maps related to fluid intelligence, handedness and disease.

For the longitudinal prediction of brain imaging data, it remains largely unknown whether we can predict individualised longitudinal changes of brain imaging features based only on the baseline brain images, which would be of great value for longitudinal brain studies and early design of personalised treatments. We explore this possibility of using longitudinal data of 50 modalities in UK Biobank brain imaging studies with around 3500 subjects. We designed a deep U-Net based model for predicting longitudinal changes in multimodal brain images, and proposed a new mathematical framework for evaluating the longitudinal prediction results, including detecting the existence of changes in data, deriving baseline prediction performance and distinguishing change prediction from data denoising. Our results show that the proposed model can predict to a modest degree individualised longitudinal changes in almost all 50 modalities, and outperforms other popular models. Further, the predicted longitudinal changes have a similar or even improved accuracy of predicting subjects' non-imaging phenotypes, and have a high between-subject discriminability. Our results show the potentials of our model for longitudinal data imputation, and show several caveats when performing longitudinal data analysis.

# Contents

**List of Figures** **vii**

**List of Abbreviations** **xiv**

**1 Introduction** **1**
1.1 Background and motivation . . . . . 1
1.2 Technical background . . . . . 3
1.2.1 Bayesian probabilistic modelling . . . . . 3
1.2.2 Machine learning . . . . . 4
1.2.3 Unsupervised data decomposition approaches . . . . . 5
1.2.4 Elastic-net regression for supervised prediction . . . . . 10
1.2.5 Deep Learning . . . . . 10
1.3 Structure of this thesis . . . . . 11

**2 Phenotype Discovery from Population Brain Imaging** **12**
2.1 Introduction . . . . . 14
2.2 Methods . . . . . 18
2.2.1 Brief overview of the proposed approach: BigFLICA. . . . . 18
2.2.2 FLICA model. . . . . 19
2.2.3 Multimodal extension of MELODIC's Incremental Group Principal component analysis for subject-space dimension reduction. . 20
2.2.4 Sparse dictionary learning for voxel-space dimension reduction. 21
2.2.5 Evaluation of BigFLICA in simulations. . . . . 23
2.2.6 HCP and UK Biobank data. . . . . 24
2.2.7 Comparing BigFLICA with the original FLICA on real data. . . 25
2.2.8 Statistical significance of difference of prediction accuracy between two approaches. . . . . 26
2.2.9 Parameter settings of running BigFLICA in the full HCP and UKB datasets. . . . . 27
2.2.10 Comparing BigFLICA with multiple independent single-modality ICA decomposition. . . . . 27
2.2.11 Comparing BigFLICA with hand-curated imaging-derived phenotypes. . . . . 28

2.3 Results . . . 29
2.3.1 Evaluation of BigFLICA in simulations. . . . 29
2.3.2 Computation time comparison. . . . 30
2.3.3 Real data: Comparing BigFLICA with the original FLICA based on the prediction accuracy of nIDPs. . . . 30
2.3.4 Comparing BigFLICA with multiple independent single-modality ICA decomposition. . . . 33
2.3.5 Comparing BigFLICA with hand-curated image-derived phenotypes. . . . 35
2.3.6 Examples of BigFLICA modes in the 14k UKB dataset. . . . 36
2.3.7 BigFLICA comparison with mCCA and reproducibility. . . . 38
2.3.8 Reproducibility of BigFLICA. . . . 39
2.3.9 Stability of BigFLICA prediction. . . . 40
2.3.10 Contribution of different modalities in a BigFLICA decomposition. 41
2.3.11 Relationship between different modalities in a BigFLICA decomposition. . . . 42
2.4 Discussion . . . 43

**3 Supervised Phenotype Discovery from Multimodal Brain Imaging 47**
3.1 Introduction . . . 48
3.2 Methods . . . 52
3.2.1 SuperBigFLICA from an optimisation perspective. . . . 52
3.2.2 SuperBigFLICA in a Bayesian framework. . . . 54
3.2.3 Model parameter optimization. . . . 56
3.3 Experiments . . . 57
3.3.1 Brain imaging data. . . . 57
3.3.2 Non-imaging derived phenotypes. . . . 58
3.3.3 Confounding variables and missing values. . . . 58
3.3.4 Imaging data pre-reduction using dictionary learning. . . . 59
3.3.5 Baseline: nIDP prediciton using hand-curated IDPs and modes of BigFLICA. . . . 60
3.3.6 Predicting nIDPs using single-task and multi-task SuperBigFLICA. 61
3.3.7 Evaluation of the generalisability of SuperBigFLICA modes on unseen nIDPs. . . . 61
3.4 Results . . . 62
3.4.1 Comparing SuperBigFLICA with hand-curated IDPs and modes of unsupervised BigFLICA. . . . 62
3.4.2 Analysis of SuperBigFLICA algorithm. . . . 64
3.4.3 Real data further qualitative analysis. . . . 67
3.5 Discussion . . . 70

**4 Individualised Prediction of Longitudinal Changes in Brain Images 73**
4.1 Introduction . . . 74
4.2 Methods . . . 77
4.2.1 A general mathematical framework for longitudinal brain images. 77
4.2.2 Deep U-Net Model for longitudinal prediction . . . 80
4.2.3 Data . . . 83
4.2.4 Implementation details. . . . 84
4.2.5 Prediction of follow-up image with baseline image . . . 85
4.2.6 Prediction of longitudinal changes with baseline image . . . 86
4.3 Results . . . 89
4.3.1 Prediction performance of follow-up image . . . 89
4.3.2 Prediction performance of longitudinal changes . . . 92
4.3.3 Test of whether a model performs change prediction or data denoising . . . 94
4.3.4 Discriminability of prediction . . . 95
4.3.5 Relations between predicted images and nIDPs . . . 96
4.3.6 Real data qualitative analysis . . . 98
4.4 Discussion . . . 98

**5 Summary and future works 102**
5.1 Work summary . . . 102
5.2 Future work . . . 105
5.2.1 Multimodal fusion for data-driven IDP discovery. . . . 105
5.2.2 Longitudinal brain image prediction. . . . 106

**Appendices**

**A 108**
A.1 Supplementary materials for Chapter 1 . . . 109
A.1.1 Optimization algorithm for dictionary learning . . . 109
A.1.2 Optimization algorithm for Elastic-net regression . . . 110
A.2 Supplementary materials for Chapter 2 . . . 111
A.2.1 Confounding variables regressed out in Chapter 2 . . . 111
A.3 Supplementary materials for Chapter 3 . . . 126

**Bibliography 128**

# List of Figures

2.1 **Overview of the proposed approach for jointly analyzing a biobank-scale multimodal neuroimaging dataset.** Currently for the UKB dataset (voxel-level data, 14503 subjects, 47 modalities), the total data size is approximately 800 GB, and if we directly feed these data into FLICA and extract 750 components, we will need approximately 1066 GB CPU memory and 1680 hours computation time. Our new approach, BigFLICA, used multimodal MIGP and dictionary learning to preprocess the multimodal data; this is efficient and memory friendly, and much of this preprocessing can be easily parallelized. BigFLICA only used 50 GB memory and 73 hours to analyze the same dataset using a 24-core compute server. . . . . 17

2.2 **Evaluation of multimodal extension of MIGP (mMIGP) and dictionary learning (DicL) as the data preprocessing steps for the FLICA using simulations.** BigFLICA achieves similar performance as compared with original FLICA that uses the full data. a, Evaluation of mMIGP preprocessing. We compared the correlations (Z-transformed) of extracted components with ground truth across 50 simulations using the original FLICA (the left column of each figure) and the mMIGP preprocessed FLICA (other columns). The mMIGP dimensions vary between 50 and 400; the SNRs are between 4 and 0.04 (left to right), and the number of FLICA and ground truth components are 25, 35, 45 (top to bottom). As there are 500 subjects, the reduction factor is from 10 to 1.25. b, Evaluation of DicL preprocessing. We compared the correlations of extracted components with ground truth using the original FLICA (FLICA results given in the titles of each figure) and the DicL preprocessed FLICA with different sparsity parameters and dictionary dimensions (cells of the heatmaps). The SNRs are between 4 and 0.04 (left to right), and the number of FLICA and ground truth components are 25, 50, 100 (top to bottom). As there are 27,000 original features per modality, the reduction factor is from 270 to 9. . . . . 31

2.3 **Comparison of prediction accuracy of nIDPs between BigFLICA and the original FLICA.** Overall, for most of the comparisons, the differences of prediction accuracy are not significant. In each of the Bland–Altman plots, each point represents the prediction of one nIDP, where the x-axis is the average prediction correlation of the two approaches while the y-axis is the difference, i.e., BigFLICA - FLICA. The z- and p-values in the titles reflected the statistical significance of the differences. The Bonferroni correction 0.05 threshold corresponds to a raw p-value of 1.7e-3. a, Comparing FLICA with mMIGP preprocessing with the original FLICA. We used a subset of 1,036 subjects in the UKB dataset (top) and the HCP (bottom). The number of estimated FLICA components is set to 50, and mMIGP dimensions are set from 100 to 500. b, Comparing FLICA with DicL preprocessing with the original FLICA. We used a subset of 1,036 subjects in the UKB dataset (top) and the HCP (bottom). The dictionary dimension is set to a high value of 2000, and the sparsity parameter is set to $\lambda = 1$ for all modalities. The number of estimated FLICA components are set from 25 to 300. c, Comparing FLICA with both mMIGP and DicL preprocessing combined, with the original FLICA. The mMIGP dimension is set to 500, and other settings are the same as in b. We use only a subset of UKB here so that running the original FLICA is computationally feasible. The lighter the blue, the higher the density of points. . . . . 32

2.4 **Comparison of prediction accuracy of nIDPs between BigFLICA against single-modality ICA and the IDPs.** Overall, for high-dimensional BigFLICA decompositions in the UKB dataset, BigFLICA achieved statistical significant increases of prediction accuracy of nIDPs compared with single-modality ICA and IDPs. Combining BigFLICA and IDPs together further improves compared with IDPs alone. In each of the Bland–Altman plots, each point represents the prediction of an nIDP, where the x-axis is the average prediction correlation of the two approaches, while the y-axis is the difference. The z- and p-values in the titles reflected the statistical significance of the differences. The Bonferroni correction 0.05 threshold corresponds to a raw p-value of 1.7e-3. a, Comparing BigFLICA with the concatenation of single-modality ICA outputs. Top: UKB; Bottom: HCP. The number of FLICA components is set from 25 to 750. b, Comparing BigFLICA with IDPs. Top: UKB; Bottom: HCP. The number of IDPs is 3,913 in UKB and 5,812 in the HCP. c, Comparing the concatenation of BigFLICA and IDPs against IDPs only. Top: UKB; Bottom: HCP. The lighter the blue, the higher the density of points. . . . . 34

2.5 **Examples of BigFLICA modes in the 14k UKB dataset.** For each subfigure, each row shows one IC (BigFLICA mode or independent component) with top 4 most strongly associated modalities. **a,** Two BigFLICA modes that significantly correlate with *fluid intelligence* (IC25: $r = -0.14$; IC57: $r = -0.12$). **b,** Two BigFLICA modes that significantly correlate with *Age started wearing glasses or contact lenses* (IC164: $r = -0.10$; IC13: $r = -0.05$). **c,** Three BigFLICA modes that significantly correlate with *handedness* (IC235: $r = -0.23$; IC569: $r = 0.07$; IC232: $r = -0.04$). **d,** Three BigFLICA modes that significantly correlate with *hypertension* (IC259: $r = 0.12$; IC13: $r = -0.11$; IC319: $r = -0.09$). The Bonferroni corrected 0.05 threshold corresponds to an uncorrected p-value of $9.2 \times 10^{-9}$ (corrected for number of components (750) and number of nIDPs (7245)). All of the above correlations passed the Bonferroni threshold except for IC232 with uncorrected $p = 2.1 \times 10^{-7}$. 39

3.1 Overview of the proposed SuperBigFLICA approach for supervised multimodal fusion and phenotype discovery. . . . . 52

3.2 **Comparison of SuperBigFLICA with hand-curated IDPs and modes of unsupervised BigFLICA for the 12 nIDPs listed in Table 3.1.** 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. . . . . 63

3.3 **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 3.1. Top.** Single-task learning setting. **Mid.** Multi-task learning setting with 24 auxiliary tasks. **Bot.** Generalisability test. . . . . 64

3.4 **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. . . . . 66

3.5 **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). . . . 68

4.1 The proposed U-Net structure for longitudinal brain image prediction. The network has four downsampling and upsampling layers with skip connections. The numbers $a \times b^3$ represents the number of channels (a) and spatial voxel dimensions (b) of the inputs and outputs. . . . . 81

4.2 **Comparing the proposed U-Net approaches with other variants of U-Nets (Top) and other state-of-the-art methods (Bottom).** Correlations are between the actual residualised second scan against the predicted residualised second scan. "Residualisation" refers to the act of regressing out the group mean map (here the group mean first-scan image), to focus on aspects of subject maps that are distinct from each other. The mean correlation among all modalities are shown on the rightmost. . . . . 90

4.3 Prediction correlation of follow-up images (i.e., ResB) with multimodal and single modal U-Net, and a baseline of copying baseline image (i.e., ResA) as the prediction, i.e., r(ResB, PredResB | ResA). The mean correlations among all modalities are shown on the right. . . . . 91

4.4 For 25 resting-state modalities, the prediction correlation of follow-up images (i.e., ResB) with multimodal and single modal U-Net, and a baseline of copying baseline image (i.e., ResA) as the prediction, and the test-retest reproducibility by splitting the time series in half. The mean correlations among all modalities are shown on the right. Among the 25 modalities, component 4, 23, 24, 25 have been identified as non-neural components. . . . . 92

4.5 Prediction correlation of temporal difference images (i.e., ResB-ResA) with multimodal and single modal U-Net, i.e., r(ResB-ResA, PredResB-ResA | ResA). The mean correlations among all modalities are shown on the right. . . . . 93
4.6 For 25 resting-state modalities, the prediction correlation of temporal difference image (i.e., ResB-ResA) with multimodal and single modal U-Net., and a baseline of copying baseline image (i.e., -ResA) as the prediction, and the estimated null correlation when there are no longitudinal changes by the mathematical model. The mean correlations among all modalities are shown on the right. . . . . 93
4.7 Prediction correlation of residual temporal difference images (i.e., ResB\ResA) with multimodal and single modal U-Net. Here the baseline correlation is zero because ResA has been regressed out, i.e., r(ResB\ResA, PredResB\ResA | ResA). The mean correlations among all modalities are shown on the right. . . . . 94
4.8 The results of a test which distinguishes whether a model is carrying out change prediction vs. data denoising. Given a trained model, this compares the correlation of the output with the baseline image fed in and the real follow-up image, i.e. *r*(PredResB, ResB) or r(PredResB\ResA, ResB\ResA), against the correlation of output with the follow-up image fed in and the real baseline image, i.e., *r*(PredResA, ResA) or r(PredResA\ResB, ResA\ResB). . . . . 96
4.9 The subject identification rate of the predicted maps. We correlate each of the predicted maps to all the true maps. This shows the proportion of subjects whose within-subject correlation is the largest among all between-subject correlations. The mean identification rates among all modalities are shown on the right. . . . . 97
4.10 The Bland-Altman plots of prediction performance of nIDPs using the predicted and true residual temporal difference images. Each point in the figure represents one nIDP, where the x-axis is the mean of its prediction correlation and the y-axis is the difference of the correlation between predicted images and true images. . . . . 98
4.11 For 25 resting-state modalities, the improvements of prediction correlation of nIDPs between predicted and true residual temporal difference images positively correlated with the difference of follow-up image prediction correlation and test-retest reproducibility ($r = 0.59, p = 0.002$). 99
4.12 Real data examples of predicting residual temporal difference maps for 5 modalities (columns). A,B are two examples (but may not be the same subjects across modalities). . . . . 100

A.1 The relationship between the difference of prediction accuracy of BigFLICA and FLICA across different number of components. . . . . 113
A.2 Comparing the prediction performance of the 750-dimensional FLICA with the 25-dimensional single-modality ICA concatenated across 47 modalities in the UKB data. . . . . 113
A.3 **Relationships between FLICA and IDPs. (a,b)** The plots show the results of predicting each IDP using BigFLICA modes in **a** the HCP and **b** the UKB dataset. The IDPs are appearing in order along the x axis, and are grouped and coloured by modality types. **(a,b)** The histograms of predicting BigFLICA modes using all IDPs in **c** the HCP and **d** the UKB dataset. . . . . 114
A.4 **Comparing FLICA and mCCA in the UKB data.** Comparing the predictive performance of FLICA with mCCA (or equivalently the subject-by-component matrix obtained in the mMIGP step) across different numbers of extracted components in the UKB dataset. The FLICA and mCCA dimensions are the same in each figure. . . . . 114
A.5 The difference of the proportion of zeros weights (BigFLICA-mCCA) in predicting nIDPs across 5 dimensions of decomposition in the UKB data. 115
A.6 **Split-half reproducibility of BigFLICA and mCCA spatial maps in the UKB dataset.** The split-half reproducibility of BigFLICA and mCCA in the UKB dataset by first computing the correlation between modality-wise concatenated spatial maps after eliminating low-weight voxels and then greedy matching. . . . . 115
A.7 Split-half prediction accuracy of all nIDPs of BigFLICA in the UKB dataset. . . . . 116
A.8 Split-half reproducibility of BigFLICA using different sample size in the UKB dataset. Left: sample size = 1500. Right: sample size = 3500. . . 116
A.9 **Stability of prediction accuracy of BigFLICA against different train-test subject splits.** We estimated the prediction correlation of each nIDP using 5 different train-test subject splits, and calculated the mean accuracy of prediction correlations for each nIDPs. We then computed the difference of the mean prediction correlation and the prediction correlation of one of the five random predictions across different number of components. . . . . 117
A.10 **Comparison of prediction accuracy of nIDPs when removing 4 noise resting-state modalities with including them in the model.** An increased prediction accuracy was observed for some nIDPs in the 25-dimensional decomposition when compared with using all 47 modalities, and a similar prediction accuracy was observed for 100- and 250-dimensional decomposition, while a decreased prediction accuracy was observed for 500- and 750-dimensional decompositions. . . . . 117

A.11 **Group average maps of task activations and resting-state networks in the UKB dataset.** These are provided to help interpret the population variability maps (modulations of these maps) shown in **Fig. 2.5**. Top: group average of emotion task activation z-statistic maps (task z1: "shapes", task z2: "face", task z5: "faces>shapes"). Group average task contrast (effect size) maps c1, c2 and c5 are highly similar to z-stat maps so that they are not shown. Bottom: group average resting-state networks from a 25-dimensional ICA parcellation in the UKB data. The six maps shown here are the networks from **Fig. 2.5**. . . . . 118

A.12 **The contribution of each modality in each BigFLICA mode (independent component).** . . . . . 119

A.13 **The relative contribution of different modalities of a BigFLICA decomposition (nIC=25-750) in the UKB data.** For each modality, we take the sum of its overall contribution (estimated by BigFLICA) across all components. . . . . 120

A.14 **The relationships between different modalities in the UKB data.** a). The cosine similarity of modality contributions across 750 components (estimated by BigFLICA) between every pair of modalities. b). The amount of shared variance between two 50-dimensional single-modality ICA decompositions in each pair of modalities. . . . . 121

A.15 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). . . . . 127

A.16 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). . . . . 127

# List of Abbreviations

**1D, 2D, 3D** . . . One- or two- or three-dimensional, referring in this thesis to spatial dimensions in an image.

**MRI** . . . . . Magnetic Resonance Imaging.

**fMRI** . . . . . functional Magnetic Resonance Imaging.

**rfMRI** . . . . . resting-state functional Magnetic Resonance Imaging.

**tfMRI** . . . . . task functional Magnetic Resonance Imaging.

**dMRI** . . . . . diffusion Magnetic Resonance Imaging.

**HCP** . . . . . Human Connectome Project.

**UKB** . . . . . UK Biobank.

**IDP** . . . . . image-derived phenotype.

**nIDP** . . . . . non-image-derived phenotypes.

**SVD** . . . . . singular value decomposition.

**PCA** . . . . . principal component analysis.

**mCCA** . . . . . multi-set canonical correlation analysis.

**ICA** . . . . . independent component analysis.

**FLICA** . . . . . FMRIB's Linked ICA.

**MIGP** . . . . . MELODIC's Incremental Group Principal component analysis.

**DicL** . . . . . dictionary learning.

**SNR** . . . . . signal-to-noise ratios.

# 1 Introduction

## Contents


**1.1 Background and motivation . . . . . 1**
**1.2 Technical background . . . . . 3**
1.2.1 Bayesian probabilistic modelling . . . . . 3
1.2.2 Machine learning . . . . . 4
1.2.3 Unsupervised data decomposition approaches . . . . . 5
1.2.4 Elastic-net regression for supervised prediction . . . . . 10
1.2.5 Deep Learning . . . . . 10
**1.3 Structure of this thesis . . . . . 11**


## 1.1 Background and motivation

One of the major goals for brain imaging studies is to understand the complex relationships between our brain and demographic, behavioural, clinical and cognitive measures (i.e., non-imaging derived phenotypes or nIDPs). With advances in the development of neuroimaging techniques, we are able to acquire multimodal, high-resolution brain imaging data on large scale. Typical such studies involves UK Biobank [Miller et al., 2016], the Human Connectome Project [Van Essen et al., 2013] and the Adolescent Brain Cognitive Development (ABCD) study [Jernigan et al., 2018]. These studies involve multimodal data, meaning that several distinct types of MRI data are acquired, mapping activity, functional networks, structural connectivity, white matter microstructure, and organisation

and volumes of different brain tissues and sub-structures [Miller et al., 2016]. At the same time, they also acquired a rich set of nIDPs and genetic data, giving us great opportunities to study the brain.

The neuroimaging field is shifting somewhat from studying *population average* patterns to studying *individual differences* [Gordon et al., 2017, Lebreton et al., 2019]. The study of population average patterns has been a primary focus of many traditional neuroimaging studies. For example, many studies used the statistical parametric mapping (SPM) approach to estimate the spatial patterns of activations when subjects are performing certain tasks [Penny et al., 2011], which aims to answer the questions like: What would be the average pattern of task activations across the population? This method uses a voxel-wise general linear model approach with random field theory based multiple comparison correction methods to locate significant brain regions in certain tasks. Similarly, for other modalities, similar approaches are developed and used, such as voxel-based morphometry (VBM)[Ashburner and Friston, 2000] for studying the grey matter volume from structural MRI (sMRI), and Tract-based spatial statistics (TBSS) [Smith et al., 2006] for studying white matter microstructure from diffusion MRI (dMRI). To move a step further, we are interested here in how individual differences in the brain's structure and functions may relate to nIDPs, this may enable us to generate individualised markers for early disease diagnosis and prediction of disease progression and the understanding of human cognition differences. There are many ways of studying individual brain differences, and one of the main approaches is to use the predictive modelling approach. In this approach, we build a predictive model of, e.g., brain imaging to behaviour scores, with the aim of achieving accurate out-of-sample predictions. There are a huge amount of such studies, we only list a few examples here, such as brain age prediction [Peng et al., 2021], early prediction of Alzheimer's disease [Zhu et al., 2021] and prediction of individual task activations from resting-state connectivities [Tavor et al., 2016]. Although there are many successful studies of individual differences in brain structure and function, there are still limitations from a methodological perspective. For example, traditional approaches may not be practical for current datasets simply because of the increases in sample sizes and the number of modalities, and the richness of information that might

be extracted at the individual subject level is far greater than is achieved by traditional methods. Therefore, powerful and efficient methods to analyse large-scale datasets such as UK Biobank are needed.

In this thesis, I will focus on two important topics in modelling multimodal brain imaging data. The first is how to fuse multimodal brain imaging data to extract brain imaging-derived phenotypes from them, and the second is how to model the longitudinal changes of brain images. I will introduce our motivations for these two topics in detail in subsequent chapters. Here, in brief, we found that most of the current approaches are either too simple and lose much information from multimodal data, or are too computationally demanding to be directly applied to our biobank-scale dataset. Therefore, we developed more powerful and efficient approaches to address the problems, and compared them with traditional approaches. For the first topic, we explore two ways of modelling multimodal data, one based on unsupervised approaches while the other is based on supervised modelling approaches, to extract low-dimensional imaging features. For the second topic, we will explore the use of a deep neural network approach to model the longitudinal changes of brain images, i.e., to predict follow-up images from the baseline brain images. These works involve knowledge of Bayesian probabilistic modelling, data decomposition approaches, predictive modelling approaches and deep neural networks. I will briefly review them in the following *technical background* section, although the majority of the relevant background material is covered within the 3 main research chapters.

## 1.2 Technical background

### 1.2.1 Bayesian probabilistic modelling

Bayes' Rule is the fundamental concept in Bayesian probabilistic modelling. Given two random variables $x$ and $y$, we have:

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} \tag{1.1}$$

In probabilistic machine learning, we usually have an explicit forward model to map the input to the output. Denoting these as $x$ and $y$, we have:

$$y = f(x|w) \tag{1.2}$$

where $w$ are the model's parameter (to be estimated), which will usually be assumed to follow some probabilistic distribution. Then, given the observed data pairs $(x, y)$, we have:

$$p(w|y,x) = \frac{p(y|w,x)p(w)}{p(y)} \tag{1.3}$$

We call $p(w)$ the prior of the model parameters independent of the data $x$, $p(y|x,w)$ the likelihood of model, and $p(w|y,x)$ the posterior distribution of model parameters. Unlike traditional approaches that aim to maximize the model likelihood, Bayesian approaches estimate the posterior distribution of model parameters, which can involve both point estimation of $w$ (e.g. the maximum a posterior value) and the uncertainty of $w$. The involvement of priors on $w$ can let us incorporate human knowledge in the modelling, and specific prior distributions can regularize the model parameters toward certain properties (e.g., sparsity).

### 1.2.2 Machine learning

The machine learning algorithm is a type of artificial intelligence that learns to make predictions or decisions with training data. Machine learning algorithms are used in a wide variety of applications, such as in medicine, nature language processing, and computer vision. Classical machine learning algorithms are usually categorized into four classes: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. In this thesis, the approaches that we developed and applied are mainly unsupervised learning and supervised learning algorithms. The unsupervised learning algorithms model a set of data that only have the input, and try to find the structure within the data, such as decomposing the data into low dimensional features or clustering the samples into difference groups. The supervised learning algorithms model the data with both input and the target. The algorithms try to learn a function to model the relationship between the input and the target using training data, for example, to classify if a picture contains a dog or cat. Below, we will briefly review some of the unsupervised and supervised learning approaches used in this thesis.

### 1.2.3 Unsupervised data decomposition approaches

Various approaches can be classified as unsupervised learning approaches. Here, we will primarily introduce unsupervised data decomposition approaches. We will assume we have one or more data matrices $X$ of size $N \times P$, where $N$ is the number of observations and $P$ is the number of features. The methods introduced in this section all try to find a set of vectors, a basis, for the data. So we can write any point (vector) in the data as a linear combination of the basis. The only difference among these approaches is the model constraints.

**Principal component analysis (PCA)**

The PCA approach has many equivalent formulations. Given a data matrix $X$ (assuming that each observation has been mean centered for simplicity), PCA seeks to minimize the L2 data reconstruction error:

$$\min_{U,V} ||X - UV^{\tau}||_2 \tag{1.4}$$

where $U$ is a set of L-dimensional orthonormal basis, and $V$ is the loading. Here, 'orthonormal' means the columns of $U$ are orthogonal to each other and each of their L2 norms is one. The $U$ is estimated by an eigendecomposition on the matrix $XX^{\tau}$ or a singular value decomposition (SVD) on the matrix $X$.

The PCA can learn the set of orthogonal principal components that explain a maximum amount of the variance in a multivariate dataset. It is one of the most popular data exploration tools.

**Dictionary learning**

Dictionary learning models a data matrix as linear combinations of sparse dictionary basis elements [Mairal et al., 2010]. The difference between PCA and dictionary learning is that, in dictionary learning, we no longer impose orthogonal constraints, instead, we put an L1 loss on the basis to find a sparse representation of data. This is because, in many cases such as image processing, the underlying components can be assumed

to be sparse. Therefore, the components found by dictionary learning are sometimes more interpretable than the PCA.

In details, given a data matrix $X$, the dictionary learning linearly decomposes the data:

$$X = AD \tag{1.5}$$

where $D \in \mathbb{R}^{L \times P}$ is the *sparse dictionary basis, and* $A \in \mathbb{R}^{N \times L}$ *is the* feature loadings. By minimizing an $l_1$-regularized sparse-coding objective function, a local optimal solution can be obtained:

$$\begin{aligned} &\min_{A,D} \sum_{i=1}^{N} \frac{1}{2} \|X_i - AD_i\|_2^2 + \lambda \|D_i\|_1 \\ &s.t.\ A_j^{\tau} A_j \leqslant 1 \end{aligned} \tag{1.6}$$

where subscript $i, j$ represents the $i, j$-th column of a matrix, and $\lambda$ is a regularization parameter. The $l_1$-regularization term enforces that the learned dictionary basis $D$ is sparse. The objective function can be efficiently optimized by a block-coordinate descent optimizer with warm restarts or variants of stochastic gradient descent (SGD) algorithms [Mairal et al., 2010]. We show details in the Appendix.

**Independent component analysis (ICA)**

ICA models observed data as linear combinations of independent components [Hyvärinen and Oja, 2000, Hyvarinen, 1999]. Compared with PCA, it imposes a stronger independent constraint on the low dimensional components. In other words, we assume that the data is a linear mixture of independent signals.

In detail, given data matrix $X$, the ICA model is:

$$X = SA \tag{1.7}$$

where $S$ is the independent components and $A$ is the feature loadings.

The definition of independence originates from a probabilistic distribution: let $x_1$ and $x_2$ be two random variables, they are independent if and only if:

$$p(x_1, x_2) = p_1(x_1) p_2(x_2) \tag{1.8}$$

The basic principle of ICA estimation is nongaussianity of components, maximised when they are independent (with the assumption that the components are not gaussian distributed) [Hyvärinen and Oja, 2000]. Therefore, to estimate independent components $S$, we make a change of variable: $Y = XW$ and maximize the nongaussianity of $XW$ ($W = A^{-1}$).

One classical measure of nongaussianity is the *kurtosis* or the fourth-order cumulant:

$$\text{kurt}(y) = E(y^4) - 3(E(y^2))^2 \tag{1.9}$$

Kurtosis has been widely used as a measure of nongaussianity in ICA and related fields, but the main problem of kurtosis is that it can be very sensitive to outliers. Another measure of nongaussianity is *negentropy*. It is defined as the difference between the entropy of a gaussian random variable $y_{gauss}$ and another random variable $y$:

$$\begin{aligned} J(y) &= H(y_{gauss}) - H(y) \\ H(y) &= -\int f(y) \log f(y) dy \end{aligned} \tag{1.10}$$

where $f(y)$ is the density of $y$. This is because a gaussian variable has the largest entropy among all random variables of equal variance. Negentropy is related to *mutual Information* up to a constant.

However, negentropy is hard to estimate, because estimating negentropy using the definition would require an estimate (possibly nonparametric) of the probability density function [Hyvärinen and Oja, 2000], so approximation techniques are needed. An approximation of negentropy used in FastICA [Hyvarinen, 1999] is:

$$J(y) \propto (E(G(y)) - E(G(y_{gauss})))^2 \tag{1.11}$$

where $G$ is some nonquadratic function [Hyvärinen and Oja, 2000], such as $G(y) = |y|$, $G(y) = y^3$ and $G(y) = -\exp(-y^2/2)$.

In summary, the objective function for ICA problem is:

$$\min_W J(y, W), \; s.t. \; WW^\tau = I \tag{1.12}$$

where $J(y)$ is the negentropy which measures the nongaussianity, i.e., independence, of the estimated components, and $WW^\tau = I$ is the orthonormality constraint used to prevent

the bases from becoming degenerate [Le et al., 2011]. FastICA uses an Newton's method to optimize this objective function [Hyvarinen, 1999].

A recent variant of ICA, called reconstruction ICA [Le et al., 2011], showed that the orthonormality constraint can be replaced by a more flexible (soft) reconstruction loss. Then the ICA objective function becomes:

$$\min_W J(y,W) + \lambda ||XWW^\tau - X||_2^2 \tag{1.13}$$

Here $WW^\tau$ will approximate the identity matrix if the data reconstruction is good. The authors proved that orthonormality constraint and reconstruction loss are mathematically equivalent under certain conditions. The benefit of using reconstruction loss is that the optimization problem is unconstrained so that we can directly apply modern SGD optimizers to it; also, it is easy to expand the cost function with other terms, such as supervised aspects.

**Bayesian ICA and variational inference**

Bayesian ICA uses probabilistic models to model independent sources [Choudrey, 2002, Roberts and Choudrey, 2004]. In Bayesian ICA, we assume the following probabilistic model:

$$X = SA + E \tag{1.14}$$

where $S$ is the independent components and $A$ is the feature loadings, and the noise term $E$ is assumed to be Gaussian distributed, with zero mean and diagonal precision matrix, and the source is column-wise independent and follows a mixture-of-Gaussian prior distribution:

$$\begin{aligned} p(X|S) &= |\det\left(\frac{R}{2\pi}\right)|^{\frac{1}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}\left((X-SA)^\tau R(X-SA)\right)\right] \\ p(S) &= \Pi_{i=1}^L p(S_i) = \Pi_{i=1}^L \sum_{c=1}^{C_i} \pi_{i,c} N(s_i|u_{i,c},\beta_{i,c}) \end{aligned} \tag{1.15}$$

where $L$ is the latent dimension, $C_i$ is the number of mixture-of-Gaussian components, $u_{i,c}$ and $\beta_{i,c}$ are the mean and variance of each individual Gaussian component. In practice, we don't know the source distributions, so the choice of mixture-of-Gaussian

prior should be capable of encompassing both super- and sub-Gaussian distributions (distributions with positive and negative kurtosis respectively) and complex multimodal distributions [Choudrey, 2002]. To this end, the estimation of independent components and source distributions are done simultaneously in the optimization process. This is strongly connected to mutual information-based ICA [Hyvärinen and Oja, 2000].

Bayesian ICA estimates the posterior distribution of source distribution using Bayes' rule:

$$p(S|X) \propto P(X|S)P(S) \tag{1.16}$$

The exact posterior distribution is usually hard to estimate (which involves the production of Gaussian and Gaussian mixture), so a computationally efficient method is used called Variational Bayes (VB). Instead of maximizing the posterior probability, VB maximizes a lower bound of marginal data likelihood $F(X)$:

$$F(X|W) = E(\log p(X,W))_{q(W)} + H(q(W)) \leqslant p(X) \tag{1.17}$$

where $q(W)$ is the approximate posterior distribution of model parameters, and $H(\cdot)$ is the entropy of a distribution. For more details about VB, see e.g. [Blei et al., 2017, Wainwright and Jordan, 2008, Bishop, 2006]. In VB, the trick is to choose an appropriate factorisation for the posterior, which sometimes is more art than science [Choudrey, 2002]. A classical way of using VB is based on taking the derivative of $F(X|W)$ with respect to $q(W)$ and setting the resulting derivative to 0, to get an analytical expression for the approximate posterior. This approach is used in Bayesian ICA [Choudrey, 2002] and FMRIB's Linked ICA (FLICA) [Groves et al., 2011]. Recent advanced VB techniques have allowed direct optimization of lower bound $F$, with mini-batch SGD optimizers [Hoffman et al., 2013, Ranganath et al., 2014, Kingma and Welling, 2013] and auto-differentiation techniques [Baydin et al., 2018]. These advances increase the computation efficiency of traditional VB methods, and bypass the need for deriving analytical gradients of F(X|W), and allow more flexible design of model architectures.

### 1.2.4 Elastic-net regression for supervised prediction

There are many classes of supervised learning approaches in the literature. For neuroimaging data, the Elastic-net regression approach was found to perform well in various tasks [Zou and Hastie, 2005], so we briefly introduce it here. Elastic-net regression is a general linear model with L1 and L2 regularization. Given the features $X$ and the response $Y$ (assuming that $X$ has been column-wise z-scored for simplicity), Elastic-net regression models them as:

$$Y = Xw + E \tag{1.18}$$

and the loss function for regression is:

$$L = ||Y - Xw||_2^2 + \lambda(\alpha||w||_1 + (1-\alpha)||w||_2^2/2) \tag{1.19}$$

where $w$ is the regression weight, and $\lambda$ and $\alpha$ are regularization parameters. The L1 regularization enforces the model to get sparse coefficients. However, when there are high inter-feature correlations, L1 regularization tends to select only one feature from a group of correlated features. When combined with L2 regularization, the model can select all relevant features for prediction [Zou and Hastie, 2005]. The details for optimizing Elastic-net loss function is shown in the Appendix section.

### 1.2.5 Deep Learning

Deep neural networks have been widely used in many fields and achieved state-of-the-art performance in many tasks including image classification, natural language processing, biomedical image segmentation, etc. The core of deep neural networks is their multi-layer design that can fit a large family of complex distributions. The development of hardware, i.e., GPU, also benefits the application of deep learning. Our collaborative works on brain age prediction demonstrated the power of convolutional neural networks for predicting age using brain sMRI data [Peng et al., 2021, Gong et al., 2021c]. We will not introduce specific types of neural networks here because this is largely beyond the scope of this thesis. Interested readers can refer to [Goodfellow et al., 2016] for more details.

In this thesis, the main deep learning approach we used is the U-net, which has originally developed for image segmentation [Ronneberger et al., 2015]. The U-net structure is a specific type of fully convolutional network. It consists of a downsampling path followed by an upsampling path, with skip connection connecting feature maps with the same spatial dimensions. Given an input image, the U-net first downsamples the image to extract features, and then upsampling the image that tries to recover the information within the original input image. The loss function for U-net can be voxel-wise cross-entropy loss for segmentation, or mean square error for prediction. We will provide more details in Chapter 4.

## 1.3 Structure of this thesis

I will describe my DPhil work in the following three chapters. In Chapter 2, I will introduce BigFLICA, a new multimodal "big data" ICA method for unsupervised discovery of brain imaging phenotypes [Gong et al., 2021b]. In Chapter 3, a supervised extension of BigFLICA is proposed, called SuperBigFLICA, which further improves the performance of brain imaging phenotype extraction [Gong et al., 2021a]. In Chapter 4, I will introduce our initial results for the longitudinal prediction of brain imaging. We have developed a deep network architecture and a mathematical framework to predict and validate an interesting task: how to use baseline brain imaging data to predict the longitudinal changes of imaging features? In Chapter 5, I will summarize my DPhil work and propose some exciting future directions.

# 2

# Phenotype Discovery from Population Brain Imaging

## Contents


**2.1 Introduction . . . . . 14**
**2.2 Methods . . . . . 18**
2.2.1 Brief overview of the proposed approach: BigFLICA. . . . . 18
2.2.2 FLICA model. . . . . 19
2.2.3 Multimodal extension of MELODIC's Incremental Group Principal component analysis for subject-space dimension reduction. . . . . 20
2.2.4 Sparse dictionary learning for voxel-space dimension reduction. 21
2.2.5 Evaluation of BigFLICA in simulations. . . . . 23
2.2.6 HCP and UK Biobank data. . . . . 24
2.2.7 Comparing BigFLICA with the original FLICA on real data. . 25
2.2.8 Statistical significance of difference of prediction accuracy between two approaches. . . . . 26
2.2.9 Parameter settings of running BigFLICA in the full HCP and UKB datasets. . . . . 27
2.2.10 Comparing BigFLICA with multiple independent single-modality ICA decomposition. . . . . 27
2.2.11 Comparing BigFLICA with hand-curated imaging-derived phenotypes. . . . . 28
**2.3 Results . . . . . 29**
2.3.1 Evaluation of BigFLICA in simulations. . . . . 29
2.3.2 Computation time comparison. . . . . 30
2.3.3 Real data: Comparing BigFLICA with the original FLICA based on the prediction accuracy of nIDPs. . . . . 30

2.3.4 Comparing BigFLICA with multiple independent single-modality ICA decomposition. . . . . 33
2.3.5 Comparing BigFLICA with hand-curated image-derived phenotypes. . . . . 35
2.3.6 Examples of BigFLICA modes in the 14k UKB dataset. . . . 36
2.3.7 BigFLICA comparison with mCCA and reproducibility. . . . 38
2.3.8 Reproducibility of BigFLICA. . . . . 39
2.3.9 Stability of BigFLICA prediction. . . . . 40
2.3.10 Contribution of different modalities in a BigFLICA decomposition. . . . . 41
2.3.11 Relationship between different modalities in a BigFLICA decomposition. . . . . 42
**2.4 Discussion . . . . . 43**

# Summary

Neuroimaging allows for the non-invasive study of the brain in rich detail. Data-driven discovery of patterns of population variability in the brain has the potential to be extremely valuable for early disease diagnosis and understanding the brain. The resulting patterns can be used as imaging-derived phenotypes (IDPs), and may complement existing expert-curated IDPs. However, population datasets, comprising many different structural and functional imaging modalities from thousands of subjects, provide a computational challenge not previously addressed. Here, for the first time, a multimodal independent component analysis approach is presented that is scalable for data fusion of voxel-level neuroimaging data in the full UK Biobank (UKB) dataset, that will soon reach 100,000 imaged subjects. This new computational approach can estimate modes of population variability that enhance the ability to predict thousands of phenotypic and behavioural variables using data from UKB and the Human Connectome Project. A high-dimensional decomposition achieved improved predictive power compared with widely-used analysis strategies, single-modality decompositions and existing IDPs. In UKB data (14,503 subjects with 47 different data modalities), many interpretable associations with non-imaging phenotypes were identified, including multimodal spatial maps related to fluid intelligence, handedness and disease, in some cases where IDP-based approaches failed.

## 2.1 Introduction

Large-scale multimodal brain imaging has enormous potential for boosting epidemiological and neuroscientific studies, generating markers for early disease diagnosis and prediction of disease progression, and the understanding of human cognition, by means of linking to clinical or behavioural variables. Recent major studies have been acquiring brain magnetic resonance imaging (MRI), genetics and demographic/behavioural data from large cohorts. Examples are the UK Biobank (UKB) [Miller et al., 2016], the Human Connectome Project (HCP) [Van Essen et al., 2013] and the Adolescent Brain Cognitive Development (ABCD) study [Jernigan et al., 2018]. These studies involve multimodal data, meaning that several distinct types of MRI data are acquired, mapping activity, functional networks, structural connectivity, white matter microstructure, and organisation and volumes of different brain tissues and sub-structures [Miller et al., 2016]. However, the multimodal, high-dimensional and noisy nature of such big datasets makes many existing analytical approaches for extracting interpretable information impractical [Smith and Nichols, 2018].

Traditionally, large-scale neuroimaging studies first summarize the imaging data into interpretable image-derived phenotypes (IDPs) [Miller et al., 2016, Elliott et al., 2018], which are scalar quantities derived from raw imaging data (e.g., regional volumes from structural MRI, mean task activations from task MRI, resting-state functional connectivities between brain parcels). This knowledge-based approach is simple and efficient, and effectively reduces the high-dimensional data into interpretable, compact, convenient features. However, there may well be a large loss of information, due to such "expert-hand-designed" features not capturing important sources of subject variability (or even just losing sensitivity by the pre-defined spatial sub-areas being suboptimal), as well as ignoring cross-modality relationships. Further, such uni-modal compartmentalised analyses do not utilise the fact that for many biological effects of interest we expect there to be biological convergence across different data modalities, iė; changes in the underlying biological phenotype likely manifest themselves across multiple quantitative

phenotypes, so that a joint analysis effectively increases both the power of detecting such effects and the interpretability of the findings.

In contrast to such uni-modal analyses, data-driven multivariate approaches (i.e., unsupervised machine learning) have been proposed, which perform simultaneous decomposition of voxel-level data directly, generally representing data as the summation of a number of “components” or “modes”. Each mode is formed as the outer product of two vectors: one is a vector of subject weights (describing the relative strength of expression of that mode in each subject), and a vector of voxel weights (in effect a spatial map for each data modality, describing the spatial localisation of the mode). The subject weight vectors (one per mode) can be considered "features" (similar to IDPs, but being data-driven) for use in further modelling, such as for the prediction of non-imaging variables. They are often either based on eigendecomposition, such as multi-set canonical correlation analysis (mCCA) [Kettenring, 1971, Klami et al., 2015], or based on variations of independent component analysis (ICA) [Calhoun et al., 2006, Liu et al., 2009, Beckmann and Smith, 2005, Groves et al., 2011]. Among them, FMRIB’s Linked ICA (FLICA) [Groves et al., 2011] is an efficient approach which has been successfully applied to identify brain systems that are involved in lifespan development and diseases [Groves et al., 2012, Douaud et al., 2014], attention deficit hyperactivity disorder [Ball et al., 2018], preterm brain development [Ball et al., 2017] and cognition and psychopathology [Alnæs et al., 2018]. FLICA has advantages compared with uni-modal analysis on IDPs, including: (1) It leverages the cross-modality information of multimodal data, so has the potential to detect patterns that are not discoverable in any single modality; (2) It is a data-driven objective approach which automatically discovers meaningful patterns in voxel-level multimodal data by searching for spatial non-Gaussian sources that have been shown to likely reflect real structured features in neuroimaging data [Griffanti et al., 2014]. While this approach has been applied successfully to medium-sized cohort data [Groves et al., 2012, Douaud et al., 2014, Ball et al., 2018, Ball et al., 2017, Alnæs et al., 2018], the original algorithms for carrying out FLICA do not scale well with increasing data size, and are unable to analyze large datasets such as UKB, where dozens of different modalities over tens of thousands of subjects are available.

Importantly, because the core FLICA algorithms are multivariate, acting in a complex way simultaneously across all subjects, modalities and voxels using Variational Bayesian updates of parameters, this problem cannot be solved through simple parallelisation or other algorithmically simple methods for distributing computations across a large cluster, and so cannot be addressed simply by increasing the number of processors or memory available.

To tackle this problem, we propose an approach that embeds advanced data compression techniques across the different data dimensions into the FLICA approach. We use a multimodal extension of MELODIC's Incremental Group Principal component analysis [Smith et al., 2014] (mMIGP, applied across modalities) and online dictionary learning [Mairal et al., 2010] (DicL, applied within-modalities) to efficiently reduce the size of multimodal neuroimaging data. The reduced data are then characterised through FLICA in terms of underlying modality-specifc maps and subject loading vectors. Here we refer to this combination of techniques as Big-data FLICA, or BigFLICA for short. Two important advantages of the proposed approach are: (1) Preserving key information in original data but also reducing the effects of stochastic domain-specific noise; (2) Increasing the computational efficiency of the FLICA algorithm for extremely large population datasets. BigFLICA is scalable for simultaneously analyzing all the multimodal data of the full 100,000-subjects UKB dataset using only a modest computing cluster (**Fig. 2.1**).

We first demonstrate the effectiveness of our approach through extensive simulations. Then, in real data, we quantify performance in terms of the prediction accuracy of non-imaging-derived phenotypes (nIDPs) [Liégeois et al., 2019, Kong et al., 2018], such as health outcome measures. Using voxel-level imaging data of 81 modalities from 1,003 subjects in the HCP and 47 modalities from 14,053 subjects in the UKB, we show that BigFLICA can perform comparably with original FLICA [Groves et al., 2011] in terms of the prediction accuracy for nIDPs (158 in HCP and 8,787 in UKB). Most importantly, we systematically investigated whether there are benefits to jointly fusing multimodal data together, instead of analysing them separately. We show that significant improvements in the prediction accuracy of nIDPs are found when comparing a high-dimensional BigFLICA with other widely-used data analysis strategies: (1) doing single-

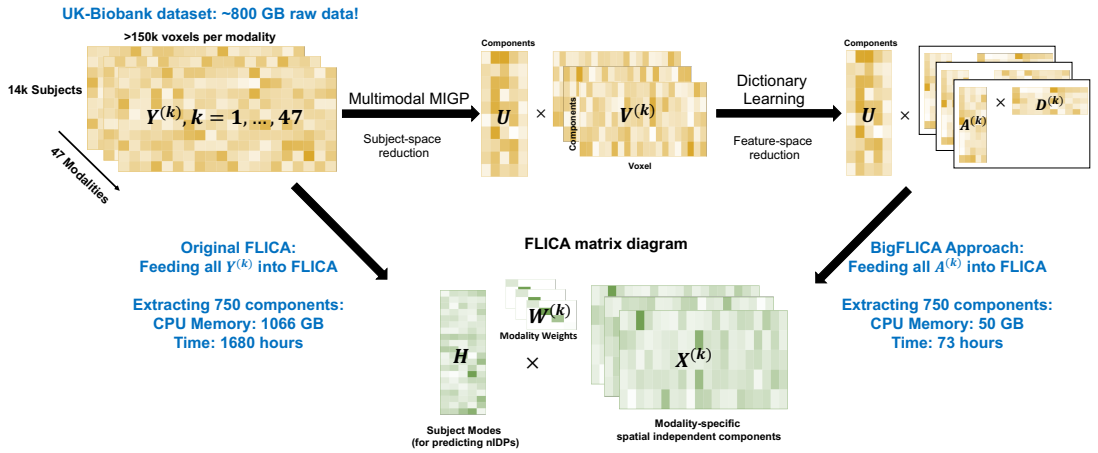



**Figure 2.1: Overview of the proposed approach for jointly analyzing a biobank-scale multimodal neuroimaging dataset.** Currently for the UKB dataset (voxel-level data, 14503 subjects, 47 modalities), the total data size is approximately 800 GB, and if we directly feed these data into FLICA and extract 750 components, we will need approximately 1066 GB CPU memory and 1680 hours computation time. Our new approach, BigFLICA, used multimodal MIGP and dictionary learning to preprocess the multimodal data; this is efficient and memory friendly, and much of this preprocessing can be easily parallelized. BigFLICA only used 50 GB memory and 73 hours to analyze the same dataset using a 24-core compute server.

modality ICA and concatenating the results across modalities and (2) using existing IDPs (5,812 in HCP and 3,913 in UKB). In particular, the improvements in prediction of many health outcome and cognitive variables are large, more than doubling prediction accuracy for some variables. Furthermore, we investigate the relationship between modes derived by BigFLICA and IDPs. We find that although the modes were estimated from the same set of voxel-level data, they have complementary information which can be combined together to further increase the prediction accuracy of nIDPs. Finally, we applied BigFLICA to analyze the UKB data and extracted 750 components. Existing multimodal ICA cannot estimate this many modes from this many subjects. We found several interpretable associations between modes of BigFLICA and nIDPs, including modes that relate to *fluid intelligence*, *handedness*, *age started wearing glasses or contact lenses* and *hypertension*. In many cases BigFLICA can find associations with nIDPs with greater statistical sensitivity than was possible with IDPs. Overall, BigFLICA demonstrated the advantages of data-driven joint multimodal modelling in the analysis of biobank-scale multimodal datasets.

## 2.2 Methods

### 2.2.1 Brief overview of the proposed approach: BigFLICA.

FLICA [Groves et al., 2011] is a Bayesian ICA approach for multimodal data fusion. The input of FLICA is $K$ modalities' data matrices $Y^{(k)}$ with dimensions $N \times P_k, k = 1,\ldots,K$, where $P_k$ is the number of features (e.g., voxels) and $N$ is the number of subjects. FLICA aims to find a joint $L$-dimensional decomposition of all $Y^{(k)}$: $Y^{(k)} = HW^{(k)}X^{(k)} + E^{(k)}$, where $H_{(N\times L)}$ is the shared subject mode (mixing matrix) across modalities (a vector of subject weights for each mode), so is a 'link' across different modalities, $W^{(k)}_{(L\times L)}$ is a positive diagonal mode-weights matrix (one overall weight per modality per mode), $X^{(k)}_{(L\times P_k)}$ is the independent (spatial) feature maps for the $L$ components of a modality (one map per modality per mode), and $E^{(k)}_{(N\times P_k)}$ is the modality-specific Gaussian noise term (**Fig. 2.1**). We propose two efficient approaches that can either be used separately or combined together to reduce the size of the original data matrices, and therefore reduce the computational load of the original FLICA. An overview of BigFLICA is shown in **Fig. 2.1**.

The first approach, termed multimodal extension of MELODIC's Incremental Group Principal component analysis [Smith et al., 2014] (mMIGP), aims to reduce the subject dimension to a linear combination of the original subjects. mMIGP is a time- and memory-efficient approximation of principal component analysis (PCA) on feature-concatenated multimodal data. To this end, if we aim to get a $L^\star$ decomposition, we first apply MIGP [Smith et al., 2014] separately within each modality to estimate $\widetilde{U}^{(k)}_{(N\times L^\star)}$, which is an approximation of an $L^\star$-dimensional PCA decomposition of one modality $Y^{(k)}$. This step can be done in parallel across modalities. Then, we concatenate all $\widetilde{U}^{(k)}$ in the component dimension and apply another MIGP to get $U_{(N\times L^\star)}$, which is an $L^\star$-dimensional approximate PCA decomposition of all modalities together. Finally, we project the original data of each modality $Y^{(k)}$ to the PCA-reduced space using $U$. If no further reduction (e.g., dictionary learning as below) is to be applied, the data that could then be fed into the core FLICA would be the $K$ component-by-feature matrices $V^{(k)}$ of size $L^\star \times P_k$, and FLICA would then extract $L$ ($L < L^\star$) components from these

(**Methods**). This step adds little computational cost compared with the original FLICA, because a similar PCA step is needed to initialize the parameters of the original FLICA, but this approach is feasible for large numbers of subjects and modalities. Although different modalities usually have different overall signal-to-noise ratios (SNR), which is largely ignored by this mMIGP step, the subsequent FLICA can take this into account by the modality-specific noise terms, and a high-dimensional mMIGP is used to capture modes with even small variations in each modality.

It is known that voxels are correlated in both a local fashion (local spatial autocorrelation) and across brain networks (long range correlation); hence, effective feature subsampling could hope to capture all important information in the data but also reduce the cost of spatial modelling in FLICA [Hoyos-Idrobo et al., 2019]. Therefore, we incorporate an approach, termed sparse online Dictionary Learning [Mairal et al., 2010] (DicL), to reduce the dimension of feature (e.g., voxel) space that can capture both local and distant spatial correlation structure. Specifically, for each modality, we use DicL to model the $V^{(k)}$ as a sparse linear combination of $L^{\star\star}$ basis elements: $V^{(k)} = A^{(k)}D^{(k)}$, where $D^{(k)}$ is the *sparse spatial dictionary basis*, and $A^{(k)}$ is the *feature loadings*. By minimizing the reconstruction error, and enforcing sparsity in the dictionary basis $D^{(k)}$, we aim to achieve an optimal subsampling of feature space. The inputs of FLICA are then $K$ smaller matrices $A^{(k)}$, which are only of dimension $L^{\star} \times L^{\star\star}$, and FLICA then extracts $L$ ($L < \min(L^{\star}, L^{\star\star})$) components from these (**Methods**). Compared with doing FLICA with the original $K$ large $N \times P_k$ matrices, using the DicL preprocessed data can greatly reduce the computation load of FLICA. DicL can easily be parallelized across modalities and is memory friendly, which further increases efficiency (**Fig. 2.1**).

### 2.2.2 FLICA model.

The input to FLICA is $K$ modalities' data matrices $Y^{(k)}$ with each modality's dimensions being $N \times P_k, k = 1, \ldots, K$, where $P_k$ is the number of features (e.g., voxels) in modality $k$ and $N$ is the number of subjects. FLICA aims to find a joint $L$-dimensional decomposition of all $Y^{(k)}$:

$$Y^{(k)} = HW^{(k)}X^{(k)} + E^{(k)},\ k = 1, \ldots, K \tag{2.1}$$

where $H_{(N\times L)}$ is the shared subject mode (mixing matrix) across modalities, so is a 'link' across different modalities, $W^{(k)}_{(L\times L)}$ is a positive diagonal mode-weights matrix, $X^{(k)}_{(L\times P_k)}$ is the independent (spatial) feature maps for the $L$ components of modality $k$, and $E^{(k)}_{(N\times P_k)}$ is the Gaussian noise term.

### 2.2.3 Multimodal extension of MELODIC's Incremental Group Principal component analysis for subject-space dimension reduction.

We propose a multimodal extension of our previous MIGP approach [Smith et al., 2014], termed mMIGP, to reduce the subject dimension of multimodal data. MIGP has been extensively validated in simulations and real neuroimaging data for finding an approximate PCA decomposition in a time- and memory-efficient way [Smith et al., 2014]. Suppose that our multimodal data are $K$ matrices $Y^{(k)}, k = 1,\ldots,K$ with dimensions $N \times P_k$, where $N$ is the number of subjects and $P_k$ is the number of features (e.g. voxels) in a modality. In mMIGP, each feature is z-score normalized first. Then, an MIGP is applied to each modality separately to find an $L^\star$-dimensional approximate PCA decomposition. Specifically, we want to find an approximation of a singular value decomposition (SVD) of each $Y^{(k)}$:

$$Y^{(k)} = \widetilde{U}^{(k)}\widetilde{S}^{(k)}(\widetilde{V}^{(k)})^\tau,\ k = 1,\ldots,K \tag{2.2}$$

where $\widetilde{U}^{(k)}_{(N\times L^\star)}$ and $\widetilde{V}^{(k)}_{(P_k\times L^\star)}$ are the left and right singular vectors, while $\widetilde{S}^{(k)}_{(L^\star\times L^\star)}$ are the singular values. A naive SVD on $Y^{(k)}$ scales quadratically with $N$, which is not efficient when $N$ is large. To find the approximation, MIGP sequentially feeds a subset of (columns of) $Y^{(k)}$ into an SVD, so that these subsets are reduced to a low-dimensional representation. The low-dimensional representation is then concatenated with another subset of (columns of) $Y^{(k)}$, and is fed into another SVD to find the low-dimensional representation of them. The final SVD approximation is found after one pass of all data. The computational complexity of MIGP scales linearly with $N$. For a detailed description, please see Appendix A of the MIGP paper [Smith et al., 2014].

The third step is to concatenate all $\widetilde{U}^{(k)}$ in the component dimension and apply another MIGP for finding a $L^\star$-dimensional approximate PCA decompositions $U$ of size $N \times L^\star$,

which is a low-dimensional representation of multimodal data in the analysis. Finally, the z-score normalized data $Y^{(k)}$ of each modality is projected onto the $U$ by:

$$V^{(k)} = U^{\tau} Y^{(k)} \ , k = 1, 2, \ldots, K \tag{2.3}$$

the $V^{(k)}, k = 1, 2, \ldots, K$ are the inputs of the subsequent FLICA algorithm. Therefore, the total size of data output by this stage is $L^{\star} \sum_{k=1}^{K} P_k$, which is smaller than the original input size $N \sum_{k=1}^{K} P_k$. The fractional reduced data size is $L^{\star}/N$, and the $L^{\star}$ can be fixed when more subjects are introduced, so it is scalable in the big-data analysis. In practice, we usually choose $L^{\star}$ based on the percentage of explained variance of SVD in the third step.

If we feed $V^{(k)}, k = 1, 2, \ldots, K$ into FLICA to estimate $L^{\star}$ FLICA modes, the output subject mode matrix $H^{\star}$ is of the size $L^{\star} \times L$, so we then simply multiply this by $U$ to get the final subject-mode matrix:

$$H = U H^{\star} \tag{2.4}$$

The mMIGP approach is equivalent to performing an approximate PCA on feature-concatenated data. The advantage is that it does not need to fit all data into the memory, and even can be parallelized across modalities [Smith et al., 2014]. This approach is also equivalent to applying mCCA across all modalities [Parra, 2018].

### 2.2.4 Sparse dictionary learning for voxel-space dimension reduction.

If the resolution of the data is high and the number of modalities is large, applying just the mMIGP reduction still leaves FLICA as being memory and computationally expensive. Therefore, we propose a method that can effectively reduce the voxel dimension, and preserve the important spatial information for subsequent FLICA spatial modelling. Although the most obvious ways of voxel subsampling are either to apply regular spatial downsampling (similarly, local voxel clustering) or apply PCA *within each modality*, the former only focuses on the local patterns [Hoyos-Idrobo et al., 2019] (and does not adapt downsampling to local variations in redundant information across voxels) while the later empirically finds more global and noise patterns in neuroimaging data, and does not

work at all well empirically in this context (see also Allen, et al. [Allen et al., 2014] and references therein).

The method we used here is sparse Dictionary Learning (DicL) [Mairal et al., 2010], which effectively performs 'voxel grouping' in both local and global fashion. It can be used directly on each of the original z-score normalized modalities, i.e., $Y^{(k)}, k = 1,2,\ldots,K$, or on the mMIGP reduced data, i.e., $V^{(k)}, k = 1,2,\ldots,K$. Taking the former as an example, the sparse DicL is adopted here:

$$Y^{(k)} = A^{(k)} D^{(k)}\ , k = 1,2,\ldots,K \tag{2.5}$$

where $D^{(k)}$ is *sparse spatial dictionary basis*, and $A^{(k)}$ is the *feature loadings* with each column representing a linear combination of information from a group of voxels which might either be a local cluster or spatially distributed network. By minimizing an $l_1$-regularized sparse-coding objective function, a local optimal solution can be obtained:

$$\begin{aligned} & l(A^{(k)}, D^{(k)}) = \sum_{i=1}^{P_k} \|Y_i^{(k)} - A^{(k)} D_i^{(k)}\|_F^2 + \lambda \|D_i^{(k)}\|_1 \\ & s.t.\ \mathscr{C} = \{A^{(k)} \in \mathbb{R}^{N \times L^{\star\star}} s.t.\ \forall j = 1,\ldots,L^{\star\star}, (A_j^{(k)})^{\tau} A_j^{(k)} \leqslant 1\} \end{aligned} \tag{2.6}$$

where subscript $i$ represents the $i$-th column of the corresponding matrix, and $\lambda$ is a regularization parameter, $\|\|_F$ is the frobenius norm. The $l_1$-regularization term enforces that the learned spatial loadings $D^{(k)}$ are sparse. The objective function can be efficiently optimized by a block-coordinate descent optimizer with warm restarts. It has been implemented in the SPAMS package (`http://spams-devel.gforge.inria.fr/`). Compared with simply using PCA in this step, sparse DicL has three advantages: (1) the spatial loading matrix $D^{(k)}$ can be sparse, so a smaller number of voxels are involved in each column of the dictionary; (2) the columns of the dictionary do not need to be orthogonal to each other, which is more flexible; (3) an "overcomplete" dictionary is allowed, i.e., the number of dictionary basis vectors can exceed the minimum of $N$ and $P_k$, which further increases the flexibility.

After the above modality-wise DicL, the final inputs to FLICA are the matrices $A^{(k)}, k = 1,2,\ldots,K$, of size $N \times L^{\star\star}$ if we use $Y^{(k)}, k = 1,2,\ldots,K$, or $L^{\star} \times L^{\star\star}$ if we use $V^{(k)}, k = 1,2,\ldots,K$. Note that (unlike the typical approach of feeding spatial

PCA eigenvectors into ICA) we are not feeding the spatial dictionary basis ($D^{(k)}$) into the FLICA core modelling, but the feature loadings ($A^{(k)}$). We could change the order by applying DicL first and then mMIGP, but this empirically has a lower computation efficiency.

To estimate the spatial loading matrices from FLICA, we do voxel-wise multiple regression, where the target variable is a voxel's feature values and the design matrix is the BigFLICA's subject mode, so that the regression coefficient (an L-by-1 vector) is the resulting spatial loadings for this voxels across components.

### 2.2.5 Evaluation of BigFLICA in simulations.

We simulated 500 subjects, and each had two modalities, which were both $30 \times 30 \times 30$ images. We first simulated $K$ ground-truth (independent) spatial maps $X$; each of these was a $30 \times 30 \times 30$ image. The spatial maps were a weighted sum of two Gaussian white noise images, where the first one was $30 \times 30 \times 30$ with weight 0.05, and the second was a $5 \times 5 \times 5$ cube randomly located in the full image with weight 0.95. Then, random positive component weights $W$, Gaussian random subject loadings $H$ and Gaussian white noise terms $E$ were simulated. Finally, after vectorizing each spatial map and noise term, the data for a single modality $Y$ was generated as $Y = HWX + \sigma E$, where $\sigma$ was a parameter to control the signal-to-noise ratio (SNR). A small amount of spatial smoothing using a Gaussian kernel was applied to spatial maps $X$ and noise terms $E$ to mimic real image data. Each of the two modalities also had 5 unique spatial maps that were not shared by each other. The voxels were z-score normalized before feeding into the subsequent FLICA analysis. The SNR was defined as: $var(vec(HWX))/var(vec(\sigma E))$.

Performance evaluation: When FLICA was applied to the simulated data, the number of independent components was always set to the ground truth $K$. The performance was measured by the similarity between estimated subject-mode matrix $H^\star$ and the ground truth $H$. The similarity was measured by the greedy matching of the components based on maximum correlation and then estimating the mean correlations across components. Here, greedy matching means we match the pair with the highest correlation and then match the second pair with the second highest correlation, etc.

Evaluation of mMIGP for subject-space dimension reduction: After generating simulated data, we reduced the data to varying dimensions ($L^{\star} = 50, 100, 200, 300, 400$) using mMIGP, and then fed the reduced data into FLICA. This was compared with the original FLICA. The number of ground-truth components was set to $25, 35, 45$ and the SNR was set to $4, 1, 0.25, 0.04$. All simulations were repeated 50 times.

Evaluation of DicL for voxel-space dimension reduction: To evaluate the influence of the DicL parameters on the subsequent FLICA results, we performed the DicL on simulated data using varying parameter combinations ($\lambda = 0.1$ to 16 and $L^{\star\star} = 100$ to 3000) followed by FLICA (nIC $= 25, 50, 100$). This was compared with the original FLICA. The SNR was set to $4, 1, 0.25, 0.04$, and the number of iterations for the DicL was set to 50, because we empirically find that this number of iterations is sufficient for DicL to converge to a stable result in simulation and real data. All simulations were repeated 50 times.

### 2.2.6 HCP and UK Biobank data.

The voxel/vertex-wise neuroimaging data of 81 different modalities of 1,003 subjects from the HCP S1200 data release were used in this chapter [Van Essen et al., 2013]. The preprocessing was conducted by the HCP team using an optimized pipeline [Glasser et al., 2013]. The 81 modalities included (1) 25 resting-state ICA dual-regression spatial maps (z-score normalized); (2) 47 unique task contrast maps as z-statistics from 7 different fMRI tasks; (3) 3 T1-image derived modalities (grey matter volume, surface area, surface thickness); (4) 6 Tract-Based Spatial Statistics (TBSS) features from diffusion MRI (FA, L1, L2, L3, MD, MO) [Smith et al., 2006]. In addition, 158 nIDPs were used here, which was the same as our previous study [Smith et al., 2015].

The UK Biobank imaging data were mainly preprocessed by FSL [Smith et al., 2004, Jenkinson et al., 2012] and FreeSurfer [Fischl, 2012] following an optimized pipeline [Alfaro-Almagro et al., 2018] (`https://www.fmrib.ox.ac.uk/ukbiobank/`). The voxel-wise neuroimaging data of 47 modalities of 14,053 subjects were used in this chapter, including: (1) 25 "modalities" from the resting-state fMRI ICA dual-regression spatial maps (z-score normalized); (2) 6 modalities from the emotion task fMRI: 3 contrasts

(shapes, faces, faces>shapes) of z-statistics and 3 *contrasts of parameter estimate* maps; (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); (4) 4 T1-MRI derived modalities (grey matter volume and Jacobian 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]). A detailed description is in **Table. A.6**. In addition, the 8,787 nIDPs were included, but we retained the 7,245 of those that have at least 1,000 non-missing values (subjects).

When carrying out nIDP prediction, a total of 13 and 54 confounding variables were regressed out from nIDPs and imaging data using linear regression in the HCP and the UKB datasets respectively. Subjects with a missing modality were imputed by the mean value of all other subjects. We did not impute the missing nIDPs. The names of confounding variables are provided in the Appendix.

### 2.2.7 Comparing BigFLICA with the original FLICA on real data.

On real data, we do not know the ground truth components, and the data may not follow the assumptions of ICA. Therefore, we rely on the performance of predicting nIDPs as a surrogate criterion to evaluate different methods. We applied the proposed mMIGP approach to HCP data and a subset of 1,036 UKB subjects (so that the original FLICA is computationally tractable). Elastic-net regression, from the glmnet package [Zou and Hastie, 2005], was used to predict the nIDPs using FLICA's subject modes as model regressors (features). This approach is widely-used and has been shown to achieve a robust and state-of-the-art performance in many neuroimaging studies [Cui and Gong, 2018, Jollans et al., 2019]. To evaluate the model performance, for each nIDP, we used 5-fold cross validation, and computed Pearson correlation between the predicted and true values of each nIDP across the 5 test sets. As there are tuning parameters within the Elastic-net regression, in each training set, we performed a nested

5-fold cross validation to tune the model parameters, and used the best model selected in the nested 5-fold cross validation to do the prediction in the test set. When comparing any two approaches, the same training-validation-testing split was used. The prediction accuracy was quantified as the Pearson correlation between predicted and the true values of each nIDP in the test sets.

To evaluate MIGP preprocessing, we reduced the dimension to varying $L^\star$ (from 100 to 500) using MIGP first and then used FLICA to extract $L = 50$ components. The original FLICA was also applied to extract 50 components. To evaluate DicL preprocessing, we used the DicL (dictionary dimension = 2000 and sparsity parameter $\lambda = 1$) to reduce the data dimension of each modality followed by the FLICA to extract varying numbers of components (nIC$= 25, 50, 100, 200, 300$). The original FLICA was also applied to extract the same numbers of components. The prediction accuracy of BigFLICA was compared with the original FLICA applied on non-reduced data.

### 2.2.8 Statistical significance of difference of prediction accuracy between two approaches.

To compare the overall prediction accuracy of two approaches (e.g., BigFLICA with mMIGP preprocessing vs. the original FLICA), we estimate the statistical significance of the difference between the prediction correlations across nIDPs. Starting with a total of $p$ nIDPs, we first exclude nIDPs where both methods have low prediction accuracy ($r < 0.1$, as these would likely just add noise to the comparison), resulting in $p_1$ nIDPs. Then we want to test whether the overall prediction accuracy of $p_1$ nIDPs with one method is significantly higher than another method. A naive approach would be a simple paired t-test, but the correlation structures among nIDPs makes the samples dependent with each other, so that the p-value of a t-test is not valid. Note that the paired t-test can also be formulated as a linear regression model, and in the framework of linear regression, sample correlation can be taken into account by the generalized least squares approach. Specifically, we assume: $y = x\beta + e$, where $y$ is the difference of the prediction accuracy between two methods and $x$ is a column of ones. The t statistic for coefficient $\beta$ can

be calculated as [Kariya and Kurata, 2004]:

$$\beta = (x'V^{-1}x)^{-1}x'V^{-1}y$$
$$s = (x'V^{-1}x)^{-1}y'(V^{-1} - V^{-1}x(x'V^{-1}x)^{-1}x'V^{-1})y/(p_1 - 1)$$
$$t = \beta/\sqrt{\text{diag}(s)}$$

Where $'$ means transpose, and $V$ is the sample covariance matrix, i.e., the covariance among nIDPs. Note that when $V = I$, the model is equivalent to a paired t-test. In general we do not know $V$, but can obtain a good estimate of $V$ by calculating the covariance of nIDPs using the nIDPs-by-subject matrix. We used the lscov function in Matlab to perform the above estimation.

### 2.2.9 Parameter settings of running BigFLICA in the full HCP and UKB datasets.

We applied BigFLICA approach to extract a varying number of target components in two datasets. In HCP, we used FLICA with DicL preprocessing only (dictionary dimension 2000 and $\lambda = 1$). In UKB, we used FLICA with both mMIGP and DicL preprocessing (dictionary dimension 5000, $\lambda = 1$ and mMIGP dimension 1000 (>95% explained variance)). The number of FLICA VB iterations is 1000.

### 2.2.10 Comparing BigFLICA with multiple independent single-modality ICA decomposition.

ICA is a widely-used approach for decomposing single-modality neuroimaging data, including functional MRI [Smith et al., 2015] but also in structural MRI [Zeighami et al., 2015] and diffusion MRI [Li et al., 2012]. A natural question arises whether BigFLICA is able to combine multimodal information more effectively than the single-modality approaches such as ICA (we used the fastICA algorithm [Hyvarinen, 1999]), which ignores inter-modality relationships.

We first performed ICA on each modality of HCP and UKB data separately to extract 25,100, 250, 500 and 750 components. For a given component number, we built a prediction model using the concatenated ICA subject modes (across modalities) to predict each of the nIDPs. To be fair, for BigFLICA, we extract the same number of

ICs to build the prediction model. For example, in the UKB data and a 25-dimensional decomposition, the predictor is a Subject$\times(25 \times 47)$ matrix for single-modality ICA, where 25 is the number of components in each modality and 47 is the total number of modalities. For BigFLICA, the predictor is a Subject$\times$25 matrix. This is arguably a fair comparison because each of the BigFLICA modes potentially contains information from all modalities. The method to build a predictive model and evaluate this is the same as above, except that when we used the concatenated ICA subject modes, we added a univariate screening step in the training set to select the top 300 most informative features according to their correlation with an nIDP in the training set. This step, in general, boosts the predictive accuracy because the dimensionality of concatenated ICA modes is usually very high, so that many of the modes are pure noise with respect to any given nIDP. Therefore, the univariate screening can help the elastic-net regression to filter out noisy features effectively. We did not perform univariate screening when using the BigFLICA subject modes to predict nIDPs.

### 2.2.11 Comparing BigFLICA with hand-curated imaging-derived phenotypes.

A popular choice of data analysis strategy is to extract imaging features based on expert knowledge (e.g., regional volumes and thickness, and resting-state functional connectivities between brain regions), often referred to as IDPs [Miller et al., 2016]. Brain IDPs have been shown to genetically correlate with many SNPs in our previous genome-wide association study (GWAS) in UK Biobank [Elliott et al., 2018], and they have been shown to change in many psychiatric diseases [Kelly et al., 2017, Van Rooij et al., 2017, Hibar et al., 2018].

We extracted 5,812 IDPs from the HCP, including (1) 199 structural MRI features from Freesurfer as provided by the HCP; (2) 4700 regional mean task activations from 47 independent task contrasts using a 100-dimensional parcellation atlas [Schaefer et al., 2017]; (3) 625 functional connectivities (FCs) based on a 25-dimensional ICA parcellation with partial correlation to estimate FCs; (4) 288 regional mean TBSS features (FA, L1, L2, L3, MD, MO) using the Johns Hopkins University tract atlas.

We used 3,913 IDPs from UKB, including global and local features from the 6 imaging modalities (T1, T2-FLAIR, swMRI, tfMRI, rfMRI, and dMRI) [Smith et al., 2020].

We built prediction models that use IDPs or BigFLICA modes to predict each of the nIDPs using the same strategy as above. The FLICA dimension is set to 25, 100, 250, 500, 750. In addition, we also concatenated IDPs and each of the BigFLICA subject modes together to predict the nIDPs, and the performance is compared with using IDPs alone. We used a univariate screening step to select the top 300/500 most informative IDPs according to their correlation with an nIDP in the inner-fold (i.e., training set) of HCP/UKB. Finally, we also built models that use IDPs to predict each of the FLICA subject modes and vice versa, aiming to evaluate the shared variances between features extracted by these two different approaches in the same data.

## 2.3 Results

### 2.3.1 Evaluation of BigFLICA in simulations.

We first applied BigFLICA on simulated data to evaluate the performance of mMIGP and DicL as data preprocessing approaches under different parameter settings and data signal-to-noise ratios. The mean correlation of extracted components with simulated ground truth was compared with the corresponding result from the original FLICA (**Methods section 2.2.5**).

For mMIGP, **Fig. 2.2a** shows that, in most of the situations, the BigFLICA with mMIGP preprocessing gave similar results to the original FLICA, and both FLICA and BigFLICA accurately find the underlying ground truth in most cases. This is in agreement with results of simulations in the MIGP paper [Smith et al., 2014] that it can accurately approximate a full-data PCA in different situations. The optimal dimension of mMIGP is different among simulations; sometimes a relative low dimension can achieve an accurate estimation of components (e.g. **Fig. 2.2a** first three columns), while in other cases a high dimension is needed (e.g. **Fig. 2.2a** the fourth column).

For DicL, **Fig. 2.2b** shows that in almost all circumstances: (1) increasing the dictionary dimensions will boost the performance of subsequent FLICA analysis; (2)

**Table 2.1: Comparison of computation time and amount of RAM usage of BigFLICA with the original FLICA in the UKB dataset (14,503 subjects, 47 different modalities).** BigFLICA greatly increases computational efficiency in different settings. Both BigFLICA and FLICA were run on the same computer using all 24 cores in all computation stages with Intel Xeon CPU E7-8857 v2 @ 3.00GHz and 2TB RAM.

| Approaches | | Number of components | | | | | 100K subjects |
|---|---|---|---|---|---|---|---|
| | | nIC=25 | nIC=100 | nIC=250 | nIC=500 | nIC=750 | 750 components (estimated) |
| Computation Time (hours) | The original FLICA | 160 h | 300 h | 580 h | 1,020 h | 1,680 h | 12,000 h |
| | BigFLICA (mMIGP preprocessing) | 23 h | 54 h | 135 h | 315 h | 565 h | 630 h |
| | BigFLICA (mMIGP+DicL preprocessing) | 52 h | 53 h | 58 h | 65 h | 73 h | 120 h |
| Peak RAM (GB) | The original FLICA | 801 GB | 821 GB | 879 GB | 963 GB | 1,066 GB | 6,000 GB |
| | BigFLICA (mMIGP preprocessing) | 66 GB | 88 GB | 136 GB | 215 GB | 297 GB | 297 GB |
| | BigFLICA (mMIGP+DicL preprocessing) | 50 GB | 50 GB | 50 GB | 50 GB | 50 GB | 50 GB |

the optimal sparsity parameters are usually between $\lambda = 0.5$ to 2, and they have similar performance; (3) In most cases the optimal performance given by DicL matches that of non-reduced analysis (noted in figure legends). Therefore, in the real data analysis, when using the DicL approach, we always use a very high dimensional DicL decomposition and fix the sparsity parameter to $\lambda = 1$.

### 2.3.2 Computation time comparison.

**Table 2.1** shows the comparison of the computation time and memory requirement of BigFLICA with the original FLICA in the UKB dataset. All code was implemented in Python 2.7, and both BigFLICA and FLICA were run using 24 cores on a single compute node with Intel Xeon CPU E7-8857 v2 @ 3.00GHz CPU and 2048 GB RAM. The computation time includes: (1) Preprocessing of data using mMIGP and DicL (BigFLICA only); (2) Initialization of FLICA parameters; (3) FLICA VB parameter updates. For the 100,000-subjects data, BigFLICA greatly decreases the computation time and memory usage from an unrealistic amount to a modest configuration for a modern high performance computing cluster, which therefore allows for the possibility of data-driven population phenotype discovery.

### 2.3.3 Real data: Comparing BigFLICA with the original FLICA based on the prediction accuracy of nIDPs.

As there is no ground truth available, we tested whether modes of BigFLICA have a similar prediction accuracy of nIDPs compared with the original FLICA, using data from

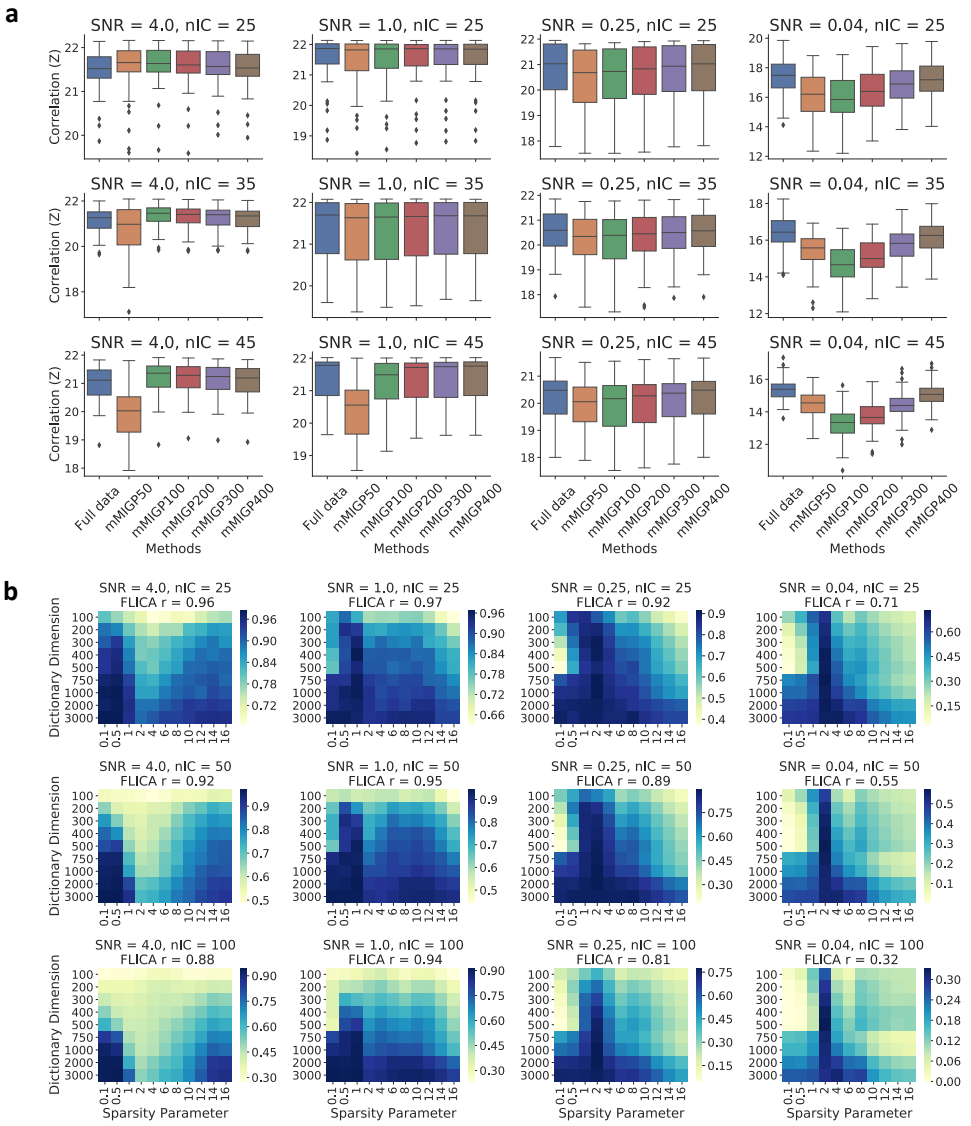



**Figure 2.2: Evaluation of multimodal extension of MIGP (mMIGP) and dictionary learning (DicL) as the data preprocessing steps for the FLICA using simulations.** BigFLICA achieves similar performance as compared with original FLICA that uses the full data. a, Evaluation of mMIGP preprocessing. We compared the correlations (Z-transformed) of extracted components with ground truth across 50 simulations using the original FLICA (the left column of each figure) and the mMIGP preprocessed FLICA (other columns). The mMIGP dimensions vary between 50 and 400; the SNRs are between 4 and 0.04 (left to right), and the number of FLICA and ground truth components are 25, 35, 45 (top to bottom). As there are 500 subjects, the reduction factor is from 10 to 1.25. b, Evaluation of DicL preprocessing. We compared the correlations of extracted components with ground truth using the original FLICA (FLICA results given in the titles of each figure) and the DicL preprocessed FLICA with different sparsity parameters and dictionary dimensions (cells of the heatmaps). The SNRs are between 4 and 0.04 (left to right), and the number of FLICA and ground truth components are 25, 50, 100 (top to bottom). As there are 27,000 original features per modality, the reduction factor is from 270 to 9.

the HCP, and a subset of 1,036 subjects from the UKB (**Methods section 2.2.7**). Elastic-net regression with nested 5-fold cross-validation was used to predict each of the nIDPs.

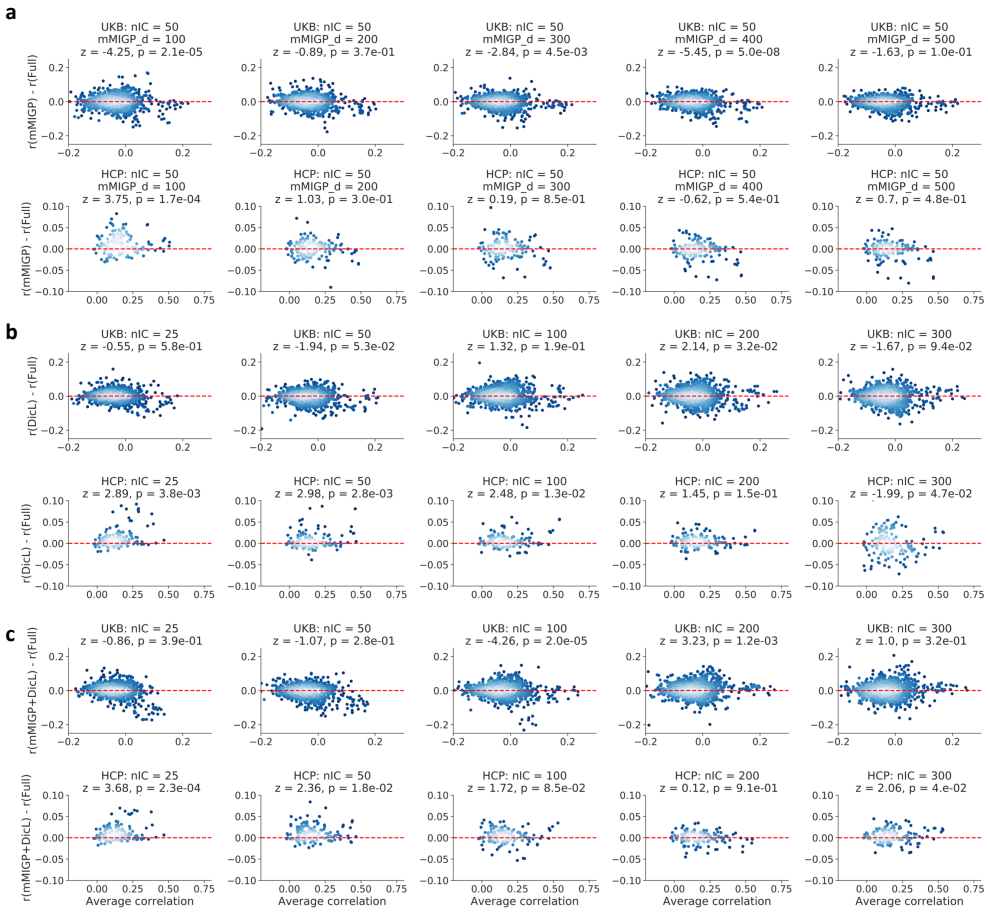



**Figure 2.3: Comparison of prediction accuracy of nIDPs between BigFLICA and the original FLICA.** Overall, for most of the comparisons, the differences of prediction accuracy are not significant. In each of the Bland–Altman plots, each point represents the prediction of one nIDP, where the x-axis is the average prediction correlation of the two approaches while the y-axis is the difference, i.e., BigFLICA - FLICA. The z- and p-values in the titles reflected the statistical significance of the differences. The Bonferroni correction 0.05 threshold corresponds to a raw p-value of 1.7e-3. a, Comparing FLICA with mMIGP preprocessing with the original FLICA. We used a subset of 1,036 subjects in the UKB dataset (top) and the HCP (bottom). The number of estimated FLICA components is set to 50, and mMIGP dimensions are set from 100 to 500. b, Comparing FLICA with DicL preprocessing with the original FLICA. We used a subset of 1,036 subjects in the UKB dataset (top) and the HCP (bottom). The dictionary dimension is set to a high value of 2000, and the sparsity parameter is set to $\lambda = 1$ for all modalities. The number of estimated FLICA components are set from 25 to 300. c, Comparing FLICA with both mMIGP and DicL preprocessing combined, with the original FLICA. The mMIGP dimension is set to 500, and other settings are the same as in b. We use only a subset of UKB here so that running the original FLICA is computationally feasible. The lighter the blue, the higher the density of points.

This approach is widely-used and has been shown to achieve a robust and state-of-the-art performance in many neuroimaging studies [Cui and Gong, 2018, Jollans et al., 2019]. Pearson correlation between each of the predicted and the true nIDPs in the outer test fold is used to quantify accuracy. The statistical significance of differences of prediction accuracy between two approaches are estimated by a weighted paired t-test approach.

(**Methods section 2.2.8**).

**Fig. 2.3** shows the Bland-Altman plots comparing the prediction accuracy of nIDPs between original FLICA and BigFLICA with mMIGP preprocessing only (**Fig. 2.3a**), and with DicL preprocessing only (**Fig. 2.3b**), and with both data reduction approaches (**Fig. 2.3c**), in the UKB and HCP datasets. In these comparisons, mMIGP reduced the data to approximately 1/10 to 1/2 of the original data size, and DicL reduced data to approximately 1/75 of the original data size. Overall, BigFLICA can estimate similar sets of modes with comparable prediction accuracy in real multimodal neuroimaging data, i.e., the difference of the correlation between two methods is centered around zero across a wide range of mean correlation values (which are also reflected in the insignificant p-values of weighted paired t-test), which demonstrates that the mMIGP and DicL approaches are effective to reduce data and preserve key information in the data.

Finally, we tested whether the set of nIDPs which were predicted better with BigFLICA than the original FLICA are relatively consistent across different number of components. We calculated the differences of the prediction accuracy between BigFLICA and original FLICA, and then tested if they are correlated across different nICs (**Fig. A.1**). However, we did not observe a significant correlation. This might mean that the nIDP-related information in extracted imaging features is different at different scales (ICA dimensionalities), but as the space spanned by the imaging features is increased with increasing dimensionality, it seems more likely that the reduction in prediction of some nIDPs with increasing ICA dimensionality is a result of overfitting.

### 2.3.4 Comparing BigFLICA with multiple independent single-modality ICA decomposition.

We also compared BigFLICA outputs against features pooled across those from separate ICA processing of each modality (**Methods section 2.2.10**). **Fig. 2.4a** shows that BigFLICA has a worse prediction performance than via running ICA separately on each modality when the dimensionality $L$ is low. This is because at low dimensional decomposition, single-modality ICA is most efficient because the constraints imposed on the degrees-of-freedom implied in the FLICA model is insufficient to capture the

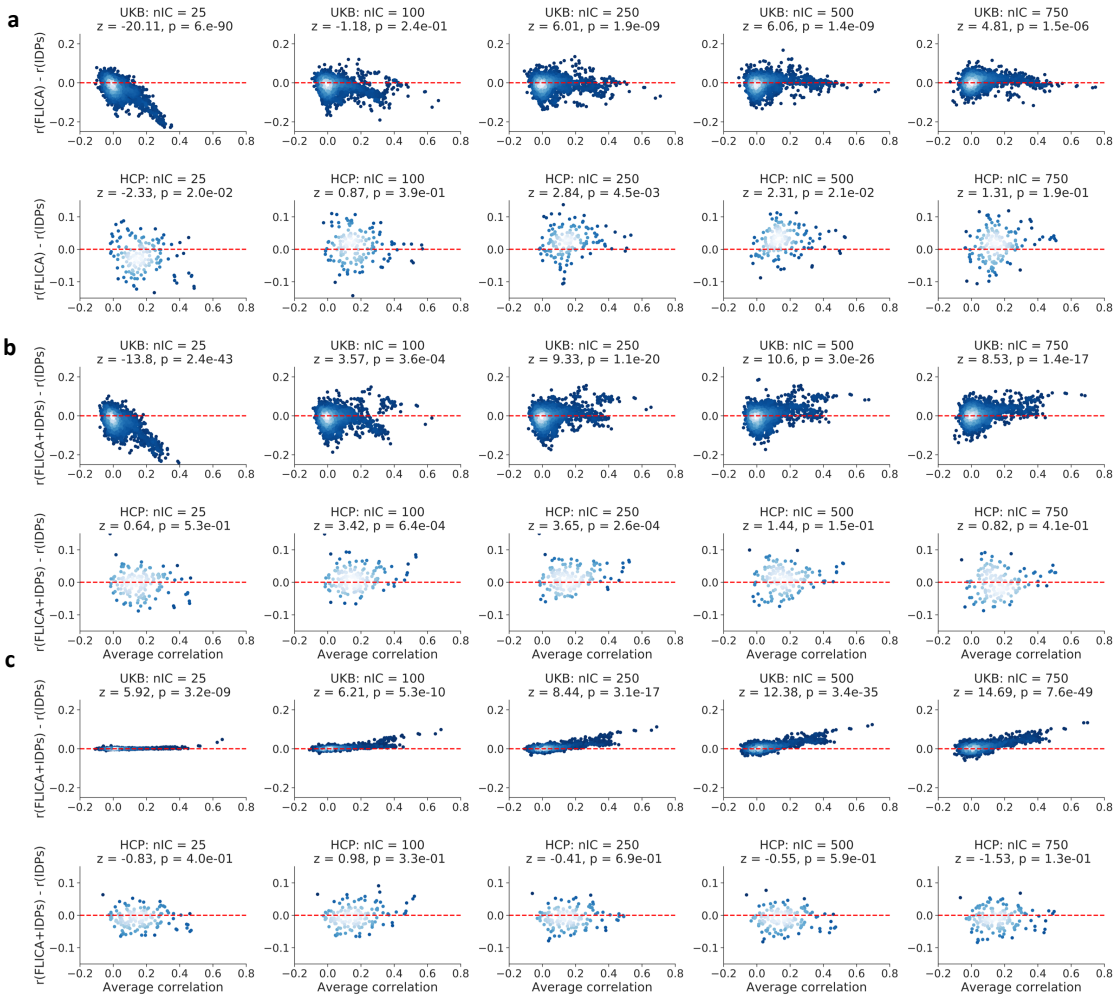



**Figure 2.4: Comparison of prediction accuracy of nIDPs between BigFLICA against single-modality ICA and the IDPs.** Overall, for high-dimensional BigFLICA decompositions in the UKB dataset, BigFLICA achieved statistical significant increases of prediction accuracy of nIDPs compared with single-modality ICA and IDPs. Combining BigFLICA and IDPs together further improves compared with IDPs alone. In each of the Bland–Altman plots, each point represents the prediction of an nIDP, where the x-axis is the average prediction correlation of the two approaches, while the y-axis is the difference. The z- and p-values in the titles reflected the statistical significance of the differences. The Bonferroni correction 0.05 threshold corresponds to a raw p-value of 1.7e-3. a, Comparing BigFLICA with the concatenation of single-modality ICA outputs. Top: UKB; Bottom: HCP. The number of FLICA components is set from 25 to 750. b, Comparing BigFLICA with IDPs. Top: UKB; Bottom: HCP. The number of IDPs is 3,913 in UKB and 5,812 in the HCP. c, Comparing the concatenation of BigFLICA and IDPs against IDPs only. Top: UKB; Bottom: HCP. The lighter the blue, the higher the density of points.

important data variation into joint components. However, when $L$ becomes large, the prediction accuracy becomes better than the single-modality ICA (e.g., $\geqslant 250$ in UKB). This is because, at high dimensional decomposition, BigFLICA effectively combines multimodal information by considering cross-modal correlation in the data decomposition

stage. Although the cross-modal correlation is considered in the final prediction stage when using single-modality ICA, the fact that BigFLICA identifies and takes advantage of correlated information between modalities at an earlier stage in feature generation helps improve the prediction performance.

In **Fig. A.2**, we also compared, in the UKB data, the 750-dimensional BigFLICA decomposition with the 25-dimensional ICA decomposition concatenated across modalities, i.e., we have $25 \times 47$ features in the single-modality ICA. In this comparison, the number of features for the two methods are almost the same, but we can see that BigFLICA clearly outperforms the single-modality ICA.

### 2.3.5 Comparing BigFLICA with hand-curated image-derived phenotypes.

We compared the predictive performance of BigFLICA with IDPs in both HCP and UKB datasets (**Methods section 2.2.11**). **Fig. 2.4b** shows that, in the UKB data, when the number of modes is low, BigFLICA has a worse predictive power than the joint performance of 3,913 IDPs, due to the same insufficient degree-of-freedom reason as above. However, when the dimensionality becomes higher, BigFLICA is clearly outperforming the IDPs, owing to jointly fusing multimodal voxelwise data by considering cross-modality correlation. In the HCP data, the performance is overall similar. These results indicate that BigFLICA can potentially explain more phenotypic and behavioural variances than IDPs.

In more detail, **Tables A.1** shows that, in the UKB dataset, the high-dimensional BigFLICA (nIC=750) has improved prediction accuracy for many nIDPs that relate to *cognition phenotypes* and *health outcomes* compared with IDPs. These tables do not include nIDPs where both methods have low predictive power ($r < 0.1$). In the HCP dataset (**Table A.2**), BigFLICA (nIC=100) also shows improved prediction accuracy in many cognitive and health outcomes variables compared with using IDPs.

Further, when we concatenated the modes of BigFLICA and IDPs together to predict nIDPs, as shown in **Fig. 2.4c**, the combined feature sets have a significant improvement of prediction accuracy than the IDPs alone in the UKB data. There are almost no differences

for the same comparison in the HCP data. This suggests that BigFLICA and IDPs may contain some complementary information of nIDPs.

To investigate the relationships between BigFLICA and IDPs further, we built prediction models that used modes of BigFLICA to predict each of the IDPs, to further characterise information overlap and complementarity between the two approaches. As shown in **Figs. A.3a** and **A.3b**, different types of IDPs can be predicted differently, and the resting-state functional connectivities always had the worst accuracy in both the HCP and the UKB datasets, because they are (relatively) noisy. However, when using BigFLICA modes to predict 6 new summary features of the connectivity matrices (derived by applying ICA to the matrix of subjects by network matrix edges) [Elliott et al., 2018], the accuracy is very high ($r$ range from 0.85 to 0.89 for a 100 dimensional BigFLICA decomposition). In addition, when we used IDPs to predict modes of BigFLICA, as shown in **Figs. A.3c** and **A.3d**, the prediction correlation almost showed a bimodal distribution, which means that some of the FLICA modes can be predicted by the IDPs (mean $r \approx 0.8$) while others cannot (mean $r \approx 0.2$). These results further demonstrate that BigFLICA and IDPs span significant complementary variance.

### 2.3.6 Examples of BigFLICA modes in the 14k UKB dataset.

We now give four examples of significant associations between BigFLICA modes and nIDPs, namely, *Fluid intelligence*, *Age started wearing glasses or contact lenses*, *Handedness* and *hypertension*. In **Fig. 2.5**, we show the top four most strongly associated modalities in FLICA modes that correlate with a given nIDP. **Fig. A.11** shows the population cross-subject mean maps for several task and rest fMRI modalities fed into FLICA. This helps give interpretive context for the FLICA mode maps, which depict subject variability in the activity/connectivity relative to these group mean maps.

For *Fluid intelligence*, using all modes (ICs) from the 750 dimensional BigFLICA decomposition as features (predictors) in multivariate elastic-net prediction, a cross-validated prediction correlation of $r = 0.26$ is achieved. When we correlated each of the BigFLICA modes and IDPs with the fluid intelligence score in the UKB, we found that several task-fMRI-related BigFLICA modes have the strongest associations (**Fig. 2.5a**).

The first (IC 25) involves task contrast "faces" and "faces>shapes" and the second (IC 57) involves contrast "shapes" and "face" (see **Tables A.6** for the full list of these modalities). As the correlation of the mode IC 25 (i.e., its subject weights vector) with fluid intelligence is negative (r=-0.14), this means that the negative-weights voxels (such as in the anterior insula) are positively correlated with intelligence. The fMRI task (Hariri faces-shapes matching [Hariri et al., 2002]) has, as expected, the greatest population average activation in sensory-motor areas (plus some amygdala involvement due to the emotionally negative nature of the faces), as seen in **Fig. A.11**. However, the main brain areas involved in these modes are distinct, including anterior cingulate cortex, frontal pole, inferior frontal gyrus, and anterior insula; it is therefore interesting that the areas found by BigFLICA to be modulated in these components (and found to associate with intelligence) are more "frontal, cognitive" areas than the sensory-motor areas primarily activated on average. The top associations between fluid intelligence and IDPs also involve task-fMRI IDPs (**Tables A.3**), but these were a factor of two weaker than associations with BigFLICA modes.

For *Age started wearing glasses or contact lenses*, BigFLICA achieved a prediction correlation of $r = 0.16$. Several resting-state connectivity and task modalities showed associations in primary visual areas (**Fig. 2.5b**), which is consistent with the fact that this is a vision-related health variable. Lower age of first wearing glasses is correlated with stronger activity in primary visual areas, and also with strength of resting-fMRI connectivity (or functional coherence) within the relevant areas of group-average connectivity; interestingly, in nearby distinct (but still primary visual) areas, there is reduction of correlation (blue voxels), suggesting greater differentiation of primary visual areas.

For *Handedness*, BigFLICA achieved a prediction correlation of $r = 0.23$. BigFLICA identified several multimodal, lateralized (or laterally asymmetric) modes, including resting-state mode 14 (left-lateralized language network), task, surface area and white matter tracts (**Fig. 2.5c**). There are several resting-state connectivity-related IDPs correlated with handedness (**Tables A.3**), consistent with a recent study [Wiberg et al., 2019] that also used UKB IDPs, while no IDPs related to other modalities are found significant;

in both cases the maximum IDP correlation only reached r=0.12, whereas the strongest association with BigFLICA modes was almost double this.

For a health variable *hypertension* (**Fig. 2.5d**), BigFLICA achieved a prediction correlation of $r = 0.22$. Several TBSS-related modalities showed consistent associations in the External Capsule tracts. Meanwhile, white matter hyperintensity (T2-Lesion volume) in the corresponding areas is also higher in people with hypertension. Several consistent findings have been reported in the literature [Moon et al., 2005, Allen et al., 2016, Hannawi et al., 2018].

### 2.3.7 BigFLICA comparison with mCCA and reproducibility.

We tested whether BigFLICA (independent components-based spatial modelling) was better than mCCA (eigendecomposition based modelling, which could be considered to be similar to the output of BigFLICA without running the final core FLICA unmixing - note that to enable mCCA to run requires the same mMIGP initial processing that we have added in this work) in three ways. The number of extracted components was the same when performing this comparison. First, for the prediction accuracy of nIDPs, **Fig. A.4** shows that, in the UKB data, BigFLICA has a (very slightly) improved prediction accuracy compared with mCCA. Then, we proposed a hypothesis that modes of BigFLICA are more parsimonious features of nIDPs compared with mCCA, or in other word, a smaller number of modes of BigFLICA can predict the nIDPs. Results shown in **Fig. A.5** validate this hypothesis: for a given number of components and a given nIDP, BigFLICA modes have a (on average) higher proportion of zero weights in the elastic-net predictions, when compared with mCCA modes. The advantage is that a more parsimonious representation usually has a better biological interpretability. Finally, we estimated and compared the split-half reproducibility of BigFLICA and mCCA. As shown in **Fig. A.6** (right), BigFLICA has a much higher between-subject reproducibility than mCCA (median BigFLICA correlation greater than 0.9 in all cases, while many mCCA dimensionalities have median correlation less than 0.5).

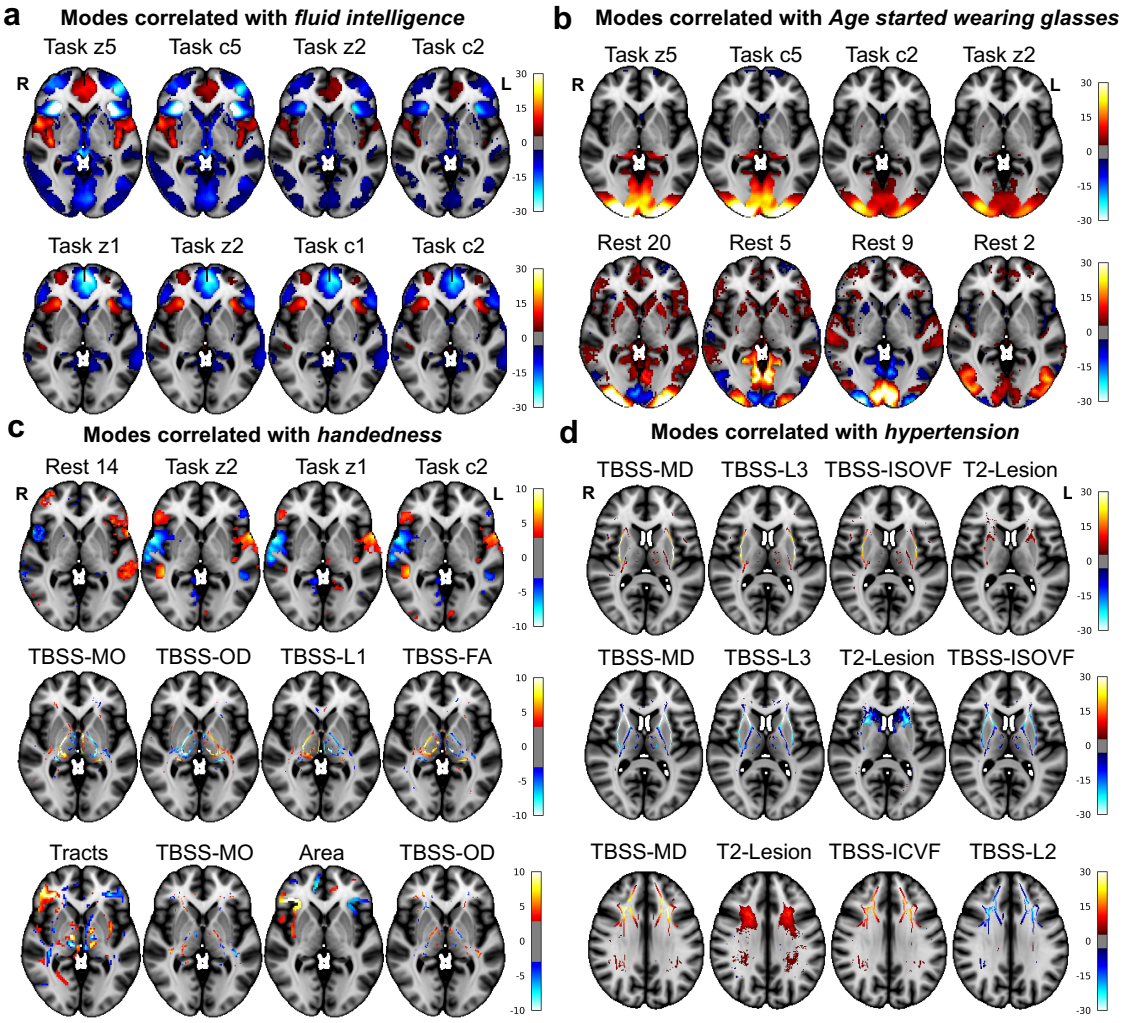



**Figure 2.5: Examples of BigFLICA modes in the 14k UKB dataset.** For each subfigure, each row shows one IC (BigFLICA mode or independent component) with top 4 most strongly associated modalities. **a,** Two BigFLICA modes that significantly correlate with *fluid intelligence* (IC25: $r = -0.14$; IC57: $r = -0.12$). **b,** Two BigFLICA modes that significantly correlate with *Age started wearing glasses or contact lenses* (IC164: $r = -0.10$; IC13: $r = -0.05$). **c,** Three BigFLICA modes that significantly correlate with *handedness* (IC235: $r = -0.23$; IC569: $r = 0.07$; IC232: $r = -0.04$). **d,** Three BigFLICA modes that significantly correlate with *hypertension* (IC259: $r = 0.12$; IC13: $r = -0.11$; IC319: $r = -0.09$). The Bonferroni corrected 0.05 threshold corresponds to an uncorrected p-value of $9.2 \times 10^{-9}$ (corrected for number of components (750) and number of nIDPs (7245)). All of the above correlations passed the Bonferroni threshold except for IC232 with uncorrected $p = 2.1 \times 10^{-7}$.

### 2.3.8 Reproducibility of BigFLICA.

To test whether BigFLICA's spatial independent components are estimated reliably, the whole UKB dataset was divided into two parts: the first part contained 7,000 subjects and the second part contained the remaining 7,503 subjects. We applied BigFLICA to the two parts separately. After estimating the subject modes, we reconstructed the z-score

normalized (voxel-wise) spatial maps of each modality by regressing the subject mode against the mMIGP-reduced data. The spatial independent components of each modality were concatenated spatially and greedily paired, based on the absolute correlation between two runs. When we computed the correlations, only voxels whose absolute z-scores that are both larger than 3 in two runs were preserved (to reduce noise, given that there are huge numbers of empty voxels across all modalities for a given FLICA component in general; this does not bias the metric of reproducibility towards finding common similar patterns). **Fig. A.6** (left) shows that the FLICA components have very high reproducibility in the split-half test across a varying number of components.

We further tested the reproducibility of prediction accuracy of BigFLICA for predicting nIDPs. We ran BigFLICA separately on two halves of the dataset as above, and then predicted each of the nIDPs using elastic net regression in each half. **Fig. A.7** shows the prediction accuracy of all nIDPs. We can see that the prediction accuracy of nIDPs are highly correlated, especially for nIDPs that have a high prediction accuracy ($r = 0.57$ for a 25 dimensional decomposition, and $r > 0.7$ for higher dimensions).

Finally, we investigated the influence of sample size on the reproducibility. We performed the same split-half reproducibility test on two random subsets of 1,500 and 3,500 subjects. **Fig. A.8** shows that across different numbers of components, BigFLICA has a reproducibility of from 0.6 to 0.7, which is expected to be lower than the 7,000 subject case. We can also see that there is a slight increase of reproducibility when the number of components increases. This different behaviour compared with the 7,000-subject cases may be due to the fact that many of the components are empty when the number of components is larger than 250 (i.e., there is not enough data to support the ICA dimensionality), so that we can only compute the reproducibility index based on the non-empty components.

### 2.3.9 Stability of BigFLICA prediction.

We evaluated the stability of prediction accuracy of BigFLICA against different train-test subject splits. We estimated the prediction correlation of each nIDP using 5 different train-test subject splits, and calculated the mean accuracy of prediction correlations for

each nIDPs. We then computed the difference of the mean prediction correlation and the prediction correlation of one of the five random predictions. As shown in **Fig. A.9**, the differences are centered around zero with extremely small spread, which demonstrated the stability of BigFLICA against different random train-test splits.

Among the 25 resting-state dual-regression spatial maps included in the analysis, four of them had been originally identified as non-neural components [Miller et al., 2016]. Non-neural components likely reflect non-neuronal physiology and therefore may help prediction, particularly for the nIDPs that relate to basic physiology (e.g., blood pressure). Therefore, we tested the prediction accuracy of nIDPs when we exclude the four non-neural modalities (rfMRI components). **Fig. A.10** shows that an increased prediction accuracy was observed for some nIDPs in the 25-dimensional decomposition when compared with using all 47 modalities, a similar prediction accuracy was observed for 100- and 250-dimensional decompositions, while a decreased prediction accuracy was observed for 500- and 750-dimensional decompositions. The set of nIDPs that are better predicted by inclusion of 4 non-neural components at high dimensions are those that relate to physical measures. Researchers may choose to include artifactual rfMRI components (e.g., where this helps maximise nIDP associations), or may exclude them (e.g., to maximise interpretability of associations).

### 2.3.10 Contribution of different modalities in a BigFLICA decomposition.

Besides using BigFLICA for exploring the relationships between imaging and non-imaging phenotypic and behavioural data, we can also use it to investigate the relationship between different modalities. For each mode, BigFLICA estimates a vector of positive numbers reflecting the contributions of different modalities (i.e., the diagonal of each $W_{(k)}$, where the higher the number, the more important is one modality to a mode). We concatenated all such vectors across all modes so that it is a mode-by-modality matrix $W$, and normalized each column to sum to one. Six examples of such matrices are shown in **Fig. A.12**, with different numbers of estimated modes in the UKB dataset.

We then calculate each row's sum (across columns) in $W$, thereby reflecting the overall contribution of each modality in the BigFLICA decomposition. As shown in **Fig. A.13**, across all FLICA dimensionalities (numbers of estimated modes), each of the 25 resting-state fMRI dual-regression spatial maps usually has a low overall contribution, followed by task fMRI maps, while modalities reflecting more about structure of the brain (e.g., structural MRI and diffusion MRI) generally have high overall contributions. The relative differences of modality contribution between functional MRI-related modalities and structural/diffusion MRI-related modalities become larger with increasing number of estimated modes. We further estimated the total shared variances between a lower dimensional BigFLICA decomposition and a higher dimensional decomposition. **Table A.4** shows that a higher dimensional decomposition explains almost all variances of a lower dimensional decomposition (upper triangle of the table), while a lower dimensional decomposition can explain a large proportion of the variances of a higher dimensional decomposition.

### 2.3.11 Relationship between different modalities in a BigFLICA decomposition.

We calculated the cosine similarity between different columns of $W$ (using the 750-dimensional BigFLICA decomposition), to measure the similarity of different modalities in terms of their contribution to the BigFLICA decomposition, i.e., the more similar information two modalities carry, the more likely they will have similar contribution to a mode. **Fig. A.14a** shows that the modality relationship matrix is clearly grouped into three large clusters. The first is all resting-state modalities, while the second is the task fMRI maps, and the third is the diffusion MRI, structural MRI-related modalities and swMRI. The white matter hyper-intensity map (T2 lesions) forms a single cluster. As a comparison, we also performed a 50-dimensional ICA decomposition within each modality, and calculated the shared variances between every pair of 50 ICs in two modalities using a simple multivariate regression model. As shown in **Fig. A.14b**, we also observed a similar pattern as **Fig. A.14a**. The main difference is that in **Fig. A.14a**, there are relatively stronger correlations within resting-state modalities and between resting-state and other modalities, but weaker correlations between task modalities

and structural related modalities. These results reflect the fact that the multimodal modelling effects of BigFLICA learn different inter-modality relationships compared with single-modality ICA.

## 2.4 Discussion

In this chapter, we presented BigFLICA, a multimodal data fusion approach which is scalable and tuneable to analyze the full UK-Biobank neuroimaging dataset, and other large-scale multimodal imaging studies. To the best of our knowledge, this is the first approach for data-driven (unsupervised) multimodal analysis in a brain imaging dataset of this size and complexity. Building on the top of the powerful FLICA model, we proposed a two-stage dimension-reduction approach that combines an incremental group-PCA (mMIGP) and dictionary learning (DicL) to effectively preprocess the multimodal dataset and reduce the computational load of the final FLICA, while maintaining or even improving performance, with as much as a 150-fold "intelligent" reduction in data size. We provide effective ways of choosing the hyper-parameters of BigFLICA, so that it is free of tuning except for choosing the final number of estimated components. Although this approach is motivated by the need for analyzing extremely big neuroimaging data, it is also applicable to other kinds of data such as genetics and behavioural measures. 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]. BigFLICA results on UKB will also be released via the UKB database as new data-driven IDPs (image features), for further epidemiological and neuroscientific research.

A strength of our work is that, unlike previous work that was limited to more moderate datasets and a few phenotypic and behavioural variables [Calhoun et al., 2006, Liu et al., 2009, Beckmann and Smith, 2005, Groves et al., 2011, Sui et al., 2012], we used two of the largest, high-quality multimodal neuroimaging datasets, and thousands of phenotypic and behavioural variables to validate the proposed approach. We demonstrated that BigFLICA is not only much faster than the original FLICA (and can be run on very large data that is simply not analysable with FLICA or other existing methods), but also estimates similar modes with a comparable performance

for predicting the non-imaging-derived phenotypes in real data (when tested on a large data subset that is just small enough to allow for comparison against FLICA). We provide insights into the advantages of data-driven multimodal fusion in big datasets by quantitative analysis [Calhoun and Sui, 2016, Uludağ and Roebroeck, 2014]. First, when comparing BigFLICA with simpler IDP-based approaches (and also single-modality ICA approaches), we demonstrated that a high-dimensional BigFLICA has improved predictive power overall. We demonstrated the value of multimodal fusion instead of analyzing each modality separately. Second, when combining high-dimensional BigFLICA-derived features with IDPs together, the predictive power increased further compared with using either method alone. In addition, when we used BigFLICA-derived features and manually created (with expert knowledge) IDPs to predict each other, they cannot predict each other perfectly (although they are derived from the same imaging data). This indicates that BigFLICA-derived features and IDPs can be complementary to each other, both therefore providing potentially important imaging biomarkers that capture different signal in the imaging data. An interesting finding is that although a high-dimensional BigFLICA has a much higher predictive power than a low dimensional decomposition, a low dimensional decomposition can still explain more than 80% of the total variance of the high dimensional decomposition. This suggests that some the phenotypic and behavioural variables are explained by only small proportions of variance of imaging data. Third, in addition to the value of using BigFLICA-derived features for relating imaging to non-imaging data, BigFLICA components (particularly at lower dimensionalities) may allow us to learn more about how the different brain imaging modalities (and hence different spatial and biological aspects of the brain's structure and function) relate to each other. Finally, when new primary data becomes available from new subjects, this data would need to have new IDPs calculated (at the subject level) and then combined with existing IDPs from previous subjects, for a complete between-subject analysis. The approach presented here, while not avoiding any of the necessary new computations, will make these efficient. Further, note that it is alternatively possible to use an existing decomposition and apply this to new subjects by projecting them onto the existing spatial bases to generate new subject weights (loadings) against the ICA features.

We see opportunities to improve the current approach. First, BigFLICA is limited to linear feature estimation, while the "ideal, true" information in imaging data may be highly nonlinear. Therefore, a nonlinear extension of BigFLICA, which might be achieved with kernel methods or deep neural networks, is an important area of further research. Second, BigFLICA is an unsupervised dimension reduction and feature generation approach. However, integrating some supervision, i.e., the target variable (such as disease outcomes), into the dimension reduction may boost the performance of the algorithm. Additionally, because BigFLICA generates data-driven features, as opposed to expert-created IDPs, the biological or anatomical interpretation of features is often likely not to be immediately obvious, requiring potentially intensive expert study. Future work could attempt to automate this interpretation process, for example by relating features to existing anatomical templates and atlases, and even by mining imaging literature. Finally, BigFLICA, or extensions, may be an effective way of discovering imaging confound factors [Li et al., 2019] that cannot be found by traditional approaches.

## Acknowledgments

We are grateful to UK Biobank and its participants (access application 8107). We thank the WU-Minn HCP Consortium for their invaluable contributions in generating the publicly available HCP data and implementing the procedures needed to acquire, analyze, visualize and share these datasets. 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.

## Data availability

BigFLICA-derived features will be available from the UK Biobank database. 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). For HCP, data can be downloaded via website (http://humanconnectome.org/data) and ConnectomeDB.

## Code availability

BigFLICA code is available at https://github.com/weikanggong/BigFLICA, and will also be released as part of an upcoming version of FSL. Matlab software for performing prediction using elastic-net regression is available at https://github.com/vidaurre/NetsPredict.

# 3

# Supervised Phenotype Discovery from Multimodal Brain Imaging

## Contents


**3.1 Introduction . . . . . 48**
**3.2 Methods . . . . . 52**
3.2.1 SuperBigFLICA from an optimisation perspective. . . . . 52
3.2.2 SuperBigFLICA in a Bayesian framework. . . . . 54
3.2.3 Model parameter optimization. . . . . 56
**3.3 Experiments . . . . . 57**
3.3.1 Brain imaging data. . . . . 57
3.3.2 Non-imaging derived phenotypes. . . . . 58
3.3.3 Confounding variables and missing values. . . . . 58
3.3.4 Imaging data pre-reduction using dictionary learning. . . . . 59
3.3.5 Baseline: nIDP prediciton using hand-curated IDPs and modes of BigFLICA. . . . . 60
3.3.6 Predicting nIDPs using single-task and multi-task SuperBigFLICA. 61
3.3.7 Evaluation of the generalisability of SuperBigFLICA modes on unseen nIDPs. . . . . 61
**3.4 Results . . . . . 62**
3.4.1 Comparing SuperBigFLICA with hand-curated IDPs and modes of unsupervised BigFLICA. . . . . 62
3.4.2 Analysis of SuperBigFLICA algorithm. . . . . 64
3.4.3 Real data further qualitative analysis. . . . . 67
**3.5 Discussion . . . . . 70**

## Summary

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 project, 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.

## 3.1 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., 2021b, 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 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., 2021b]. 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 chapter, 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., 2021b]. 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,

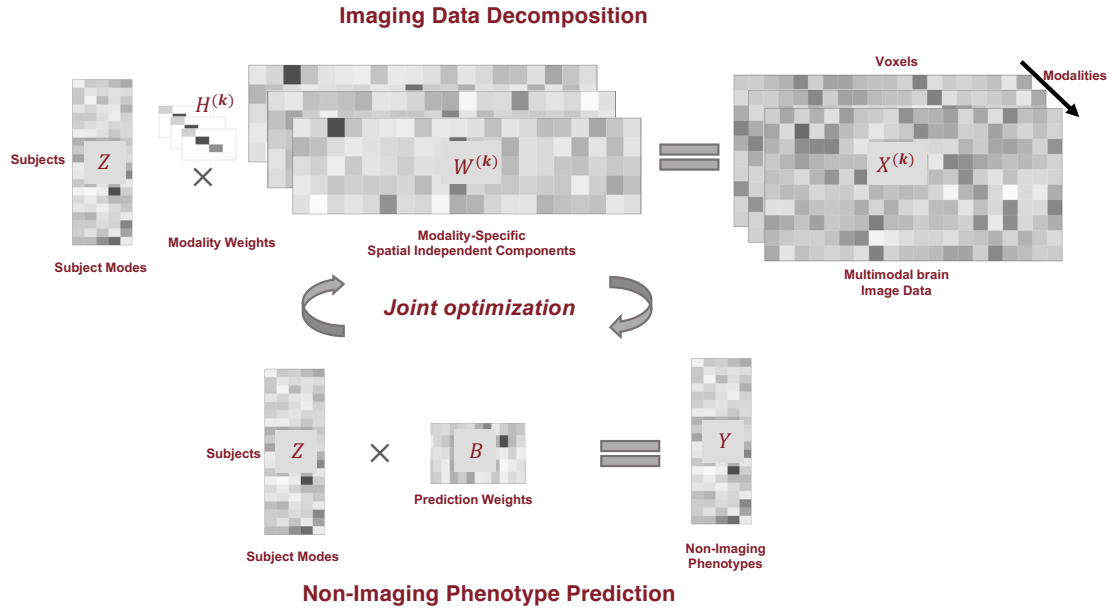



**Figure 3.1:** Overview of the proposed SuperBigFLICA approach for supervised multimodal fusion and phenotype discovery.

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 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.

## 3.2 Methods

### 3.2.1 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 $\mathbf{X^{(k)}}$ 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:

$$\mathbf{X^{(k)}} = ZH^{(k)}W^{(k)} + E^{(k)},\ k = 1, \ldots, K \tag{3.1}$$

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

$$\mathbf{Y} = ZB + E. \tag{3.2}$$

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 feature 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, $E^{(k)}_{(N\times Q)}$ and $E_{(N\times Q)}$ are independent Gaussian random error terms. **Fig. 3.1** 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:

$$Z = \frac{1}{K}\sum_{k=1}^{K} Z^{(k)} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{X^{(k)}}(H^{(k)}W^{(k)})', \tag{3.3}$$

where $'$ represents transpose. This shared subject loading is a weighted average of subject loadings per modality, i.e., $Z^{(k)}$, 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]:

$$\min_{W,H} \sum_{k=1}^{K} \|\mathbf{X^{(k)}} - Z(H^{(k)}W^{(k)})\|_2^2 \tag{3.4}$$

Note that the transpose of $H^{(k)}W^{(k)}$ in Eq. (3.3) - 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:

$$\min_{W,H,B} \underbrace{\sum_{k=1}^{K} (\lambda_1^{(k)} \|\mathbf{X^{(k)}} - Z(H^{(k)}W^{(k)})\|_2^2 + \lambda_2^{(k)} |W^{(k)}|_1)}_{\text{data reconstruction loss}} + \underbrace{\sum_{i=1}^{Q} (\lambda_3^{(i)} \|\mathbf{Y_i} - ZB_i\|_2^2 + \lambda_4^{(i)} |B_i|_1)}_{\text{nIDP prediction loss}} \tag{3.5}$$

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

### 3.2.2 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 $\lambda$ 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 $\lambda$ 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:

$$P(\mathbf{X^{(k)}}|Z,W^{(k)},H^{(k)},\sigma^{(k)}) = N(ZH^{(k)}W^{(k)},(\sigma^{(k)})^2 I) \tag{3.6}$$

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]:

$$P(W^{(k)}|b^{(k)}) = \frac{1}{2b^{(k)}} \exp\left(-\frac{|W^{(k)}|}{b^{(k)}}\right) \tag{3.7}$$

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:

$$P(Y_i|Z,B_i,\gamma_i^2) = \mathcal{N}(ZB_i,\gamma_i^2 I), \tag{3.8}$$

where $\gamma_i$ is a task-specific noise term. We also place a Laplacian prior on prediction weights $B_i$ [Park and Casella, 2008]:

$$P(B_i|c_i) = \frac{1}{2c_i} \exp\left(-\frac{|B_i|}{c_i}\right), \tag{3.9}$$

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:

$$\log P(\theta|D) \propto \log P(D|\theta)P(\theta) = \log P(\mathbf{X^{(1)}},\ldots,\mathbf{X^{(K)}}|\theta)P(\mathbf{Y}|Z,\theta)P(\theta) \tag{3.10}$$

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 $\lambda$ is shown to be useful in different kinds of prediction tasks in our experiments later. Thus, we get a modified posterior to be maximized:

$$\begin{aligned}
&\lambda \sum_{k=1}^{K} \Big( \log P(\mathbf{X^{(k)}}|\theta) + \log P(\sigma^{(k)}, W^{(k)}, b^{(k)}) \Big) + (1-\lambda) \sum_{i=1}^{Q} (\log P(\mathbf{Y_i}|Z,\theta) + \log P(\gamma_i, B_i, c_i)) \\
&= \underbrace{\lambda \sum_{k=1}^{K} \left[ \left( -\frac{1}{2(\sigma^{(k)})^2} \|\mathbf{X^{(k)}} - ZH^{(k)}W^{(k)}\|_2^2 + (2\alpha_1 - 3)\log(\sigma^{(k)}) - \beta_1 (\sigma^{(k)})^2 \right) + \left( -2\log(b^{(k)}) - \frac{1}{b^{(k)}} |W^{(k)}| \right) \right]}_{\text{data reconstruction loss}} \\
&+ (1-\lambda) \underbrace{\sum_{i=1}^{Q} \left[ \left( -\frac{1}{2(\gamma_i)^2} \|\mathbf{Y_i} - ZB_i\|_2^2 + (2\alpha_2 - 3)\log(\gamma_i) - \beta_2 (\gamma_i)^2 \right) + \left( -2\log(c_i) - \frac{1}{c_i} |B_i| \right) \right]}_{\text{nIDP prediction loss}} + \text{const}
\end{aligned} \tag{3.11}$$

We can appreciate from Eq. (3.11) that, the $\lambda$s in Eq. (3.5) 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. (3.3) 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.

### 3.2.3 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) algorithm. 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 (we found that the performance is stable across different batch sizes) and chose the optimal learning rate from $0.0001, 0.001, 0.01$, and the tuning parameters $\lambda$ 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] (Similarly, we found that the performance is stable across different dropout ratios). 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.

## 3.3 Experiments

### 3.3.1 Brain imaging data.

Voxel-wise neuroimaging data of 47 modalities from 39,770 subjects were used in this chapter, 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 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].

### 3.3.2 Non-imaging derived phenotypes.

A total of 17,485 non-imaging derived phenotypes (nIDPs) were used in this chapter. We mainly analyzed 12 of them in the 'physical', 'cognition', and 'health outcome' domains, summarised in **Table 3.1**. 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., 2021b]. The direct comparison of performance between SuperBigFLICA and BigFLICA approaches enables us to study the benefits of including the supervised loss terms.

### 3.3.3 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.

**Table 3.1:** 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 |

### 3.3.4 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., 2021b], 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.

### 3.3.5 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., 2021b]. 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., 2021b] and a 2,000-dimensional dictionary learning approach as data preprocessing [Gong et al., 2021b] 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., 2021b]. 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). The elastic-net 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.

### 3.3.6 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 $\lambda$ parameter is chosen from $1E-5$ to 0.99999. We evaluated the influence of $\lambda$, and different random model parameter initialisations on the final prediction performance in the result section.

### 3.3.7 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.

## 3.4 Results

### 3.4.1 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 3.3.5, 3.3.6 and 3.3.7). Each subfigure of **Fig.3.2** 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 ($\lambda$ 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%.

The fourth and fifth columns of each subfigure of **Fig.3.2** show prediction accuracy of multi-task SuperBigFLICA, with 24 and 99 most correlated nIDPs as auxiliary

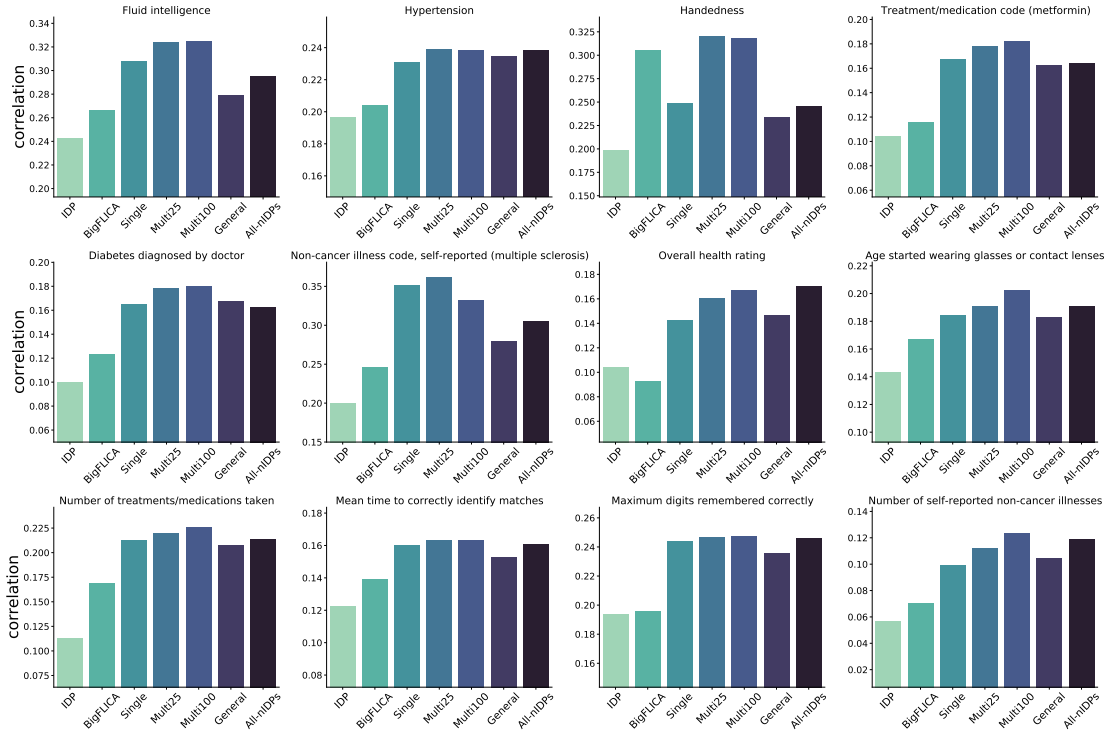



**Figure 3.2: Comparison of SuperBigFLICA with hand-curated IDPs and modes of unsupervised BigFLICA for the 12 nIDPs listed in Table 3.1.** Each figure shows - for a different nIDP - the predictive correlation of different approaches and different parameter settings. The first and second column shows 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.

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.3.2** shows the generalisability of Super-BigFLICA 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

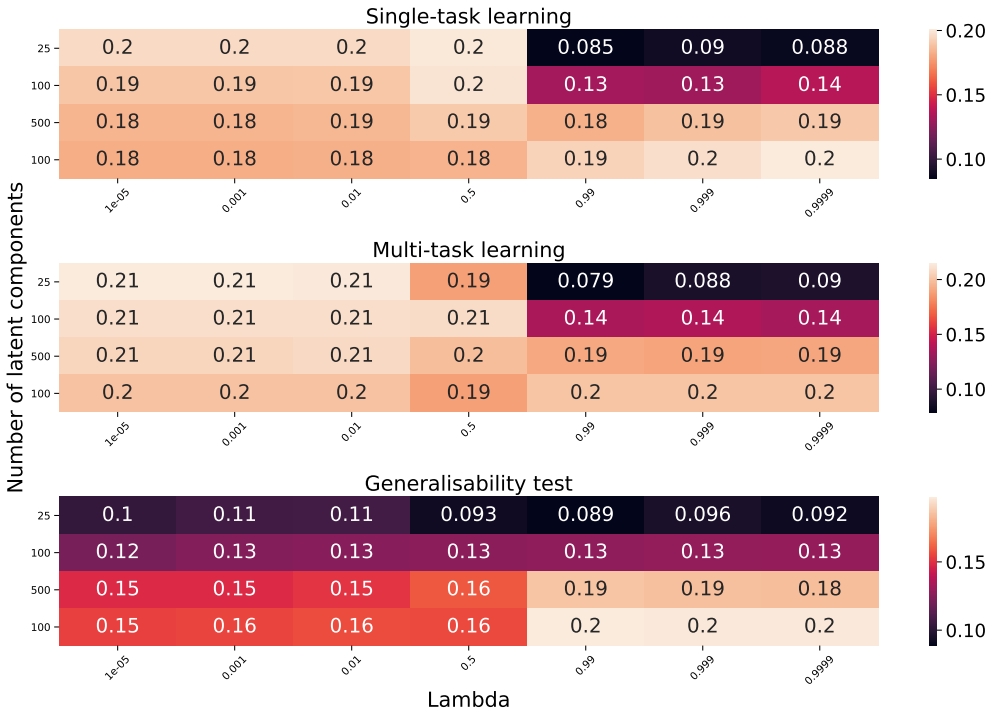



**Figure 3.3: The relationship between the number of latent dimensions (y-axis) and weights between reconstruction and prediction losses (x-axis) with the mean prediction accuracy across 12 nIDPs listed in Table 3.1. Top.** Single-task learning setting. **Mid.** Multi-task learning setting with 24 auxiliary tasks. **Bot.** Generalisability test.

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.3.2** 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.

### 3.4.2 Analysis of SuperBigFLICA algorithm.

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

We evaluated the influence of the relative weighting between reconstruction loss and prediction loss, i.e., $\lambda$, and the number of latent components, on the final prediction performance. **Fig.3.3A** shows the mean prediction correlation of 12 nIDPs listed in

**Table** 3.1 across different $\lambda$ and latent components using single-task SuperBigFLICA. When the latent dimension is small, we need to choose a small $\lambda$ (i.e., $\lambda < 0.5$, which means that the prediction loss has a higher weight than the reconstruction loss) to achieve optimal prediction. When the latent dimension is large, although a smaller $\lambda$ also achieves the best prediction accuracy, the influence of $\lambda$ becomes much smaller than using a smaller latent dimension. We can draw a similar conclusion from the results shown in **Fig.3.3B**, which is the case of multi-task SuperBigFLICA with 24 most correlated nIDPs as auxiliary tasks. Conversely, for the generalisability test, **Fig.3.3C** 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 hyper-parameters in different circumstances and demonstrates the usefulness of including both data reconstruction and prediction losses in the objective function.

**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 3.3.4). **Fig.3.4A** 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

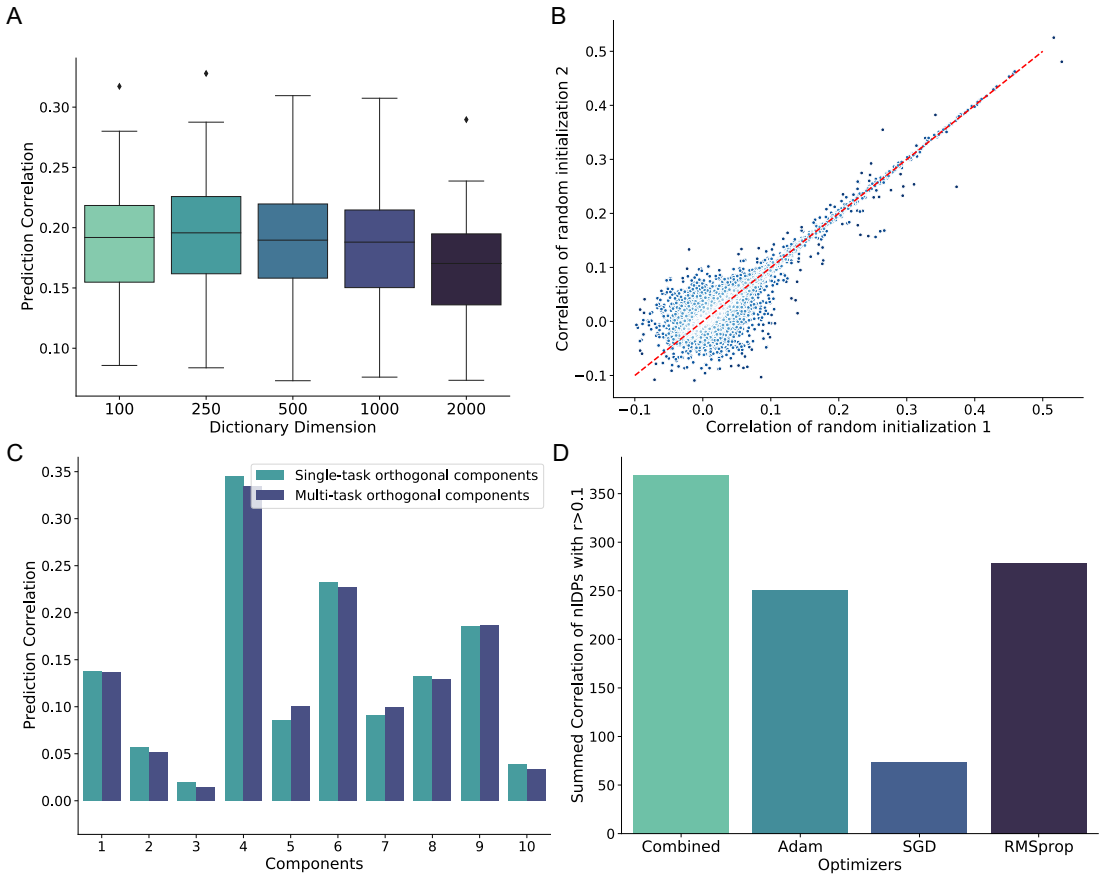



**Figure 3.4: Evaluations of SuperBigFLICA model.** **A**. The average prediction accuracy of 12 nIDPs using different dictionary learning dimensions in the image data pre-reduction stage. **B**. The prediction accuracy of 17,485 nIDPs as a result of different random parameter initializations. **C**. 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.

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.

### 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 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.3.4B** 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.

**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.3.4C**). 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.

**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.3.4D**). The overall accuracy is estimated by the sum of correlation of nIDPs larger than 0.1.

### 3.4.3 Real data further qualitative analysis.

**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.3.5A** shows the weights of each of the 25 latent

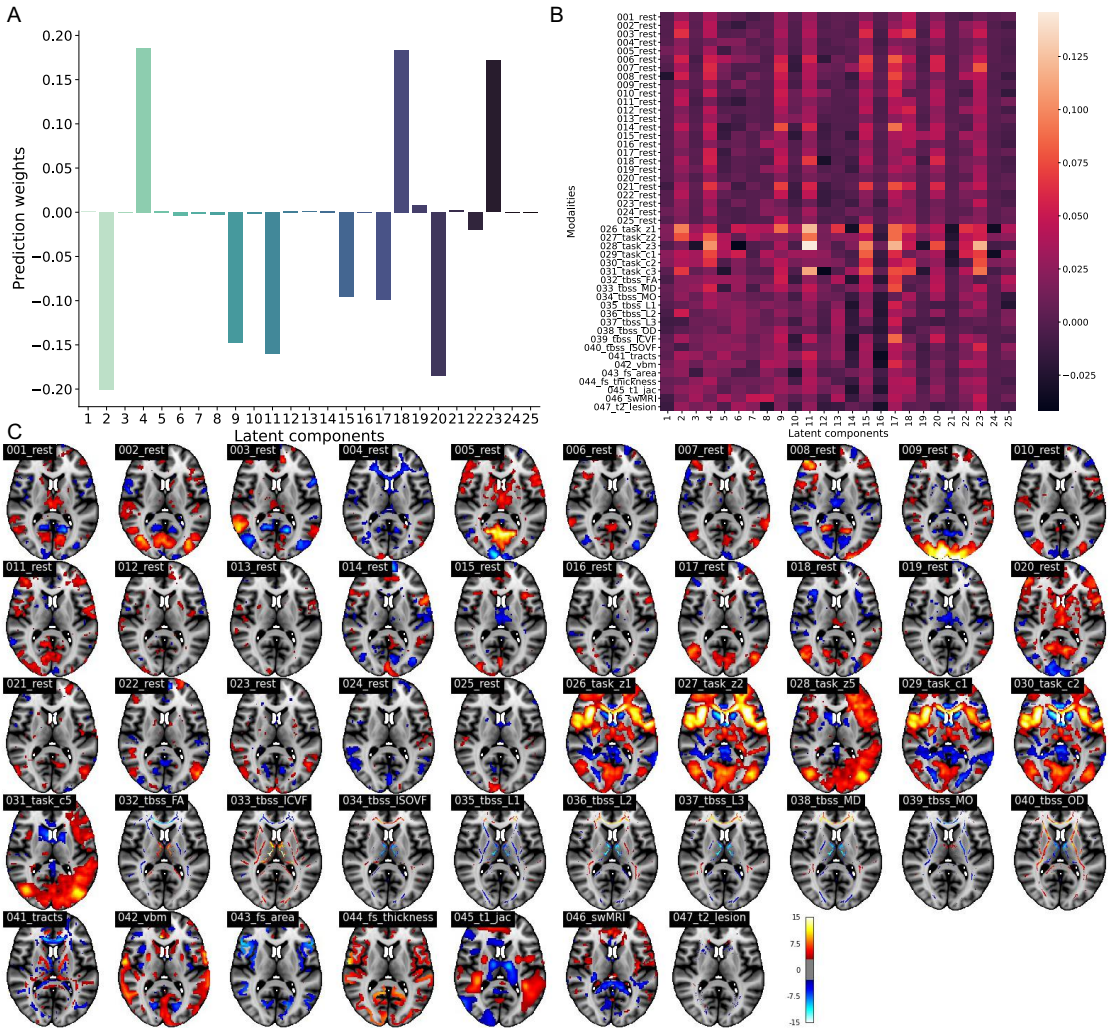



**Figure 3.5: Example results of predicting fluid intelligence with a 25-dimentional SuperBigFLICA.** **A.** The prediction weights of fluid intelligence in a 25-dimensional SuperBigFLICA analysis. **B.** The contribution of different modalities within each of 25 SuperBigFLICA components for predicting 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).

components on predicting fluid intelligence scores ($B$ in Eq. (3.2)). 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.3.5B** shows the contribution of each component and each modality to the prediction of fluid intelligence (Method section 3.2.2)). 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. 3.5C** 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., 2021b], 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].

**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.15** 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.16** 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 *age started wearing glasses or contact lenses* are mainly located in visual areas, especially for resting-state dual regression spatial maps 5, which represents the visual network.

## 3.5 Discussion

In this chapter, 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., 2021b]. 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.

## 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.

# 4

# Individualised Prediction of Longitudinal Changes in Brain Images

## Contents


**4.1 Introduction . . . . . 74**
**4.2 Methods . . . . . 77**
4.2.1 A general mathematical framework for longitudinal brain images. 77
4.2.2 Deep U-Net Model for longitudinal prediction . . . . . 80
4.2.3 Data . . . . . 83
4.2.4 Implementation details. . . . . 84
4.2.5 Prediction of follow-up image with baseline image . . . . . 85
4.2.6 Prediction of longitudinal changes with baseline image . . . . 86
**4.3 Results . . . . . 89**
4.3.1 Prediction performance of follow-up image . . . . . 89
4.3.2 Prediction performance of longitudinal changes . . . . . 92
4.3.3 Test of whether a model performs change prediction or data denoising . . . . . 94
4.3.4 Discriminability of prediction . . . . . 95
4.3.5 Relations between predicted images and nIDPs . . . . . 96
4.3.6 Real data qualitative analysis . . . . . 98
**4.4 Discussion . . . . . 98**

## Summary

It remains largely unknown whether we can predict individualised longitudinal changes of brain imaging features based only on the baseline brain images, which would be of great value for longitudinal brain studies and early design of personalised treatments. We explore this possibility of using longitudinal data of 50 modalities in UK Biobank brain imaging studies with around 3500 subjects. We designed a deep U-Net based model for predicting longitudinal changes in multimodal brain images, and proposed a new mathematical framework for evaluating the longitudinal prediction results, including detecting the existence of changes in data, deriving baseline prediction performance and distinguishing change prediction from data denoising. Our results show that the proposed model can predict to a modest degree individualised longitudinal changes in almost all 50 modalities, and outperforms other popular models. Further, the predicted longitudinal changes have a similar or even improved accuracy of predicting subjects' non-imaging phenotypes, and have a high between-subject discriminability. Our results show the potentials of our model for longitudinal data imputation, and show several caveats when performing longitudinal data analysis.


## 4.1 Introduction

The brain's structure and function gradually change over time, and these changes are reflected in (and potentially can then be predicted by) behavioural, clinical, cognitive and demographic measures (known as non-imaging derived phenotypes or nIDPs). Accurately detecting these brain changes can not only help the early design of personalised treatments for brain disorders but also bring new neuroscientific insights into brain-behaviour relationships [Mills and Tamnes, 2014, Douaud et al., 2021] and, from a practical perspective, may even allow for longitudinal data imputation. However, longitudinal studies of the brain are much more difficult to conduct than cross-sectional studies, because of the need for rescanning subjects' brain image data, which are expensive and difficult to acquire. Therefore, typically we need to use subjects with both baseline and longitudinal data to perform longitudinal analysis, which reduces the usable range of datasets and/or subjects

(for example, to date, UK Biobank has released data from over 40,000 subjects' first imaging scan, but only about 3,500 of these also had a second scan). One possible way to solve this problem is to synthesise longitudinal brain image data based only on baseline images using machine learning techniques and then perform the longitudinal analysis based on the synthesised images. However, it is largely unknown whether the underlying longitudinal changes of the brain can be predicted.

Previous work has shown that, for some neuroimaging modalities, it is possible to do longitudinal prediction. For example, [Xia et al., 2021, Da Silva et al., 2021, Gafuroğlu et al., 2018] showed that future T1-weighted (T1w) brain images could be synthesised based on one's baseline T1w image. [Rachmadi et al., 2020] showed that white matter hyperintensities' (WMH) evolution 1-year later could be predicted by the baseline WMH observed in T2-fluid attenuated inversion recovery (T2-FLAIR) images. [Hong et al., 2019] also reported that infant diffusion MRI data can be predicted longitudinally. Other studies also showed that the brain functional connectome can also be predicted from the baseline connectome [Nebli et al., 2020, Ghribi et al., 2021]. However, these studies only focus on one or two modalities with a few subjects, so their models may not be generic. In addition, these studies only focus on the prediction of follow-up images. Little attention is paid to predicting the temporal changes, which is a key feature of interest in longitudinal studies. Therefore, a more comprehensive study is needed to explore the possibility of longitudinal brain image prediction, and a more general framework should broaden our understanding of brain development and ageing.

In this study, to overcome the above-mentioned limitations, we seek to answer the following questions: *Is it possible to predict subjects' follow-up brain imaging data and their longitudinal changes based only on their baseline brain images?* Given this central question, we proposed the following sub-questions to answer: (1) What is a suitable definition of longitudinal changes in brain imaging? (2) Can we detect the existence of longitudinal changes given our observed data? (3) As baseline and follow-up images are intrinsically similar given short follow-up time intervals, what is the "null" correlation between predicted images and true images? (4) Can we know a trained model is making change prediction rather than simple data denoising? (5) Does our method outperform

other state-of-the-art approaches? (6) Do the predicted maps capture real change patterns that are related to subjects' nIDPs, and are the predictions individualised?

We used the UK Biobank brain imaging dataset to address these questions, from which we extracted 50 different modalities from subjects' structural, functional and diffusion magnetic resonance imaging (MRI) scans from over 3,500 subjects longitudinally, with one baseline and one (on average two years later) rescan available (**Table 4.1**). First, we propose a mathematical modelling framework that models the longitudinal changes as a nonlinear function of the baseline image. This forms the basis of why we were able to predict longitudinal changes. Building on this framework together with an empirical estimation of data test-retest reproducibility, we show that we can (1) detect whether there exist temporal changes in the imaging data; (2) derive empirical baseline correlation of temporal change prediction; (3) test whether a model is really achieving change prediction or denoises the data. We then proposed a deep learning-based model, based on 3D U-Net [Çiçek et al., 2016], to perform longitudinal prediction. The motivation of using U-Net based structure is because U-Net uses skip connections in corresponding feature maps with the same spatial resolutions in the downsampling and upsampling layers, which assist the model to focus on learning the changes in the follow-up images compared with the baseline.

We show that the proposed U-Net model is generic and easy to train from scratch. It attempts to predict each modality's second scan from baseline scans. It outperforms other state-of-the-art linear prediction models, such as Supervised BigFLICA [Gong et al., 2021a], and other nonlinear approaches including the Variational Autoencoder (VAE) [Kingma and Welling, 2013], other variants of U-Nets (e.g., Attention U-Net [Oktay et al., 2018], U-Net++ [Zhou et al., 2018] and Nonlocal U-Net [Wang et al., 2020]), and variants of generative adversarial networks (GANs) (e.g., Conditional GAN [Isola et al., 2017]). This is in agreement with the conclusion in a recent paper on the performance of U-Net in biomedical imaging tasks [Isensee et al., 2021]. We evaluate the model's performance by showing that it can meaningfully predict longitudinal changes in almost all 50 modalities in the UK Biobank, with the follow-up time interval being about two years. The prediction correlations are significantly larger than the chance level. We quantitively and qualitatively

show that our model can predict known change patterns in brain images, such as increased brain ventricles' sizes observed in T1-weighted imaging derived modalities. The predicted longitudinal changes have a similar or even improved out-of-sample predictive power for nIDPs, and can serve as a brain fingerprint of a subject [Finn et al., 2015]. Therefore, our prediction model can aim to impute imaging data of missing subjects in longitudinal studies and perform early prediction of changes in brain imaging biomarkers.

## 4.2 Methods

### 4.2.1 A general mathematical framework for longitudinal brain images.

Let the baseline image be $A$ and its follow-up image $B$. We assume both $A$ and $B$ have been vectorized to size $v \times 1$ vectors, where $v$ is the number of voxels. The model for baseline and follow-up images is:

$$\begin{aligned} A &= A_0 + E_A \\ B &= B_0 + E_B \end{aligned} \tag{4.1}$$

where $A_0$ and $B_0$ are true underlying signals for the baseline and follow-up images, and $E_A \sim N(0, \sigma^2)$ and $E_B \sim N(0, \sigma^2)$ are their noise terms. We assume baseline and follow-up images have the same level of noise.

The reason why we would expect we can use the baseline image $A$ to predict the follow-up image $B$ is that the temporal change might be a function of $A_0$:

$$\begin{aligned} B_0 &= A_0 + \Delta \\ \Delta &= WA_0 + f(A_0) \end{aligned} \tag{4.2}$$

where $W$ is a linear transformation matrix (size $v \times v$) of signal image $A_0$, and $f(A_0)$ is a nonlinear function of $A_0$. Here, we have decomposed the changes into linear and nonlinear ones, which will be useful later. Note that common linear decomposition methods can be seen as learning $W$. For example, if we apply a partial least square regression to predict $\Delta$ using $A$, the product of coefficients of dimension reduction parts and prediction parts will approximate $W$. We further assume that the variance of $A_0$ is 1, the mean of $A_0$ and $f(A_0)$ is 0. This means that the brain image will change based on the pattern of $A_0$.

**A model of longitudinal changes: the "null" prediction correlation**

First, we derive the "null" correlation between temporal difference, i.e., $B-A$, and baseline $A$. This means the correlation we will get when there is actually no longitudinal change and we simply copy $-A$ as the prediction of $B-A$.

Based on the model above, the correlation between temporal difference $B-A$ and baseline $A$ is:

$$\begin{aligned} r(B-A,A) &= r(WA_0+f(A_0)+E_B-E_A,A_0+E_A) \\ &\approx \frac{\mathrm{cov}(WA_0,A_0)-\sigma^2}{\sqrt{(\mathrm{var}(WA_0)+\mathrm{var}(f(A_0))+2\sigma^2)(1+\sigma^2)}} \end{aligned} \tag{4.3}$$

If there are no temporal changes ($\Delta=0$), the above correlation reduces to:

$$\begin{aligned} r(B-A,A) &= r(E_B-E_A,A_0+E_A) \\ &= -\frac{\sqrt{2}}{2}\frac{1}{\sqrt{1+1/\sigma^2}} \end{aligned} \tag{4.4}$$

which will vary between $(-\frac{\sqrt{2}}{2},0)$ with $\sigma\in(0,+\infty)$. In practice, we can estimate $\sigma$ with various methods. For example, for resting-state fMRI, we can estimate the $\sigma$ based on the split-half reproducibility. As shown in the red bars in Fig. 4.4, an estimate of test-retest reproducibility for resting-state dual regression spatial map 1 is 0.34. This means that, let $B_1$ and $B_2$ be test and retest scans:

$$\begin{aligned} B_1 &= B_0+E_{B1} \\ B_2 &= B_0+E_{B2} \\ r(B_1,B_2) &= \frac{1}{1+\sigma^2_{\text{half}}} = 0.34 \\ \sigma^2 &= \sigma^2_{\text{half}}/2 \end{aligned} \tag{4.5}$$

Therefore, we can get $\sigma^2=0.97$. Take this back into $\Delta=0$ case (Eq. 4.4), we get the "null" correlation $r(B-A,A)_{\text{null}}\approx -0.49$.

**A model of longitudinal changes: does temporal change exist?**

We then propose another question: can we judge whether temporal change exist? To do this, we rely on the difference between the observed baseline-follow up correlation and the "null" correlation. We calculate the observed $r(B-A,A)$. As shown in the Fig. 4.5, it is around -0.63 for resting-state map 1.

How does this -0.63 arise? From Eq. 4.3 ($\Delta \neq 0$), we see that we can only get a $r$ that smaller than $\Delta = 0$ if $\text{cov}(WA_0, A_0) < 0$. This means that in population, the linear part of temporal changes, i.e., $WA_0$, should be negatively correlated with $A_0$.

Similar results can be derived from the correlation between baseline and follow-up images:

$$\begin{aligned} r(B,A) &= r(A_0 + WA_0 + f(A_0) + E_B, A_0 + E_A) \\ &\approx \frac{\text{cov}((I+W)A_0, A_0)}{\sqrt{(\text{var}((I+W)A_0) + \text{var}(f(A_0)) + \sigma^2)(1+\sigma^2)}} \end{aligned} \tag{4.6}$$

The data split-half reproducibility reflects the correlation when $\Delta = 0$, which is:

$$r(B,A) = r(A_0 + E_B, A_0 + E_A) = \frac{1}{1+\sigma^2} = 0.34 \tag{4.7}$$

This correlation should be even higher if we use full data, instead of split-half data. However, the observed correlation between baseline and follow-up $r(B,A)$ is only 0.23 (Fig. 4.4). The $\Delta$ decreased if $\text{cov}(WA_0, A_0) < 0$ and/or there are nonlinear changes.

**Change prediction or data denoising?**

In predicting B as PredB with A, suppose our model is $g(\cdot) = g_1(\cdot) + g_2(\cdot)$, ($g_1(\cdot)$ does change prediction and $g_2(\cdot)$ does denoising), then :

$$\begin{aligned} r(\text{PredB}, B) &= r(g(A), B) \\ &= r(g_1(A_0) + g_2(E_A), (I+W)A_0 + f(A_0) + E_B) \end{aligned} \tag{4.8}$$

If $g_1(\cdot)$ learns nothing but is simply a scaled copy of $A_0$ and only $g_2(\cdot)$ is active, we have:

$$\begin{aligned} r(\text{PredB}, B) &= r(g(A), B) \\ &= r(g_1(A_0) + g_2(E_A), (I+W)A_0 + f(A_0) + E_B) \\ &= r(A_0 + g_2(E_A), (I+W)A_0 + f(A_0) + E_B) \\ &= \frac{\text{cov}(A_0, (I+W)A_0)}{\sqrt{(1 + \text{var}(g_2(E_A))(\text{var}((I+W)A_0) + \text{var}(f(A_0)) + \sigma^2))}} \\ &= \frac{1}{\sqrt{1 + \text{var}(g_2(E_A))}} \frac{\text{cov}(A_0 B_0)}{\text{std}(B)} \end{aligned} \tag{4.9}$$

where PredB is the predicted follow-up image. So it is possible that $r(\text{PredB}, B) > r(B,A)$ if the $g(\cdot)$ only denoises $A$.

Next, we made an assumption that the pattern of change between ages $x-2$ and $x$ is similar to that between $x$ and $x+2$ ($C$ is the 4 years follow-up image):

$$\begin{aligned} &C = C_0 + E_C = B_0 + \Delta + E_C \\ &\Delta = WB_0 + f(B_0) \\ &r(C,B) \approx r(B,A) \rightarrow r(C,B) \approx r(B,C-2\Delta) \\ &\rightarrow \frac{\text{cov}(B_0C)}{\text{std}(C)} \approx \frac{\text{cov}(B_0(C-2\Delta))}{\text{std}(C-2\Delta)} \end{aligned} \tag{4.10}$$

Then if we feed $B$ into $g(\cdot)$ (suppose the output is PredA), if $g(\cdot)$ only denoises the data, we have:

$$\begin{aligned} r(\text{PredA},A) &= r(g(B),A) \\ &= r(g_1(B_0) + g_2(E_B), C-2\Delta) \\ &= r(B_0 + g_2(E_B), C-2\Delta) \\ &\approx \frac{1}{\sqrt{1+\text{var}(g_2(E_B))}} \frac{\text{cov}(B_0(C-2\Delta))}{\text{std}(C-2\Delta)} \\ &\approx \frac{1}{\sqrt{1+\text{var}(g_2(E_B))}} \frac{\text{cov}(B_0C)}{\text{std}(C)} \\ &\approx \frac{1}{\sqrt{1+\text{var}(g_2(E_B))}} \frac{\text{cov}(A_0B)}{\text{std}(B)} \\ &\approx r(\text{PredB},B) \end{aligned} \tag{4.11}$$

If $g(B)$ is closer to $C$ than $A$ ($g$ is doing change prediction), then $r(\text{PredB},B) > r(\text{PredA},A)$.

### 4.2.2 Deep U-Net Model for longitudinal prediction

#### The proposed Model

Based on the above mathematical model, we will construct a machine learning model that can learn the longitudinal changes (i.e., $W$ and $f(\cdot)$). We did not choose a linear model here, and therefore choose to use a deep U-Net structure for this task.

U-Net is a popular deep neural network originally designed for image segmentation. We adopt a U-Net structure here with a modified voxel-wise mean square error loss to perform longitudinal prediction. **Fig. 4.1** shows the network architecture. The input of the network is a baseline image, while the output is the predicted follow-up image. The network has a downsampling and an upsampling path, each with four layers. In the downsampling path, each layer contains two $3 \times 3 \times 3$ convolutions followed by

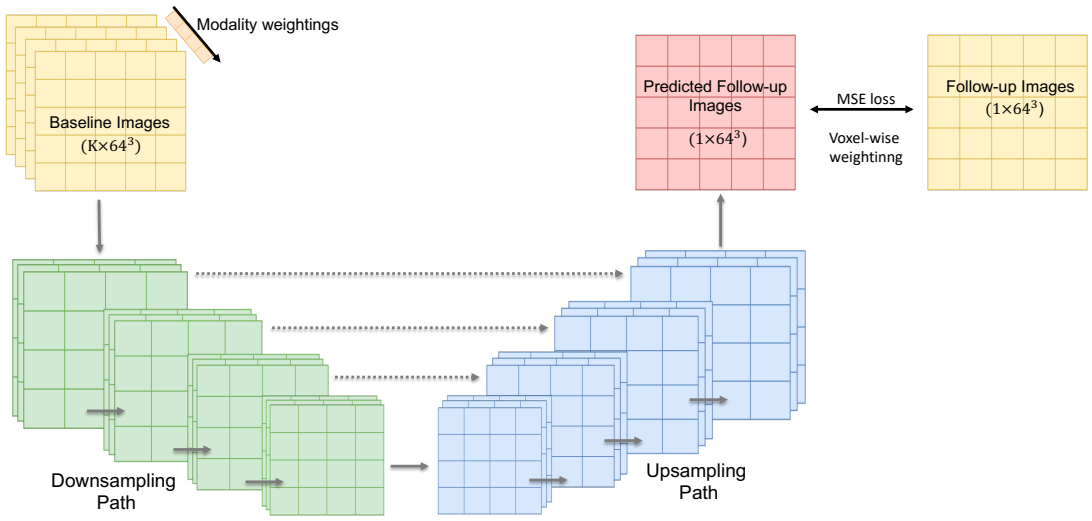



**Figure 4.1:** The proposed U-Net structure for longitudinal brain image prediction. The network has four downsampling and upsampling layers with skip connections. The numbers $a \times b^3$ represents the number of channels (a) and spatial voxel dimensions (b) of the inputs and outputs.

batch normalisation and a rectified linear unit (RELU), and followed by a $3 \times 3 \times 3$ average pooling layer with strides of 2 in each dimension. In the upsampling path, each layer contains a $3 \times 3 \times 3$ upconvolution, followed by two $3 \times 3 \times 3$ convolutions and a batch normalisation and a RELU. Skip connections are added between feature maps in the downsampling and the upsampling path with the same spatial resolution, which is one of the key features that differs from Autoencoder-based models. The number of channels doubles in each layer of the downsampling path while becoming halved in each layer of the upsampling path. In the final layer, a $1 \times 1 \times 1$ convolution reduces the number of channels to the target output.

When we use multiple (denoted as $K$) baseline modalities to do longitudinal prediction, the baseline images are simply concatenated, forming a four-dimensional input to the U-Net. Therefore, the number of channels in the first layer of the U-Net becomes $K$. We also add a positive weight vector to the network to learn the contributions of different modalities to the prediction. The length of the vector equals $K$, each of the elements is multiplied to one modality map before the first U-Net downsampling. Therefore, after the optimisation, modalities with higher contributions to the prediction of follow-up images should have higher weights. The losses are only calculated in voxels inside the brain.

We used mean square error at each voxel as the loss function, because our target image is a continuous one. To combine losses of multiple voxels, we use a weighted sum of voxel-wise mean square error. The weights of voxels are learned based on their prediction uncertainty [Kendall et al., 2018]. Specifically, let the loss function at voxel $i$ be $l_i$, then the overall loss is:

$$L = \sum_{i=1}^{v} \left( \frac{l_i}{2\sigma_i^2} + \log(\sigma_i + 1) \right) \tag{4.12}$$

where $\sigma_i > 0$ is a learnable parameter. This idea was originally developed for multi-task learning [Kendall et al., 2018]. Here, we treat the prediction at each voxel as a different 'task'. The weights among voxels 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 with larger error/uncertainty will be given lower weights [Kendall et al., 2018]. The performance is improved compared with using MSE loss without voxel-wise weighting empirically, and the learning of weighting parameters is efficient.

**Comparing with other approaches**

There are broad classes of models that can perform longitudinal prediction. In classical machine learning literature, partial least square regression can perform longitudinal prediction, where the input and output are vectorised versions of image data. We choose a recent method called SuperBigFLICA [Gong et al., 2021a] as a baseline linear method. In addition, the Variational Autoencoder (VAE) [Kingma and Welling, 2013] is a Bayesian generative model widely used for image generation and denoising. We adopt it to the longitudinal prediction as another baseline approach. Similarly, another comparison is against conditional generative adversarial networks (cGANs), which uses adversarial losses for image synthesis[Isola et al., 2017]. We also compared the proposed U-Net structure with more advanced variants of U-Net, i.e., Attention U-Net [Oktay et al., 2018], U-Net++ [Zhou et al., 2018] and Nonlocal U-Net [Wang et al., 2020].

In SuperBigFLICA, We added skip connections from the input directly to the output, and used the Adam optimiser [Kingma and Ba, 2014] for all learnable parameters (instead

of using combined optimisers as done in the original work), which empirically improves the performance. For the VAE, we adopt the same structure as the proposed U-Net, with skip connections removed, while in cGANs, the structure of the generator network is the same as our proposed U-Net, and the weight between adversarial loss and voxel-wise MSE loss varies between 0 to 1. The same backbone structure as the proposed U-Net is used in the advanced U-Nets, with different attention modules added.

### 4.2.3 Data

Voxel-wise neuroimaging data of 50 modalities from 3,572 subjects were used in this chapter, including: (1) 25 "modalities" from the resting-state fMRI ICA dual-regression spatial maps [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* (COPE) 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) 3 T1-MRI derived modalities (standard space T1w image; grey matter volume; 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 the nonlinear warp maps (which reflects the shape changes in the nonlinear registration of individual T1w image to the MNI152 standard space in each voxel in the x-, y-, and z-directions, so that it has 3 channels)); (5) 1 susceptibility-weighted MRI map (T2-star image); (6) 2 T2-FLAIR MRI derived modality (standard space T2-FLAIR image; white matter hyperintensity map estimated by BIANCA [Griffanti et al., 2016]). The UK Biobank imaging data were mainly preprocessed using FSL [Smith et al., 2004, Jenkinson et al., 2012] following an optimized processing pipeline[Alfaro-Almagro et al., 2018] (`https://www.fmrib.ox.ac.uk/ukbiobank/`). A description of these 50 modalities is shown in **Table 4.1**.

All 3,532 subjects have both baseline and follow-up T1w images, and a small proportion of them have some other missing modalities. The follow-up images are scanned on average two years after the baseline images. We also used 15,697 nIDPs from

the UK Biobank. We performed a z-score normalization across subjects for each nIDP. We did not regress out confounding variables from brain imaging data and nIDP data. We exclude an nIDP if it has more than 1000 missing values in these 3,532 subjects or is constant in the training set (e.g., for many disease variables that have only a few subjects affected, so it is possible that no disease affected subjects are in the training set).

**Table 4.1:** A description of 50 Modalities of UKB dataset used in this chapter.

| 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 (COPE) of emotion task contrast "shapes" |
| task c2 | Contrasts of parameter estimate (COPE) of emotion task contrast "face" |
| task c5 | Contrasts of parameter estimate (COPE) 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 |
| 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 |
| T1 nonlinear | T1 weighted image nonlinearly registered to MNI152 standard space |
| T2 FLAIR | T2 FLAIR image nonlinearly registered to MNI152 standard space |
| T1 warp (x,y,z) | The nonlinear warp field of T1 image registered to MNI152 standard space |

### 4.2.4 Implementation details.

The U-Net model is implemented using the Pytorch framework. The input and output are first downsampled to $64 \times 64 \times 64$ images. The intensity of the input image is then normalised to between zero and one by min-max normalisation before being fed into the model. The intensity of the output image is not normalised. We did not perform any data augmentation, because empirically, it does not improve model performance. We used the Adam optimizer [Kingma and Ba, 2014] for parameters optimization. The learning rate is set to 0.001, and all other parameters in Adam are set as default. The mini-batch size is 16 subjects. The total number of epochs (number of times the full data passes through

the model) is 40, the learning rate decreases by a cosine rule. The model weights are initialised by Gaussian-distributed random numbers of mean 0 and variance 1.

A random subset of 2,500 subjects is used in training, while another random subset of 200 subjects is used as validation, and the remaining 852 subjects are used for testing and reporting the performance. The test set size is different for different modalities due to different missing data proportions. All evaluations are based on the same train/validation/test splitting. We used mean Pearson correlation (r) and the mean absolute error (MAE) between the predicted and the true images as the evaluation criteria. Before calculating these two metrics, we excluded voxels outside the brain and vectorised the 3-dimensional image.

### 4.2.5 Prediction of follow-up image with baseline image

We denote the baseline brain image of a modality as A, and its follow-up image as B. The most intuitive approach is to feed A into our model and predict B. Although there is no problem in theory, in practice, this may significantly bias the model towards learning to predict the population average of B. This is because, for all modalities, there is a population average pattern that can be highly similar to each subject's image. For example, in task contrast maps, every subject tends to have activations in similar brain areas. The population average pattern can dominate each subject's image so that a 'clever' deep network will find that simply predicting the mean can have a low mean square error loss. There are two ways to eliminate this kind of population average effect, and both of which have shown success in predicting task activation maps using resting-state fMRI data. The first is to use contrastive loss [Ngo et al., 2021], which is a loss function that simultaneously minimises both the prediction error and the similarity among the predicted maps of different subjects. The second approach is residualisation, which regresses the population average map out of each of the individual images using linear regression, and then predicts based on the residualised maps [Zheng et al., 2021]. The first approach would require fine-tuning of a balance parameter between two losses, which is computationally inefficient. We, therefore, used the second approach to enforce learning parameters that characterise longitudinal change at an individual level.

In detail, we will denote the residualized baseline image as ResA, and the residualized follow-up image as ResB. Let $\bar{A}$ and $\bar{B}$ be the population average image in the training set, then each subject's ResA and ResB are obtained as:

$$\begin{aligned} \mathrm{A} &= \bar{A}\hat{\beta} + \mathrm{ResA} \\ \mathrm{ResA} &= A - \bar{A}\hat{\beta} \\ \mathrm{ResB} &= B - \bar{A}\hat{\beta} \end{aligned} \tag{4.13}$$

Note that here A and B are vectorised and non-brain voxel removed before being fed into the linear regression. $\hat{\beta}$ is the least-square estimate of coefficient when regressing $\bar{A}$ out of A. Then this $\hat{\beta}$ is applied to estimate ResA and ResB. In practice, we can also use the coefficient of regressing out $\bar{B}$ from B to generate ResB, but we find there is little difference. Furthermore, one added advantage of using $\bar{A}$ only is that we can reconstruct the estimated original B from the predicted ResB without knowing the true original B.

### 4.2.6 Prediction of longitudinal changes with baseline image

The success of predicting follow-up images does not necessarily mean we can accurately capture the longitudinal change patterns. In addition, in longitudinal analysis, we are often interested in how the temporal changes in imaging features relate to nIDPs or changes in nIDPs. Therefore, we explicitly explore the possibility of predicting the longitudinal changes using the baseline image. Basically, we would propose three problems to solve here: (1) What is a meaningful definition of longitudinal changes in images; (2) Can we prove the existence of longitudinal changes in the data; (3) Can we prove our model is actually doing change prediction rather than simple data denoising.

#### Definition of longitudinal changes

For the first problem, i.e., the definition of longitudinal changes, we propose two metrics: The first one is ResB-ResA, which we call the *temporal difference*, and the second is ResB\ResA, i.e., regressing ResA out of ResB, which we call *residual temporal difference*. The temporal difference is the most intuitive definition. However, as can be seen from its definition, it is correlated with -ResA, which means that, even if the model estimates ResB=0, we will get a non-zero prediction correlation. In Method section 4.2.1, we show

that based on our mathematical model, in the null case (i.e., there are no longitudinal changes), the correlation between ResB-ResA and ResA is a function of signal-to-noise ratio (SNR) of the image data. As a result the null correlation can vary between 0 to $-\frac{\sqrt{2}}{2}$. We can empirically estimate the SNR of data by examining the test-retest reproducibility of the data (Method section 4.2.1). For example, the test-retest correlation of resting-state dual regression map 1, i.e., the default mode network, is 0.34, so its estimated null correlation is -0.49. This can serve as a baseline prediction performance of our proposed model if we use this definition of temporal difference.

In order to avoid this problem, we propose another metric for defining temporal difference: ResB\ResA.

$$\begin{aligned} \text{ResB} &= \text{ResA}\hat{\beta} \\ \text{ResB}\backslash\text{ResA} &= \text{ResB} - \text{ResA}\hat{\beta} \end{aligned} \tag{4.14}$$

where $\hat{\beta}$ is the least-square estimate of coefficient when regressing ResA out of ResB. Using this definition, the ResA has zero linear correlation with ResB\ResA, so that in the null case of no temporal changes, the correlation between ResB\ResA and ResA is exactly zero. The intuitive interpretation of ResB\ResA is: the information in ResB that can not be linearly represented by ResA. Therefore, if we choose this metric, we aim to examine the prediction performance of the nonlinear part of the changes, i.e., $f(A_0)$. This is because, in the regression, we not only regress out the original ResA in the baseline data but also regress out the linear part of temporal changes. A further concern is whether we can accurately remove all linear parts related to $A_0$. A concern on ResB\ResA may arise due to the regression dilution problem: when ResA is a noisy modality (e.g. fMRI), ResB\ResA may still contain information in $A_0$ because of the underestimates of regression coefficients.

**Existence of longitudinal changes**

For the second problem, we propose a method to prove the existence of longitudinal changes in the data empirically. The method (detailed in Method section 4.2.1) compares the observed correlation between ResA and ResB (or ResB-ResA and ResA), with their null correlations. For example, with a test-retest reproducibility of 0.34 for the

default mode network map, we observed an empirical correlation between baseline and follow-up with a smaller correlation of 0.23. Therefore, by analysing the expression for the correlation between ResA and ResB with longitudinal changes (Eq. 4.6), we conclude that there is a negative correlation between $WA_0$ and $A_0$ or the existence of the nonlinear part of longitudinal change.

**Evaluation of model prediction**

Four complementary methods are proposed to evaluate the proposed model in different ways. The first is the most straightforward, which examines the model's prediction performance by computing the correlation or MAE between the predicted image and true image, and comparing its value with the null case. This evaluation is applied in both follow-up images and residual temporal difference images.

The second method overcomes a potential limitation of the first one. In the first, we can still not distinguish whether the model can truly predict longitudinal changes or simply perform data denoising. We show in the Method section 4.2.1 that even if the deep model is simply denoising the data, we can potentially get a higher correlation value than the null case. We, therefore, propose a test to evaluate whether the model is doing change prediction: Given a trained deep network $g(\cdot)$, we feed the baseline image ResA and the follow-up image ResB into it separately, getting the outputs of the network as PredResB and PredResA (respectively). We then compare the correlation values $r$(ResB, PredResB) against $r$(ResA, PredResA). If the network is purely denoising, we show in the Method section 4.2.1 that these two metrics should be very similar. On the contrary, if the former is larger than the latter, we show that the network can predict the longitudinal change. We further compare the above two correlations using the residual temporal differences maps, i.e., $r$(ResB$\setminus$ResA, PredResB$\setminus$ResA) against $r$(ResA$\setminus$ResB, PredResA$\setminus$ResB).

The third method evaluates the discriminability of predicted maps based on ideas originally developed in the functional connectome fingerprint [Finn et al., 2015]. A model can achieve individualised prediction if the predicted spatial maps can be a subject's fingerprint, i.e., the predicted maps correlate most to the subjects themselves but less so with other subjects. We tested if the predicted maps could be identified based on

their correlation with the true maps. We estimate the mean identification rate as the proportion of subjects whose within-subject correlation is the largest among all between-subject correlations. This evaluation is applied in both follow-up images and residual temporal difference images.

The fourth method compared the prediction performance of nIDPs between the predicted and the true residual temporal difference image. As our goal in brain image analysis is usually linking image data to non-imaging variables, we want to test whether the predicted images have a similar or even improved predictive power of nIDPs compared with the original images. Specifically, for each modality, we calculated PredResB\ResA and ResB\ResA. Then we used principal component analysis to reduce the residual temporal difference data into 500 orthogonal components, and then applied elastic-net regression, from the glmnet package [Zou and Hastie, 2005], to predict each of the nIDPs in the UK Biobank. Elastic-net regression 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 use the residual temporal difference image here because we want to eliminate all the effects of the baseline image on predicting nIDPs. In temporal difference image, i.e., ResB-ResA, we are not able to achieve this.

## 4.3 Results

### 4.3.1 Prediction performance of follow-up image

**Fig. 4.2** shows a comparison of different approaches for longitudinal prediction in each of the 50 modalities. The input is the baseline image, and the target is the follow-up image for one specific modality. Both the baseline and follow-up images have been residualised. We can see that U-Net based approaches achieve the best performances. More advanced U-Nets do not show any significant advantages compared to our proposed U-Net, and they even perform worse for many TBSS modalities. All of the U-Net based methods outperform simply copying the baseline image as the prediction of follow-up image (shown in **Fig. 4.3**). The other three methods, i.e., VAE, cGANs and SuperBigFLICA, perform significantly worse than the U-Net based approaches. Although successful

in many computer vision tasks, the failure of VAE and cGANs here may be that they are not well-suited to the noisy, high-dimensional and small-sample size brain image data. The failure of SuperBigFLICA may be due to its linear nature so that important nonlinear information is lost. Note that the **Fig. 4.2** does not demonstrate whether the methods are successful in estimating longitudinal change, as it only shows the extent to which the second scan is predicted (which might be just the aspects of the second scan that are similar to the first).

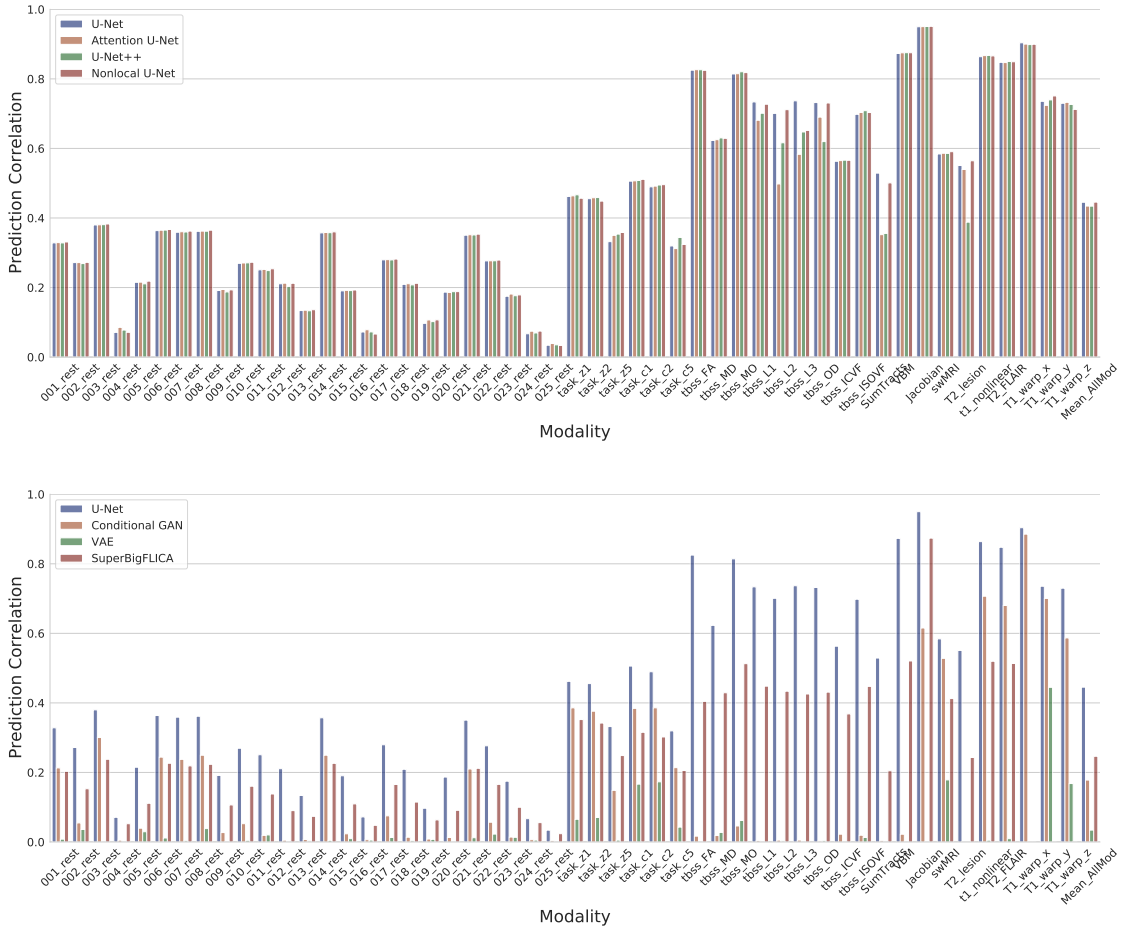



**Figure 4.2: Comparing the proposed U-Net approaches with other variants of U-Nets (Top) and other state-of-the-art methods (Bottom).** Correlations are between the actual residualised second scan against the predicted residualised second scan. "Residualisation" refers to the act of regressing out the group mean map (here the group mean first-scan image), to focus on aspects of subject maps that are distinct from each other. The mean correlation among all modalities are shown on the rightmost.

**Fig. 4.3** shows the prediction correlation of the proposed approach with all 50 multimodal images as inputs, and compares it with single-modality inputs and just copying

the baseline image. The multimodal U-Net has a small improvement in performance compared with single-modal input, indicating that the complementary features in different modalities can be learned to assist longitudinal prediction, but the primary information contributing to the prediction exists in the same modality as the target. The fact that the U-Net outperforms simply copying the input image to the output does not necessarily demonstrate successful longitudinal (change) prediction, because the improvement might also be due to image denoising.

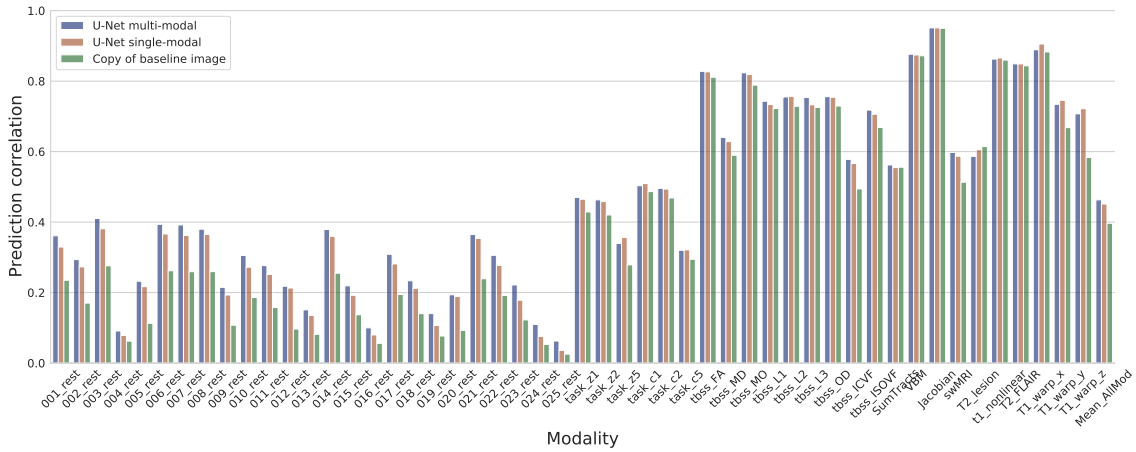



**Figure 4.3:** Prediction correlation of follow-up images (i.e., ResB) with multimodal and single modal U-Net, and a baseline of copying baseline image (i.e., ResA) as the prediction, i.e., r(ResB, PredResB | ResA). The mean correlations among all modalities are shown on the right.

**Fig. 4.4** compares the test-retest correlation of 25 resting-state components (estimated by splitting the time series of the follow-up data into two) with the prediction correlation of follow-up image and with the observed correlation. As shown by the mathematical model in the Method section 4.2.1, a lower correlation between baseline and follow-up than the test-retest reproducibility indicates temporal changes. The changes may be linearly and negatively related to the baseline and nonlinearly related to the baseline.

For the comparison between prediction correlation and test-retest reproducibility, theoretically, among 25 resting-state modalities, 8 of these have a larger prediction correlation than the test-retest reproducibility when using multimodal input, while 2 of them when using single-modal input. As the input baseline image has a similar SNR to the follow-up image, we would expect a prediction correlation that is similar to the full-data test-retest reproducibility when the underlying temporal changes are perfectly predicted. Here, as

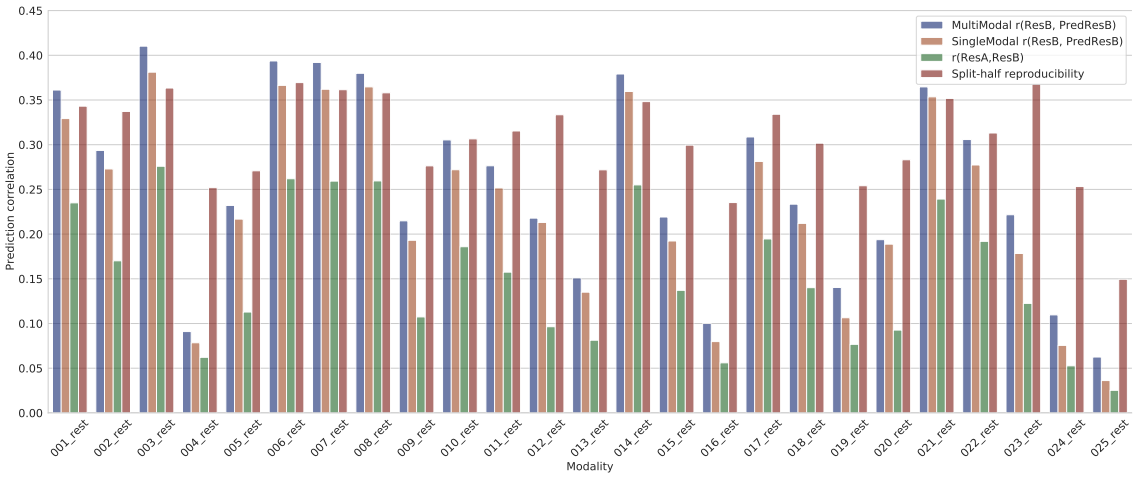



**Figure 4.4:** For 25 resting-state modalities, the prediction correlation of follow-up images (i.e., ResB) with multimodal and single modal U-Net, and a baseline of copying baseline image (i.e., ResA) as the prediction, and the test-retest reproducibility by splitting the time series in half. The mean correlations among all modalities are shown on the right. Among the 25 modalities, component 4, 23, 24, 25 have been identified as non-neural components.

we only used split-half data to estimate the reproducibility, the reproducibility should be lower than the full-data case. Therefore, we would expect most of the time, the prediction correlation of a single modality to be smaller than the full-data test-retest reproducibility. As a result, we can even observe that 23 of the 25 single modality prediction correlations are smaller than the split-half reproducibility. In multimodal prediction, we observe 8 modalities that can exceed the split-half reproducibility, indicating clear advantages of integrating complementary information across different modalities.

### 4.3.2 Prediction performance of longitudinal changes

**Fig. 4.5** shows the prediction correlation of temporal difference image, i.e., ResB-ResA, with multimodal and single modality input. Across all modalities, the accuracy of single modality prediction was lower but by only a small amount than multimodal prediction. **Fig. 4.6** further shows, for 25 resting-state components, the results for (1) the prediction performance of temporal difference image; (2) the empirically observed correlation between baseline image and the real temporal difference image; (3) the estimated null correlation between baseline image and temporal difference image when the longitudinal difference is zero; and (4) the split-half reproducibility. As we show from the mathematical model in the Method section 4.2.1, we observed a lower correlation of

the observed temporal difference than the null temporal difference, indicating a negative relation between the baseline image and the underlying longitudinal change. We observe that the prediction correlation of temporal difference is larger than the null correlation but only slightly higher than copying -ResA as the prediction. The latter indicates that the main temporal change may negatively correlate with the baseline on its own, and the nonlinear part of changes are small.

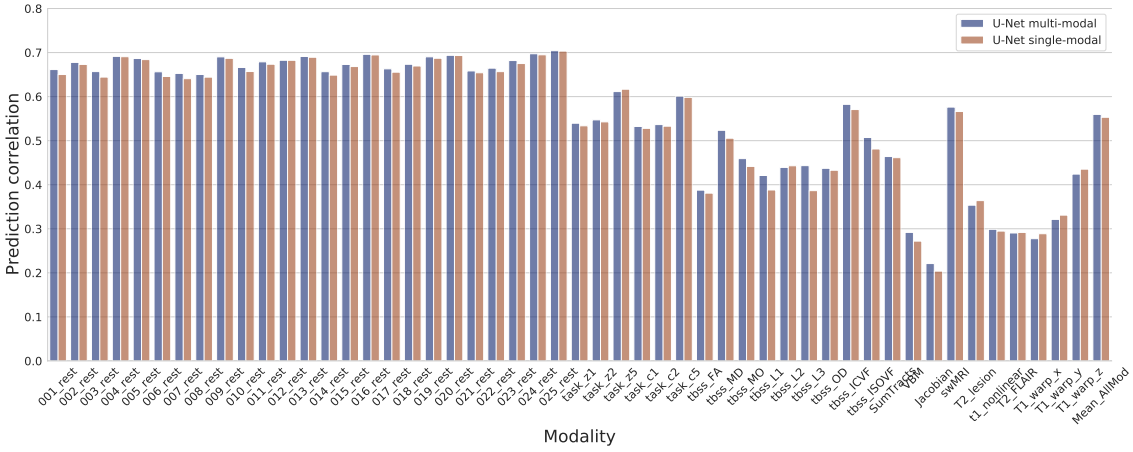



**Figure 4.5:** Prediction correlation of temporal difference images (i.e., ResB-ResA) with multimodal and single modal U-Net, i.e., r(ResB-ResA, PredResB-ResA | ResA). The mean correlations among all modalities are shown on the right.

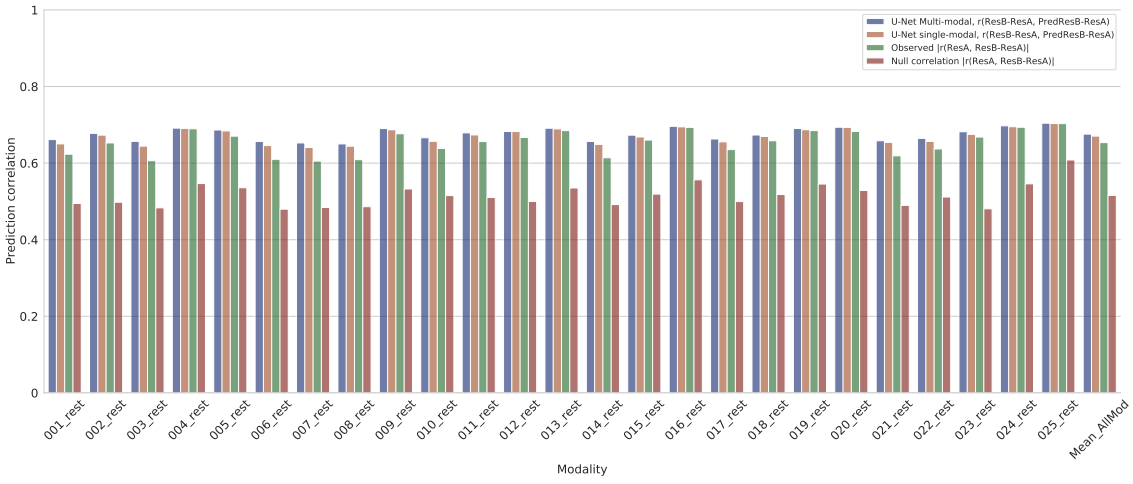



**Figure 4.6:** For 25 resting-state modalities, the prediction correlation of temporal difference image (i.e., ResB-ResA) with multimodal and single modal U-Net., and a baseline of copying baseline image (i.e., -ResA) as the prediction, and the estimated null correlation when there are no longitudinal changes by the mathematical model. The mean correlations among all modalities are shown on the right.

**Fig. 4.7** shows the results for predicting residual temporal difference, i.e., ResB\ResA. Both multimodal and single-modal U-Net have high prediction correlations. The baseline

correlation here is zero because all linear information related to the baseline image has been removed. We observe lower overall prediction correlation values for predicting residual temporal difference images, i.e., ResB\ResA, than temporal difference images, i.e. ResB-ResA. This may be owing to the residualisation also removing linear relations between baseline and longitudinal changes. In other words, we are trying to predict the nonlinear changes related to baseline images.

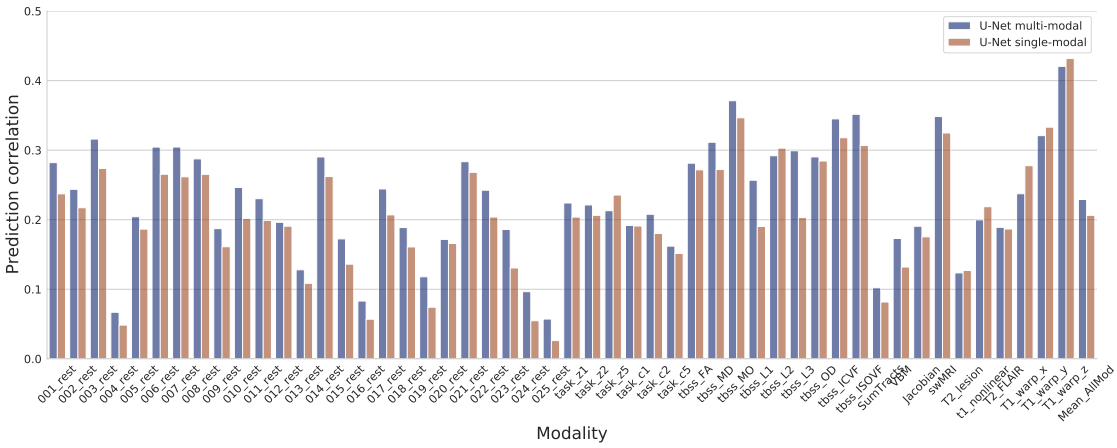



**Figure 4.7:** Prediction correlation of residual temporal difference images (i.e., ResB\ResA) with multimodal and single modal U-Net. Here the baseline correlation is zero because ResA has been regressed out, i.e., r(ResB\ResA, PredResB\ResA | ResA). The mean correlations among all modalities are shown on the right.

### 4.3.3 Test of whether a model performs change prediction or data denoising

We proposed to quantitatively test whether a trained model achieves longitudinal change prediction or denoising the data, by computing the correlation of output with baseline image fed in and the real follow-up image, i.e. *r*(PredResB, ResB) or r(PredResB\ResA, ResB\ResA), and comparing this against the correlation of output with follow-up image fed in and the real baseline image, i.e., *r*(PredResA, ResA) or r(PredResA\ResB, ResA\ResB). Put more simply, we estimate ResB from ResA and compare this against predicting ResA from ResB. As shown in Method section 4.2.1, the former will be larger than the latter if the model is doing change prediction, while the former is similar to the latter if it is doing only data denoising.

**Fig. 4.8** shows the results for follow-up images and residual temporal difference images. For most modalities, the correlations of feeding baseline images into the trained network are larger than the correlation of feeding follow-up images into the network. However, some modalities follow distinct patterns of results. For some rfMRI components (e.g., 1, 3, 6, 7, 8), A-predicts-B is a small amount better than B-predicts-A, and both are quite a lot better than A-equals-B (prediction is set equal to input); this suggests that for these modalities, there is a very small amount of succesful longitudinal prediction, and a much larger amount of input denoising. For other modalities (e.g., rfMRI 5), the order of the results is the same, but with more equal spacing, suggesting that the success of longitudinal prediction is more similar to the strength of the denoising. In some other modalities (e.g., rfMRI 4), all three analyses are similar and low, suggesting that almost no successful prediction or denoising is taking place. This example makes sense, as rfMRI 4 is one of the 4 rfMRI artefact components. Finally, in a few modalities (e.g., T1_warp_z), A-predicts-B is much higher than B-predicts-A, with A-equals-B in between; this suggests that logitudinal prediction is being highly successful, with little denoising taking place.

### 4.3.4 Discriminability of prediction

A model can achieve individualised prediction if the predicted spatial maps can be a subject's fingerprint, i.e., the predicted maps correlate most to the subjects themselves and are less correlated with other subjects. We tested whether the predicted maps could be identified based on their correlation with the true maps. **Fig. 4.9** shows that both the predicted follow-up images and the residual temporal difference images have a high discriminability approaching 1 (perfect subject identification rate) for many modalities, suggesting that our model can achieve individualised prediction. The high identification rate of PredResB may be driven by it simply looking like ResA, so we included residual temporal difference images in the same plot, which show that the PredResB\ResA are also "individualised". The PredResB\ResA has an overall lower discrimibility than PredResB; this may because the former has a worse prediction correlation so that less individual-level information is well captured. Finally, note that PredResB\ResA has a very low discriminability in some structure-related modalities, such as Jacobian maps

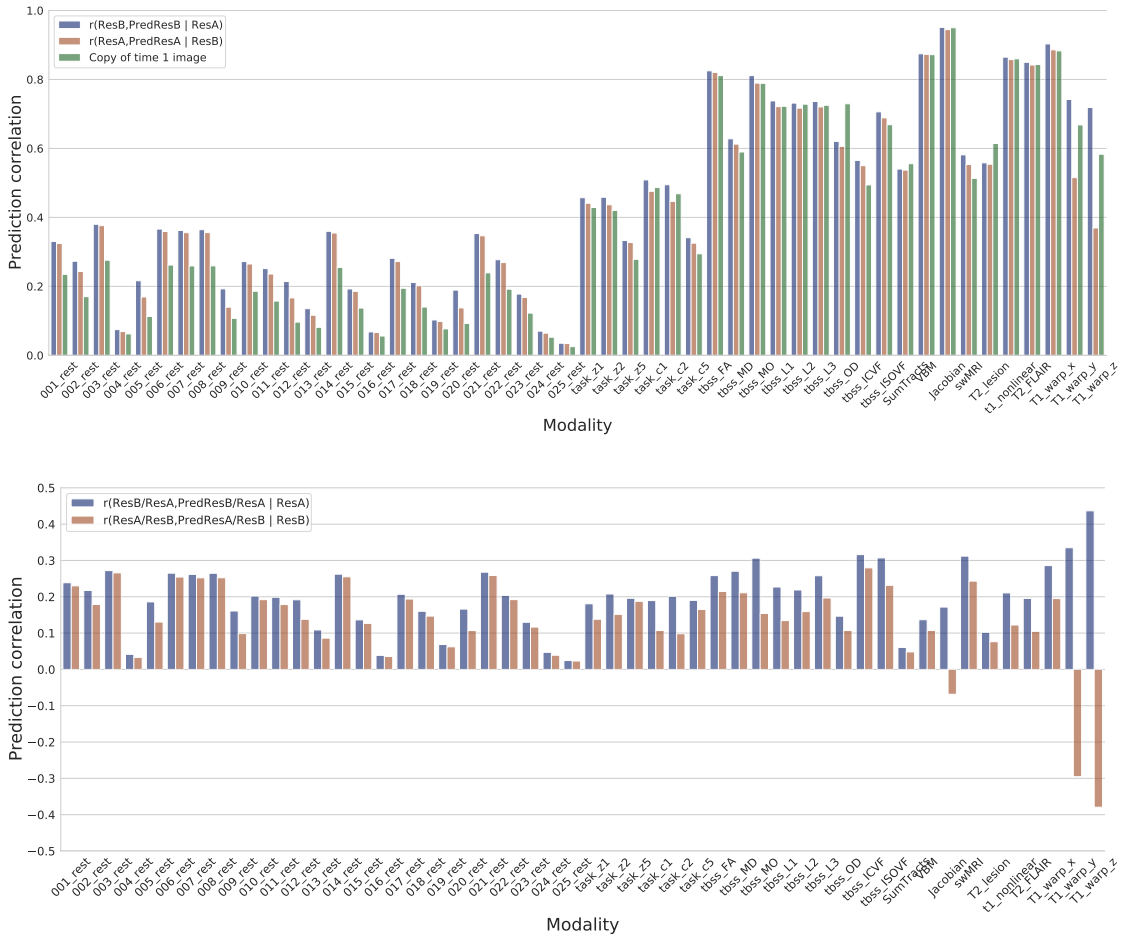



**Figure 4.8:** The results of a test which distinguishes whether a model is carrying out change prediction vs. data denoising. Given a trained model, this compares the correlation of the output with the baseline image fed in and the real follow-up image, i.e. $r$(PredResB, ResB) or r(PredResB\ResA, ResB\ResA), against the correlation of output with the follow-up image fed in and the real baseline image, i.e., $r$(PredResA, ResA) or r(PredResA\ResB, ResA\ResB).

and nonlinear warps. This may because in these modalities the age-related structural changes are more common among subjects.

### 4.3.5 Relations between predicted images and nIDPs

In longitudinal studies, we are often interested in analysing the relationships between longitudinal changes in imaging data and "future" nIDPs (or changes in nIDPs). As our method has shown (in some case) reasonable correlations between predicted and true images, we further investigated whether the changes relate to nIDPs. To this end, we used the predicted residual temporal difference images and the true temporal difference

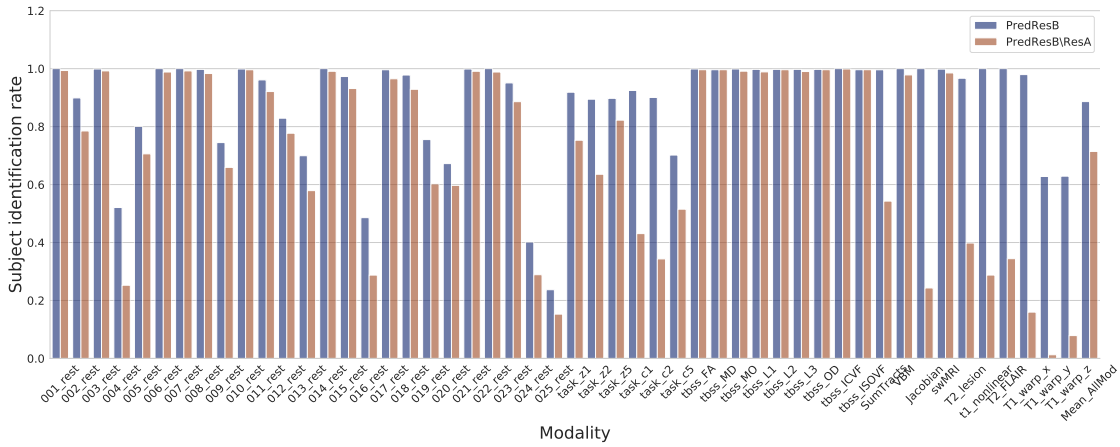



**Figure 4.9:** The subject identification rate of the predicted maps. We correlate each of the predicted maps to all the true maps. This shows the proportion of subjects whose within-subject correlation is the largest among all between-subject correlations. The mean identification rates among all modalities are shown on the right.

image to predict each of the nIDPs available in UK Biobank. **Fig. 4.10** shows the Bland-Altman plots comparing the prediction correlations of nIDPs. We see that in 34 of the 50 modalities, we have significantly improved prediction performance of nIDPs with predicted maps compared with true maps. Importantly, when we explore the relationships between longitudinal changes and nIDPs, neither of the two definition of longitudinal changes, i.e., ResB-ResA or ResB\ResA, can perfectly remove the effect of baseline images. This is because ResB-ResA is explicitly correlated with baseline image, and ResB\ResA can be correlated with the underlying true signals in ResA due to regression dilution problem (ResA contains both signal and noise, so that regressing out ResA does not necessarily means the true underlying signals have been fully regressed out).

Finally, for the 25 resting-state modalities, we explore the relationship of mean improvements of nIDP predictions by the predicted maps versus the true maps with the difference between prediction correlations of follow-up images and their test-retest reproducibility. **Fig. 4.11** shows that there is a clear linear relationship between these two metrics. This result indicates that the more we can predict the follow-up images, compared with data reproducibility, the more we will improve with the predicted maps.

In summary, the above results show that our model can reasonably capture nIDP-related patterns in temporal changes in some case, and may reduce the noise level of predicted data, resulting in similar or even improved nIDP prediction accuracy.

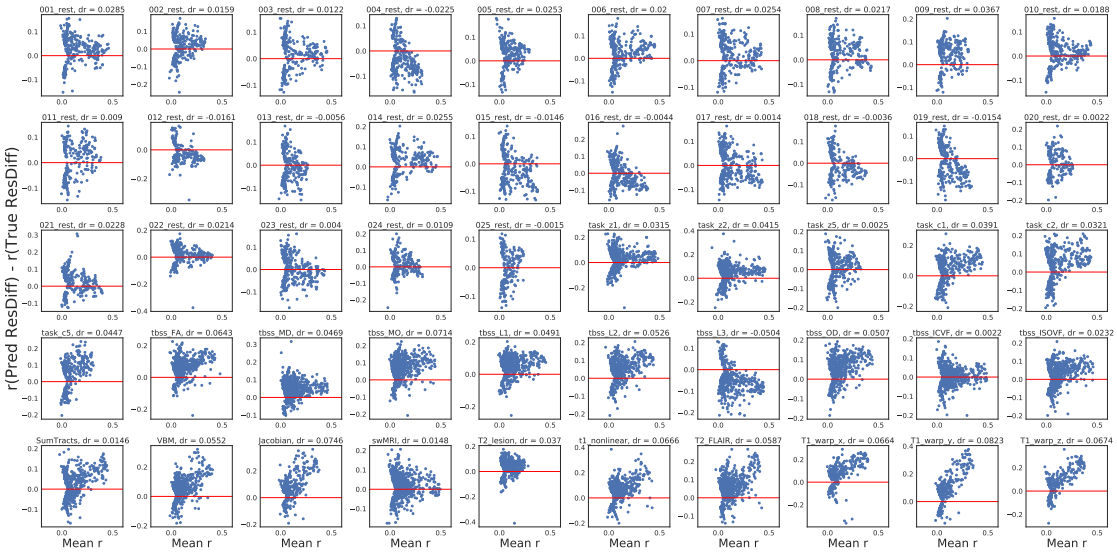



**Figure 4.10:** The Bland-Altman plots of prediction performance of nIDPs using the predicted and true residual temporal difference images. Each point in the figure represents one nIDP, where the x-axis is the mean of its prediction correlation and the y-axis is the difference of the correlation between predicted images and true images.

Therefore, our prediction model can potentially impute imaging data of missing subjects in longitudinal studies and perform early prediction of brain imaging biomarkers.

### 4.3.6 Real data qualitative analysis

**Fig. 4.12** shows some examples of subjects' predicted residual difference images together with their true residual difference images. Each of them is thresholded only to include the top 5% of voxels with the highest values. We can see that there exist a lot of individual differences, but our model can capture many of them, such as the differences in the increased ventricle sizes shown in the T1 jacobian maps and different patterns of default mode network changes.

## 4.4 Discussion

In this chapter, we propose a general framework for individualised prediction of longitudinal changes in multimodal brain images. The proposed method takes one or multiple baseline brain imaging maps as inputs and outputs the predicted follow-up images and the temporal difference images. We show that we are able to accurately predict the longitudinal changes to some degree in almost all of the 50 modalities in the UK Biobank,

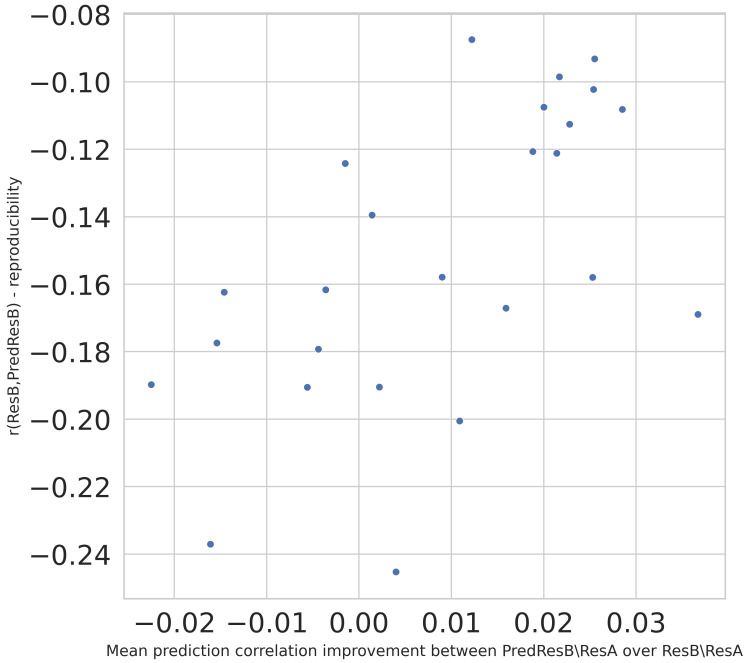



**Figure 4.11:** For 25 resting-state modalities, the improvements of prediction correlation of nIDPs between predicted and true residual temporal difference images positively correlated with the difference of follow-up image prediction correlation and test-retest reproducibility ($r = 0.59, p = 0.002$).

with two-year follow-up time intervals. For example, this includes the brain shape changes reflected in the warp fields, and the functional changes reflected in the resting-state network and task activation patterns. The method is generic and easy to train and applied to other modalities and datasets. We also proposed a mathematical framework for evaluating longitudinal predictions. The code for our model will be released on **Github**.

Compared with previous works [Xia et al., 2021, Da Silva et al., 2021, Gafuroğlu et al., 2018, Rachmadi et al., 2020, Hong et al., 2019, Nebli et al., 2020, Ghribi et al., 2021], our work here has the following advantages: First, our method is more generic, the same U-Net structure can be applied to train for 50 different modalities in one of the largest longitudinal neuroimaging datasets UK Biobank. The sample size and number of modalities used here are significantly larger than in previous studies. Second, for the first time, we proposed a mathematical modelling framework for longitudinal brain imaging studies. Based on this framework, we can (1) derive empirical baseline correlation of temporal change prediction; (2) prove the existence of temporal changes in the imaging data; (3) test whether a model does change prediction or denoises the data. These analyses are particularly important when the follow-up time interval is small (e.g., on average 2 years

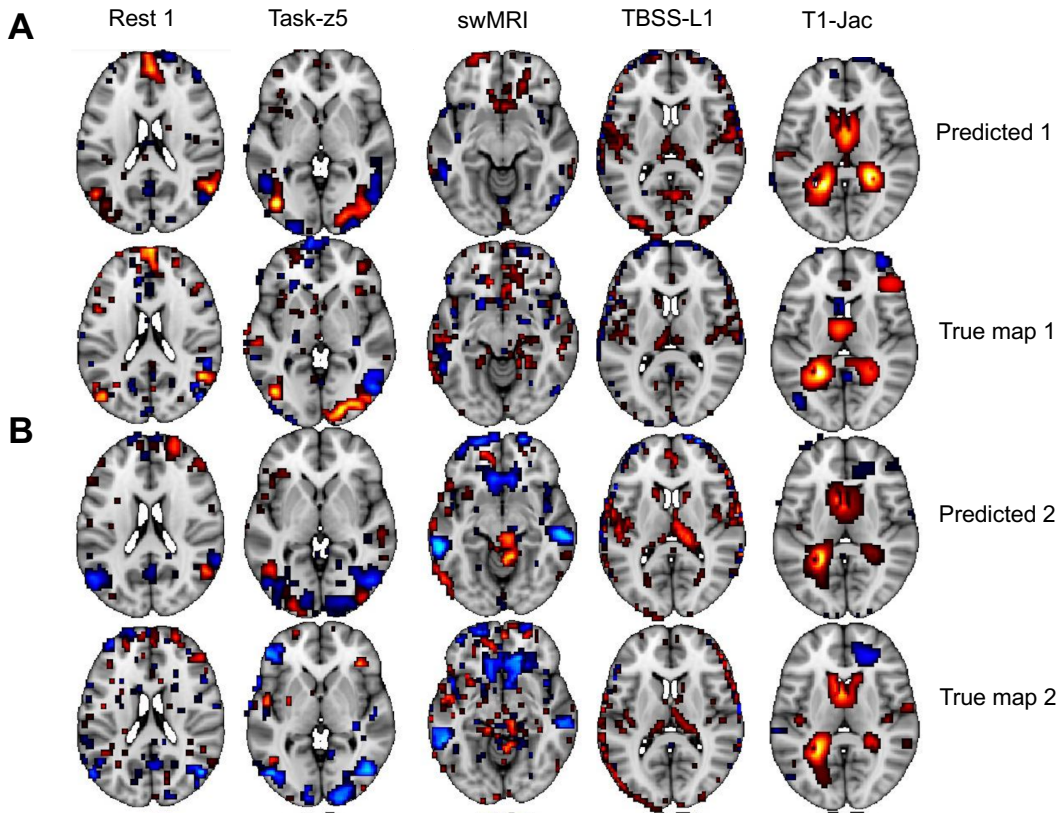



**Figure 4.12:** Real data examples of predicting residual temporal difference maps for 5 modalities (columns). A,B are two examples (but may not be the same subjects across modalities).

in the UK Biobank data). Previous studies usually only use evaluation metrics such as correlation or MSE to evaluate their proposed approaches. Third, the proposed method is further validated by comparing the predicted maps with the true maps in terms of predicting thousands of nIDPs in the UK Biobank. Previous work only used a few nIDPs such as age or status of Alzheimer's Disease. Finally, we also showed that the predicted follow-up maps and longitudinal changes could be fingerprints of a subject.

Our work points out various caveats of longitudinal prediction. We need a good definition of longitudinal changes. We proposed the *temporal difference* and the *residual temporal difference* as two definitions. However, both have their pros and cons. The temporal change is the most straightforward one, but we need to carefully derive a "null" prediction correlation because it intrinsically contains a baseline image term. The residual temporal difference partially overcomes this problem, but may overly regress out temporal changes that are linearly related to the baseline images, and it may also surfer from the regression dilution problem so that the (true underlying) signals in the

baseline image are still there. However, due to the high noise in both baseline and follow-up neuroimaging data, a "perfect" definition that can fully remove baseline image information from longitudinal changes is hard to find. Therefore, when interpreting longitudinal results, we need carefully think about if the findings are truly related to the changes or just confounding factors.

We still can improve the current work in the following ways. First, the prediction performance of follow-up images or temporal difference images may be improved if we directly incorporate nIDP information in the training process. Second, we may use more advanced deep learning models to integrate multimodal brain imaging data, such as transformer [Dosovitskiy et al., 2020]. Third, the ability to transfer longitudinal prediction models to different domains, i.e., sites, is still unexplored. We may train a more generic cross-site model by multi-centre data and advanced domain adaptation techniques [Ganin et al., 2016].

## Acknowledgments.

We are grateful to UK Biobank and its participants (access application 8107). 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.

## 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`).

# 5 Summary and future works

## Contents


**5.1 Work summary . . . . . 102**
**5.2 Future work . . . . . 105**
5.2.1 Multimodal fusion for data-driven IDP discovery. . . . . 105
5.2.2 Longitudinal brain image prediction. . . . . 106


## 5.1 Work summary

The work in this thesis arises from our group's role leading the brain imaging in UK Biobank, currently halfway through MRI scanning of 100,000 participants, creating the largest brain imaging resource to date. In 2016, our group announced the resource [Miller et al., 2016], and in 2018 our group published its first large-scale imaging-genetics study, with more than 3,000 GWAS of brain structure and function [Elliott et al., 2018]. These papers used our hand-created Imaging-derived Phenotypes (IDP), and until the end of 2021 around 100 papers have been published using these IDPs. However, there is huge scope for expanding and improving these IDPs, aiming to better predict future disease and brain development and to find genetic associations of brain structure and function, in a data-driven way rather than being hand-created.

In this thesis, we have developed new methodologies for modelling multimodal population brain imaging data. The main focus of Chapters 2 and 3 are new methods for fusing multimodal data to generate low dimensional summary features, i.e., data-driven IDPs. The main focus of Chapter 4 is a new approach for the longitudinal prediction of brain imaging data, i.e., using baseline brain images to predict future brain development. A detailed summary and discussion have been written in the corresponding chapters, so we only briefly summarize them here.

In Chapter 2, we developed BigFLICA, a new multimodal ICA approach for unsupervised fusion of brain imaging data. The main contributions are:

- We propose a two-stage dimension-reduction approach to effectively preprocess a multimodal dataset and reduce the computational load of the original FLICA algorithm.
- The approach is scalable to extreme high-dimensional data sets.
- We provide effective ways of choosing the hyper-parameters of BigFLICA.
- We demonstrated that a high-dimensional BigFLICA has improved predictive power compared with hand-curated IDPs and ICA.
- We demonstrated the value of multimodal fusion instead of analyzing each modality separately.
- We studied between-modalities relationships based on BigFLICA decomposition.
- We studied BigFLICA's reproducibility and compared it with other approaches.
- We demonstrated BigFLICA can find interesting modes related to nIDPs such as handedness, fluid intelligence, hypertension and age started wearing glasses.

In Chapter 3, we developed SuperBigFLICA, a supervised multimodal fusion approach for analyzing brain imaging data. The main contributions are:

- SuperBigFLICA inherits the Bayesian framework from the previous FLICA model. Additionally, it incorporates an additional prediction term to enable supervised learning of the target variable of interests (i.e., multiple nIDPs).
- The approach is scalable to extreme high-dimensional data sets, e.g., UK Biobank scale neuroimaging datasets, through the use of advanced stochastic optimisation algorithms.
- SuperBigFLICA largely bypasses the need for parameter tuning.
- SuperBigFLICA outperforms hand-curated IDPs and BigFLICA as new image-derived phenotypes.
- SuperBigFLICA can learn a generalisable latent space by applying it to predict unseen nIDPs.
- SuperBigFLICA is robust in different prediction tasks.
- SuperBigFLICA can find interpretable modes predictive of health outcome and cognitive nIDPs.

In Chapter 4, we developed a new deep learning model for longitudinal brain image prediction. The main contributions are:

- We developed a general framework that can be applied to 50 different modalities.
- We can predict, to a modest degree, individualised longitudinal changes in almost all the 50 modalities in the UK Biobank, with two-year follow-up time intervals.
- Our approach outperforms other linear and deep learning models in this task.
- We proposed a mathematical framework for evaluating longitudinal prediction.
- We raised and investigated several caveats for longitudinal data analysis.

## 5.2 Future work

### 5.2.1 Multimodal fusion for data-driven IDP discovery.

Multimodal fusion is a hot topic in the neuroimaging field, in terms of both methodology development and application. There exists a lot of literature that proposed different ways of fusing multimodal data. However, there are no "fair" comparisons across different methods in benchmark datasets, for various reasons. For example, there is a data-sharing restriction; there is no publicly available code for some approaches; different methods are evaluated on different datasets, etc. A potential future direction is to compare different multimodal fusion approaches comprehensively, using a dataset like UK Biobank with many nIDPs and imaging modalities available.

In addition, the adaptation of multimodal algorithms to multi-site data is also interesting but challenging. Currently, most multimodal approaches assume that the data are independent and identically distributed random variables. However, in multi-site data, this assumption no longer holds. One potential way is to add an adversarial loss to the original multimodal fusion algorithm. This adversarial loss is high if the learned latent component can classify the site correctly.

Finally, deep learning-based multimodal fusion has rarely been explored for neuroimaging data, which could have huge potential for the discovery of powerful nonlinear IDPs. We can use convolutional neural networks that directly extract features from the voxel-wise imaging data, either as a preprocessing step or trained end-to-end before the nIDP predictor. This convolutional neural network can have a similar structure to the variational auto-encoder.

On the clinical application side, applying BigFLICA and SuperBigFLICA to different real clinical and neuroscience problems is promising. For example, we can explore common brain features related to multiple psychiatric diseases and their polygenic risk scores (PRS), which can potentially be more powerful than using existing IDPs.

### 5.2.2 Longitudinal brain image prediction.

The acquisition of longitudinal brain imaging data is difficult, but the longitudinal studies of the brain are important for the early design of personalised treatments and understanding of ageing. When there are few longitudinal scans available, we don't have enough statistical power to perform longitudinal analysis. Our preliminary longitudinal prediction study has explored an interesting question on whether the baseline image can predict the follow-up images, which has the potential for longitudinal data imputation. When there are more follow-up images and nIDPs available in the UK Biobank, we could have these future analyses that extend this study by exploring (1) whether the baseline image can predict the follow-up nIDPs; (2) whether the baseline image can predict the change of nIDPs over time; (3) which brain regions are important for make prediction of follow-up nIDPs and images; (4) How the changes in brain images relate to changes in nIDPs. For the first two points, we can design a convolutional neural network that extracts features from the baseline image and predicts the follow-up nIDPs, and for the last two points, we need to explore how the trained model makes the prediction by methods such as saliency maps and class activation map.

# Appendices

# A

# A.1 Supplementary materials for Chapter 1

## A.1.1 Optimization algorithm for dictionary learning

Although the objective function is not convex wrt. both $A$ and $D$, it is convex when either of them is fixed, so we use an alternating minimization approach to obtain a local optimal estimation. First, at $t-$th step, given an estimates of $A_{t-1}$, the objective for finding $D_i(t)$ is a *LASSO* problem:

$$\min_{D_i(t),i=1,\ldots,N} \sum_{i=1}^{N} \frac{1}{2}\|X_i - A_{t-1}D_i(t)\|_2^2 + \lambda\|D_i(t)\|_1 \tag{A.1}$$

which can be solved by the LARS algorithm [Efron et al., 2004]. We then fix $D$, to obtain an estimate of $A$. The objective function now becomes:

$$\begin{aligned} &\min_{A} \sum_{i=1}^{N} \frac{1}{2}\|X_i - A_{t-1}D_i(t)\|_2^2 + \lambda\|D_i(t)\|_1 \\ &= \frac{1}{N}\left(\frac{1}{2}tr(A^\tau AC_t) - tr(A^\tau B_t)\right) \end{aligned} \tag{A.2}$$

with $B_t = B_{t-1} + x_i D_i(t)^\tau$, $C_t = C_{t-1} + D_i(t)D_i(t)^\tau$ and $A_j^\tau A_j \leqslant 1$.

Taking the derivative of this objective function, we have the following update rule (based on projected gradient descent):

$$\begin{aligned} f(A) &= \frac{1}{2}tr(A^\tau AC) - tr(A^\tau B) \\ &= \frac{1}{2}\sum_{ij} C_{ij}a_i^\tau a_j - \sum_j a_j^\tau b_j \\ &= \frac{1}{2}(\sum_{i\neq j} C_{ij}a_i^\tau a_j + \sum_j C_{jj}a_j^\tau a_j) - \sum_j a_j^\tau b_j \\ \frac{\partial f(D)}{\partial a_j} &= \frac{1}{2}(2\sum_{i\neq j} C_{ij}a_i + 2\sum_j C_{jj}a_j) - b_j \\ &= Ac_j - b_j \\ u_j &= a_j - \frac{1}{C[j,j]}(Ac_j - b_j) \\ a_j &= \frac{1}{\max(||u_j||_2, 1)}u_j \end{aligned} \tag{A.3}$$

where $b_j$ and $c_j$ are $j$-th columns of $B$ and $C$.

### A.1.2 Optimization algorithm for Elastic-net regression

There are many methods for optimizing the Elastic-net loss function; the most efficient implementation in the glmnet package uses the coordinate descent approach [Friedman et al., 2010]. That is, the parameters are updated one by one iteratively. Taking the derivative of the $L$ with respect to one element of $w$, i.e., $w_j$, when $w_j > 0$, we get:

$$\frac{\partial L}{\partial w_j} = 2\sum x_j(Y - Xw) + \lambda(1-\alpha)w_j + \lambda\alpha \tag{A.4}$$

where $x_j$ is the $j$-th column of X. A similar expression exists if $w_j < 0$, and $w_j = 0$ is treated separately. Then the update rule is:

$$w_j = \frac{S\left(\sum x_j(Y - \sum_{i \neq j} w_i x_i), \lambda\alpha\right)}{1 + \lambda(1-\alpha)} \tag{A.5}$$

where $S$ is the soft-thresholding operator. For more details, please refer to [Friedman et al., 2010].

## A.2 Supplementary materials for Chapter 2

### A.2.1 Confounding variables regressed out in Chapter 2

UKB dataset: age, age squared, age X sex, age squared X sex, age (quantile normalised), age squared (quantile normalised), age X sex (quantile normalised), age squared X sex (quantile normalised), rfMRI head motion, tfMRI head motion, head size scaling, rfMRI head motion squared, tfMRI head motion squared, [4] confounds relating to bed position in scanner (x), [4] confounds relating to bed position in scanner (y), [4] confounds relating to bed position in scanner (z), [4] confounds relating to bed position in scanner (table), [4] confounds relating to bed position in scanner (x) squared, [4] confounds relating to bed position in scanner (y) squared, [4] confounds relating to bed position in scanner (z) squared, [4] confounds relating to bed position in scanner (table) squared, [10] confounds modelling slow date-related drift 1, [10] confounds modelling slow date-related drift 2, [10] confounds modelling slow date-related drift 3, [10] confounds modelling slow date-related drift 4, [10] confounds modelling slow date-related drift 5, [10] confounds modelling slow date-related drift 6, [10] confounds modelling slow date-related drift 7, [10] confounds modelling slow date-related drift 8, [10] confounds modelling slow date-related drift 9, [10] confounds modelling slow date-related drift 10, rfMRI head motion (quantile normalised), tfMRI head motion (quantile normalised), head size scaling (quantile normalised), [4] confounds relating to bed position in scanner (x) (quantile normalised), [4] confounds relating to bed position in scanner (y) (quantile normalised), [4] confounds relating to bed position in scanner (z) (quantile normalised), [4] confounds relating to bed position in scanner (table) (quantile normalised), [4] confounds relating to bed position in scanner (x) squared (quantile normalised), [4] confounds relating to bed position in scanner (y) squared (quantile normalised), [4] confounds relating to bed position in scanner (z) squared (quantile normalised), [4] confounds relating to bed position in scanner (table) squared (quantile normalised), [10] confounds modelling slow date-related drift 1 (quantile normalised), [10] confounds modelling slow date-related drift 2 (quantile normalised), [10] confounds modelling slow date-related drift 3 (quantile normalised), [10] confounds modelling slow date-related drift 4 (quantile

normalised), [10] confounds modelling slow date-related drift 5 (quantile normalised), [10] confounds modelling slow date-related drift 6 (quantile normalised), [10] confounds modelling slow date-related drift 7 (quantile normalised), [10] confounds modelling slow date-related drift 8 (quantile normalised), [10] confounds modelling slow date-related drift 9 (quantile normalised), [10] confounds modelling slow date-related drift 10 (quantile normalised), imaging centre, sex.

HCP dataset: image reconstruction version, age, age squared, sex, age X sex, age squared X sex, race, ethnicity, rfMRI motion, Height, Weight, FS_IntraCranial_Vol, FS_BrainSeg_Vol.

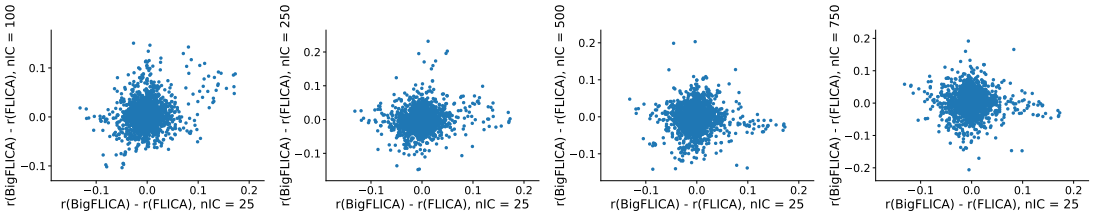



**Figure A.1:** The relationship between the difference of prediction accuracy of BigFLICA and FLICA across different number of components.

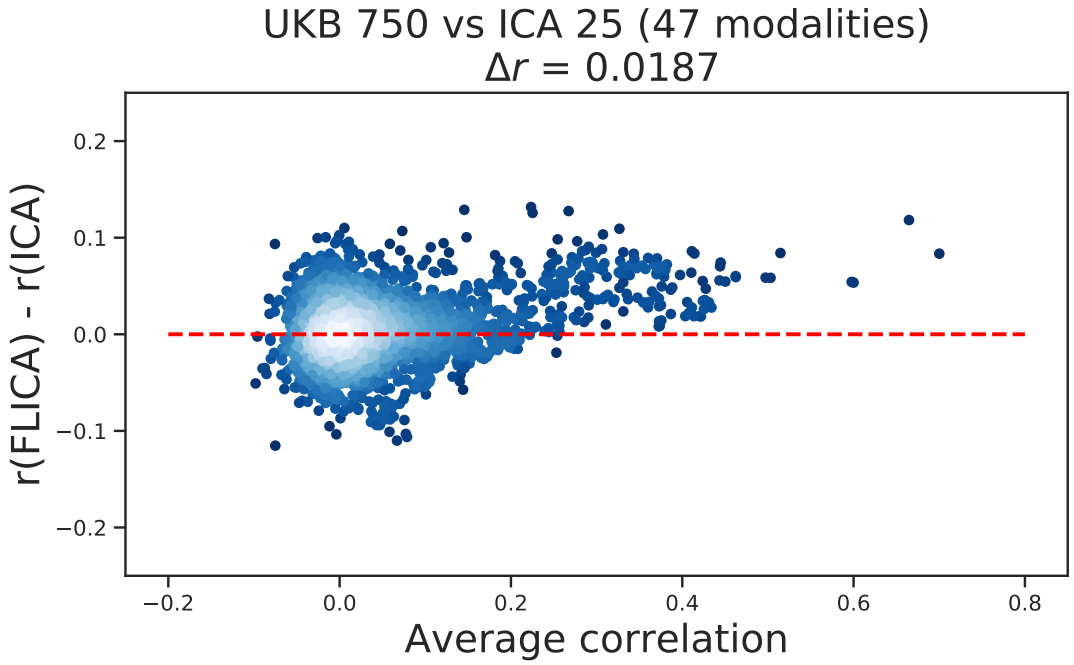



**Figure A.2:** Comparing the prediction performance of the 750-dimensional FLICA with the 25-dimensional single-modality ICA concatenated across 47 modalities in the UKB data.

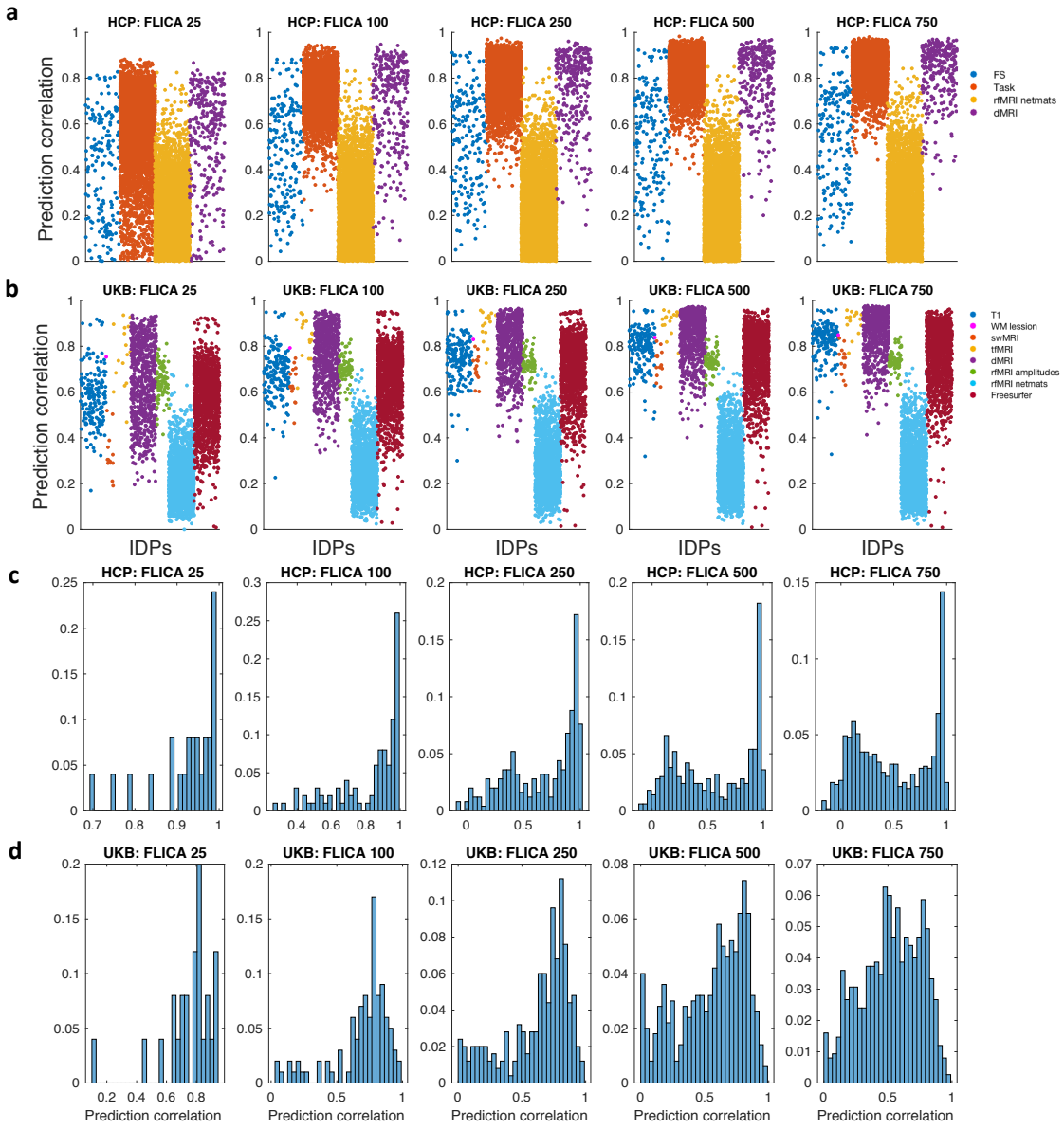



**Figure A.3: Relationships between FLICA and IDPs.** **(a,b)** The plots show the results of predicting each IDP using BigFLICA modes in **a** the HCP and **b** the UKB dataset. The IDPs are appearing in order along the x axis, and are grouped and coloured by modality types. **(a,b)** The histograms of predicting BigFLICA modes using all IDPs in **c** the HCP and **d** the UKB dataset.

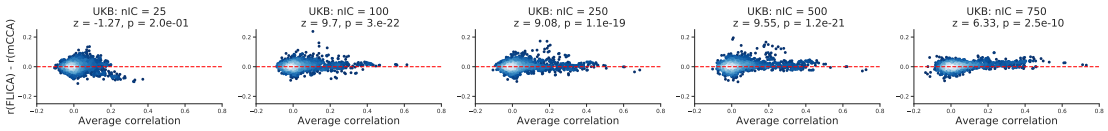



**Figure A.4: Comparing FLICA and mCCA in the UKB data.** Comparing the predictive performance of FLICA with mCCA (or equivalently the subject-by-component matrix obtained in the mMIGP step) across different numbers of extracted components in the UKB dataset. The FLICA and mCCA dimensions are the same in each figure.

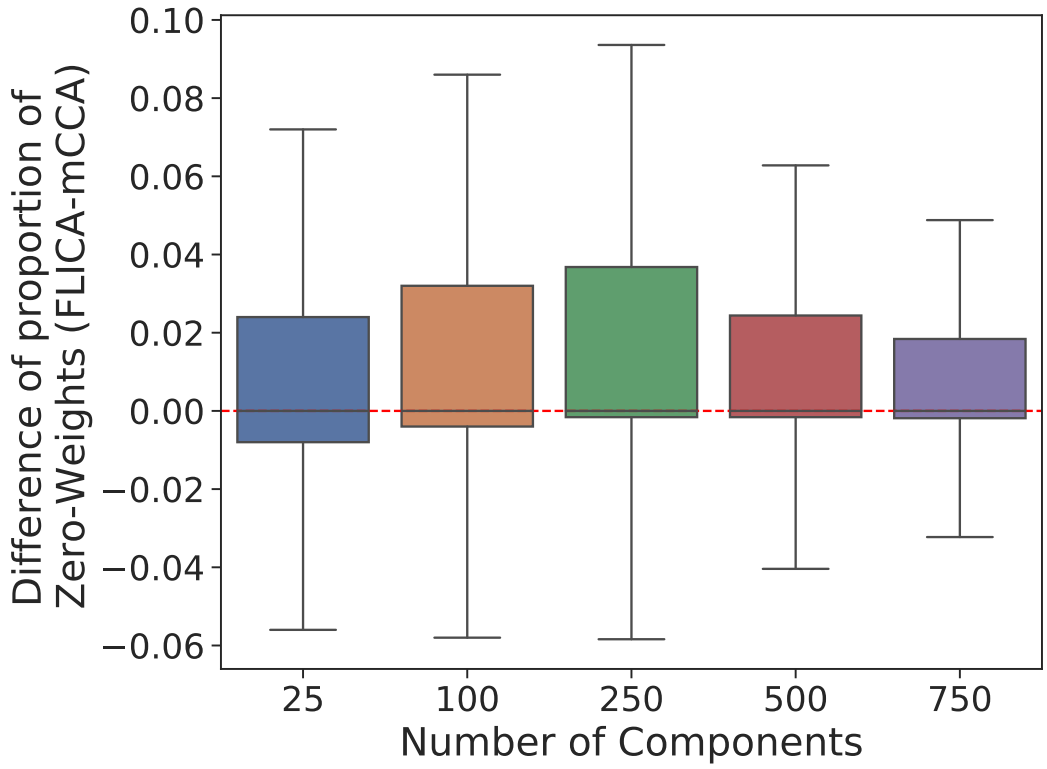



**Figure A.5:** The difference of the proportion of zeros weights (BigFLICA-mCCA) in predicting nIDPs across 5 dimensions of decomposition in the UKB data.

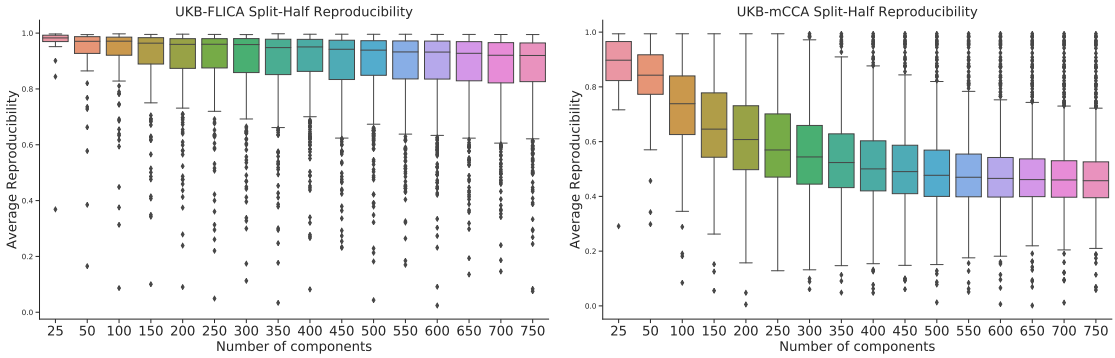



**Figure A.6: Split-half reproducibility of BigFLICA and mCCA spatial maps in the UKB dataset.** The split-half reproducibility of BigFLICA and mCCA in the UKB dataset by first computing the correlation between modality-wise concatenated spatial maps after eliminating low-weight voxels and then greedy matching.

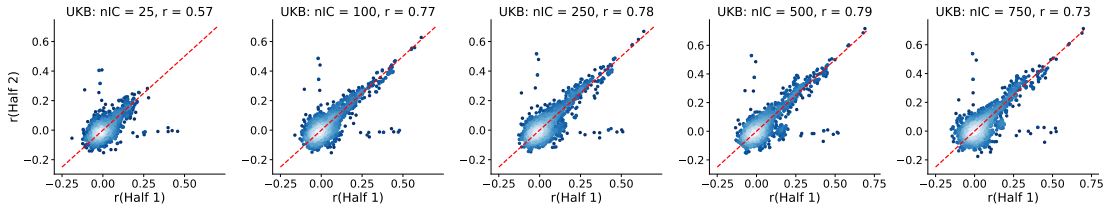



**Figure A.7:** Split-half prediction accuracy of all nIDPs of BigFLICA in the UKB dataset.

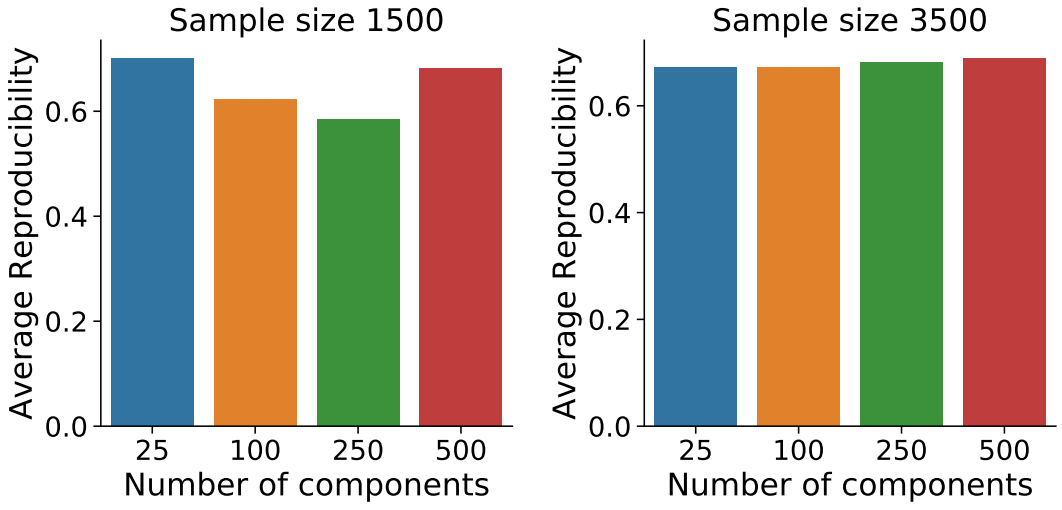



**Figure A.8:** Split-half reproducibility of BigFLICA using different sample size in the UKB dataset. Left: sample size = 1500. Right: sample size = 3500.

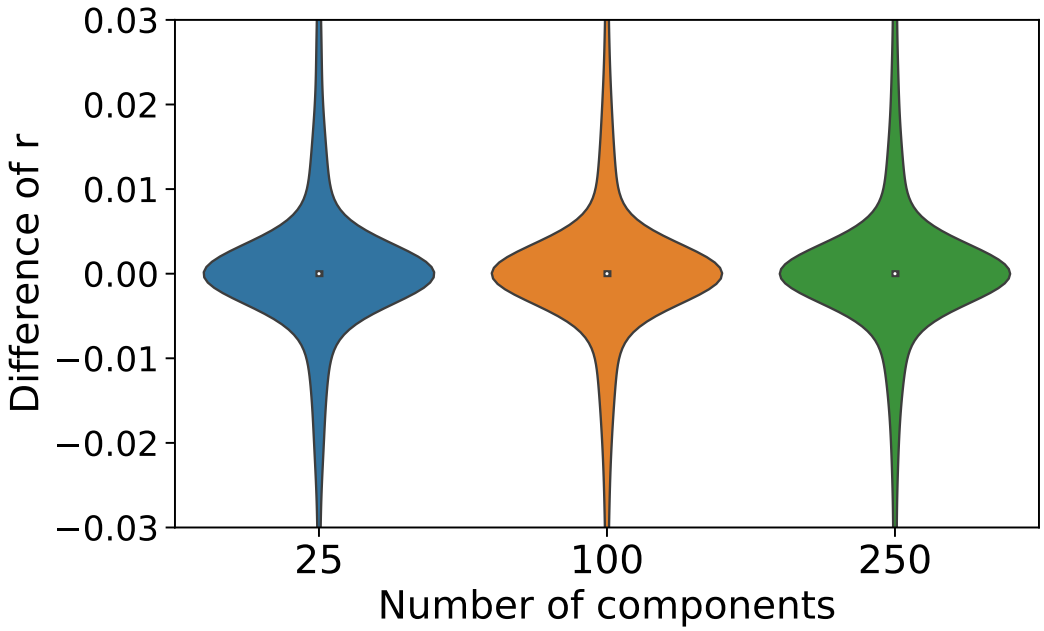



**Figure A.9: Stability of prediction accuracy of BigFLICA against different train-test subject splits.** We estimated the prediction correlation of each nIDP using 5 different train-test subject splits, and calculated the mean accuracy of prediction correlations for each nIDPs. We then computed the difference of the mean prediction correlation and the prediction correlation of one of the five random predictions across different number of components.

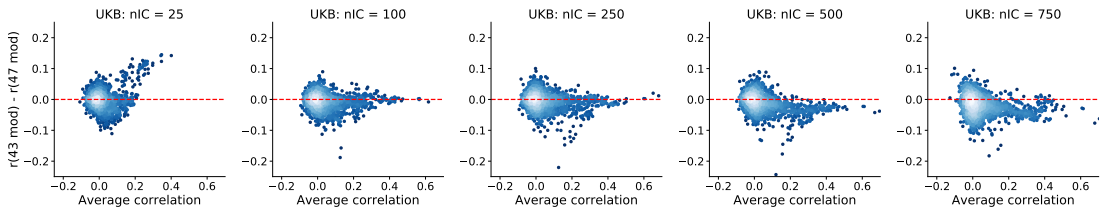



**Figure A.10: Comparison of prediction accuracy of nIDPs when removing 4 noise resting-state modalities with including them in the model.** An increased prediction accuracy was observed for some nIDPs in the 25-dimensional decomposition when compared with using all 47 modalities, and a similar prediction accuracy was observed for 100- and 250-dimensional decomposition, while a decreased prediction accuracy was observed for 500- and 750-dimensional decompositions.

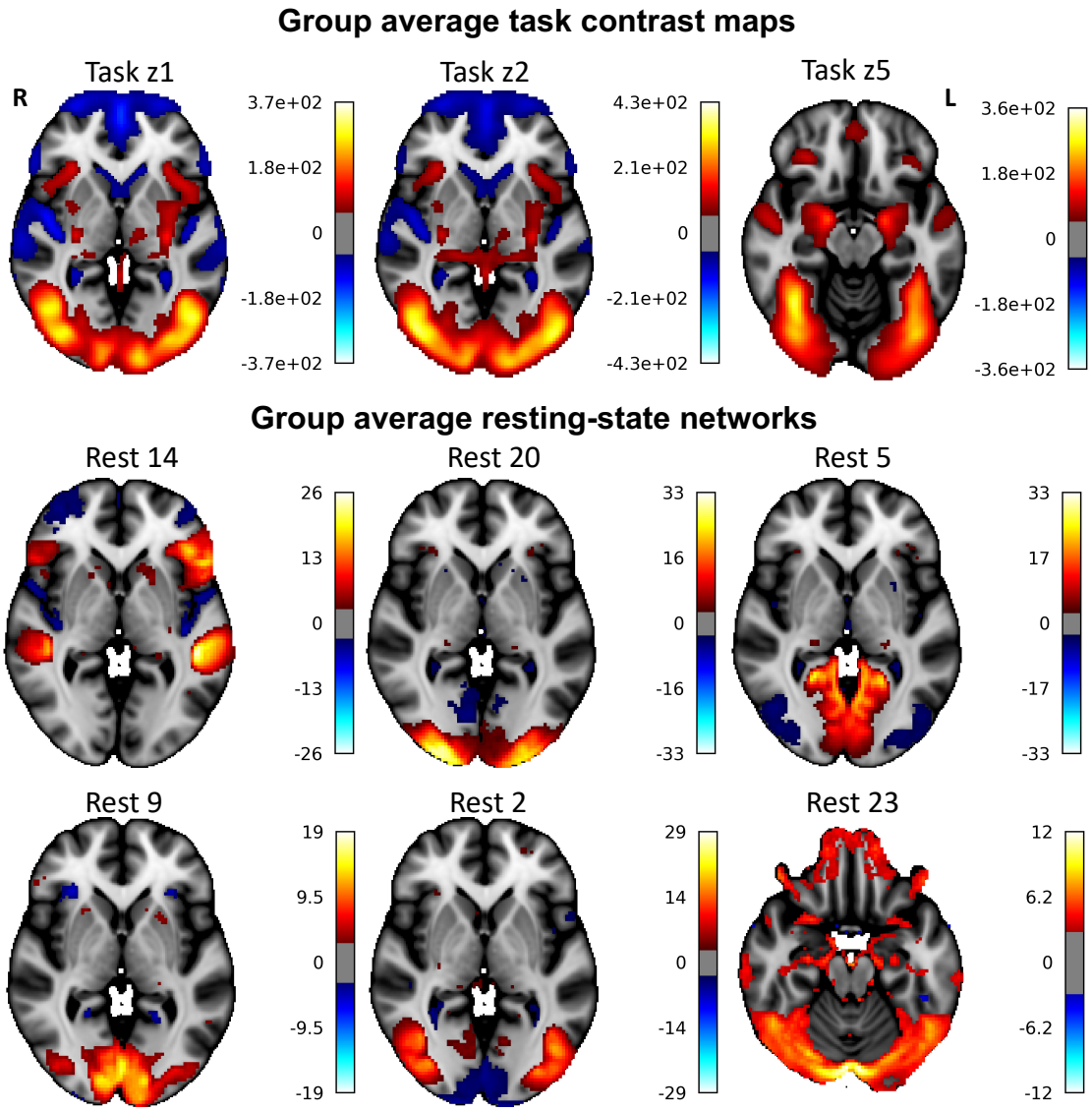



**Figure A.11: Group average maps of task activations and resting-state networks in the UKB dataset.** These are provided to help interpret the population variability maps (modulations of these maps) shown in **Fig. 2.5**. Top: group average of emotion task activation z-statistic maps (task z1: "shapes", task z2: "face", task z5: "faces>shapes"). Group average task contrast (effect size) maps c1, c2 and c5 are highly similar to z-stat maps so that they are not shown. Bottom: group average resting-state networks from a 25-dimensional ICA parcellation in the UKB data. The six maps shown here are the networks from **Fig. 2.5**.

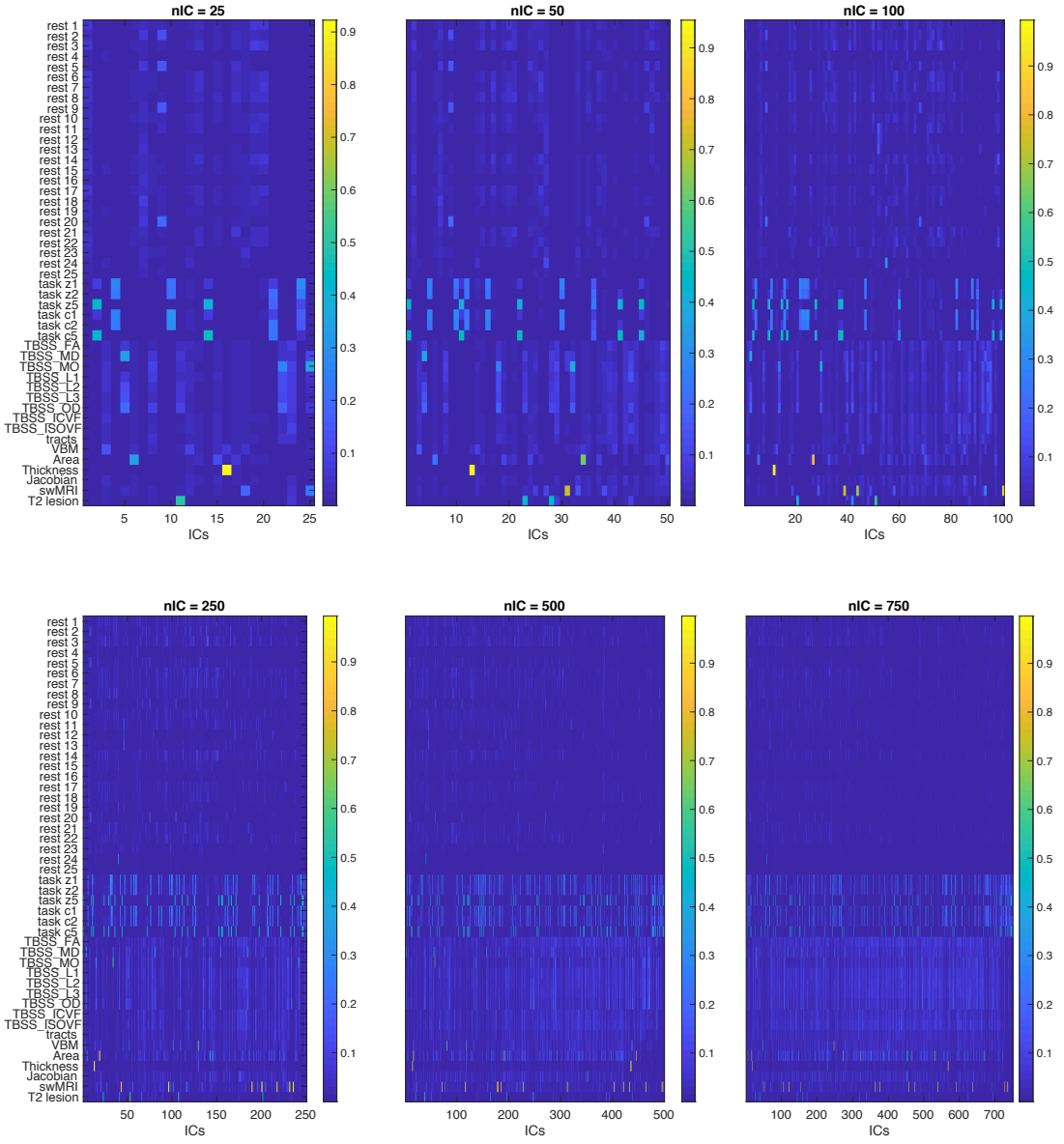



**Figure A.12: The contribution of each modality in each BigFLICA mode (independent component).**

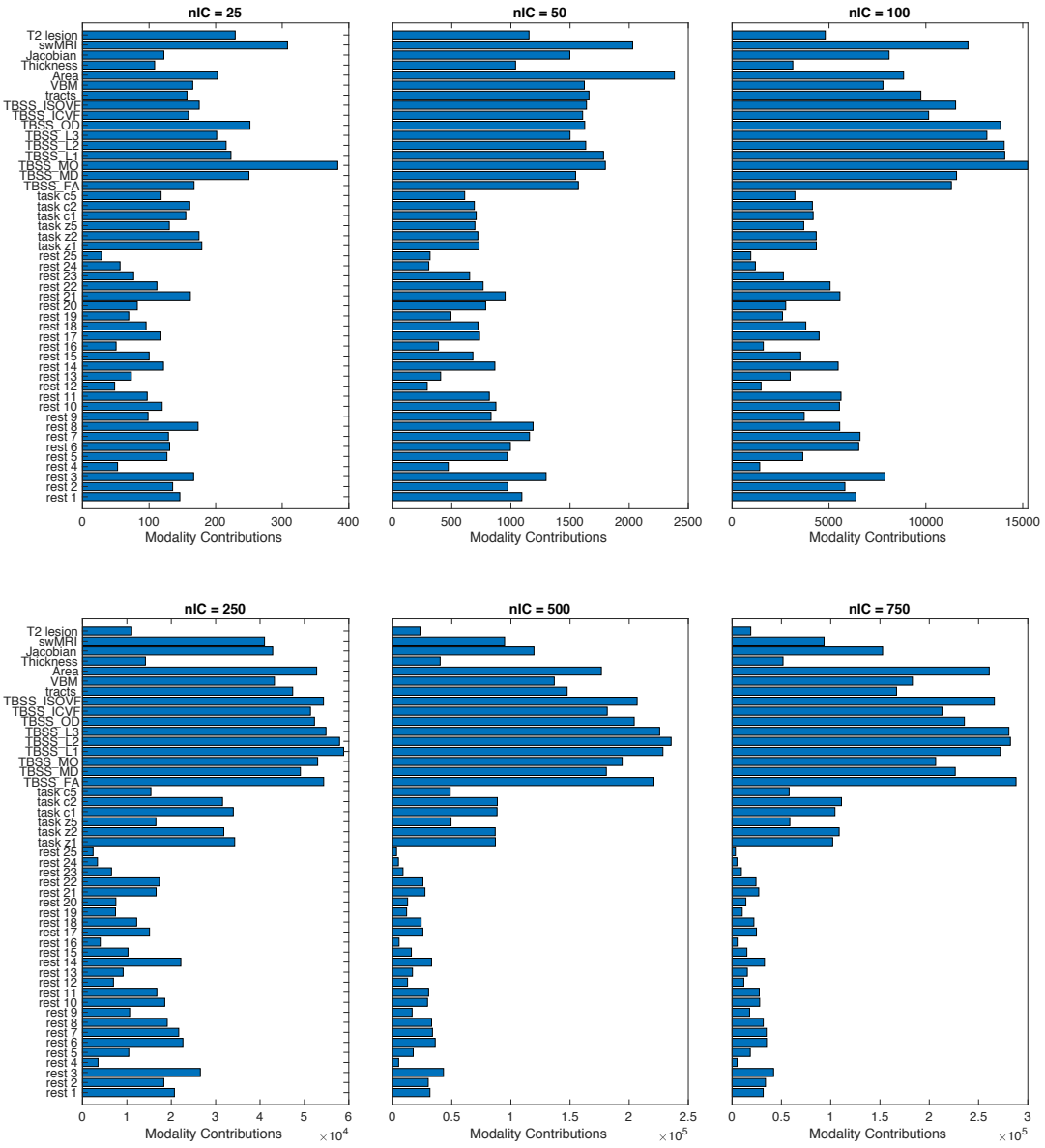



**Figure A.13: The relative contribution of different modalities of a BigFLICA decomposition (nIC=25-750) in the UKB data.** For each modality, we take the sum of its overall contribution (estimated by BigFLICA) across all components.

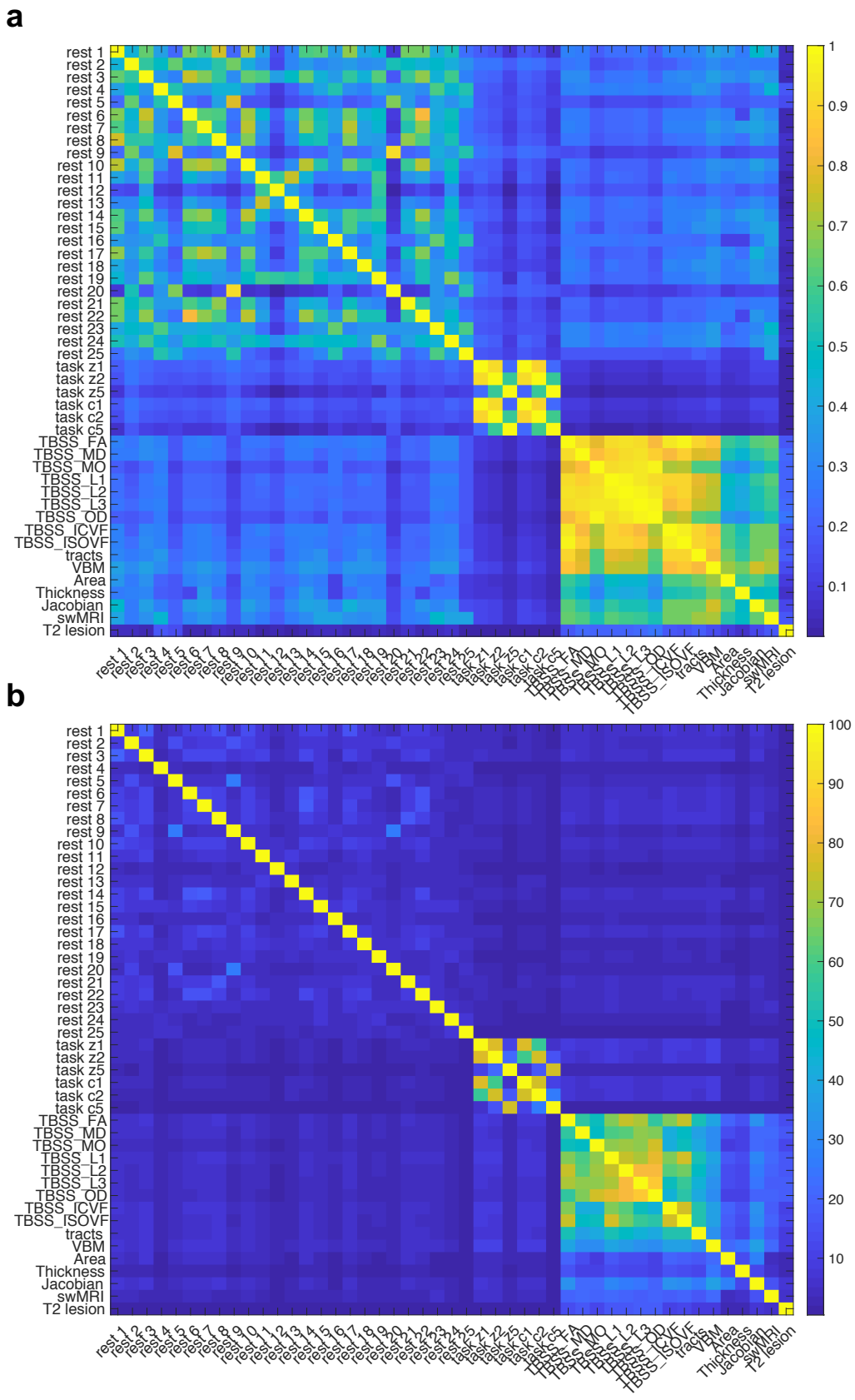



**Figure A.14: The relationships between different modalities in the UKB data.** a). The cosine similarity of modality contributions across 750 components (estimated by BigFLICA) between every pair of modalities. b). The amount of shared variance between two 50-dimensional single-modality ICA decompositions in each pair of modalities.

**Table A.1:** Comparison of prediction performance of **Cognitive Phenotypes** and **Health and Medical History Health Outcomes** between BigFLICA (nIC=750) and 3,913 IDPs in the UKB dataset. We excluded an nIDP if both methods have prediction r-value$<$ 0.1.

| Variable Names | r-value FLICA | -log10(p) FLICA | r-value IDPs | -log10(p) IDPs | Percent Improvement | Sample size |
|---|---|---|---|---|---|---|
| **cognitive phenotypes** | | | | | | |
| Digits entered correctly (0.1) | 0.104 | 3.6 | 0.003 | 0.3 | 3141.6 | 1126 |
| Time to complete round (1.2) | 0.108 | 11.7 | 0.073 | 5.9 | 47.4 | 4119 |
| Number of fluid intelligence questions attempted within time limit (1.0) | 0.107 | 11.6 | 0.078 | 6.6 | 37.6 | 4116 |
| Duration to first press of snap-button in each round (2.7) | 0.11 | 37.6 | 0.081 | 21.1 | 35.1 | 13749 |
| Mean time to correctly identify matches (0.0) | 0.108 | 38.6 | 0.083 | 23.1 | 30.6 | 14471 |
| Duration screen displayed (2.0) | 0.103 | 33.6 | 0.079 | 20.3 | 30.2 | 13809 |
| Duration to first press of snap-button in each round (2.10) | 0.104 | 33.7 | 0.08 | 20.3 | 30.1 | 13751 |
| Duration to complete alphanumeric path (trail #2) (0.0) | 0.145 | 33.7 | 0.113 | 20.8 | 28.7 | 6966 |
| Time to complete round (0.2) | 0.113 | 41 | 0.088 | 25.5 | 27.8 | 14115 |
| Number of symbol digit matches made correctly (0.0) | 0.156 | 43.5 | 0.122 | 27 | 27.7 | 7862 |
| Number of symbol digit matches attempted (0.0) | 0.167 | 49.8 | 0.132 | 31.4 | 26.6 | 7862 |
| Number of fluid intelligence questions attempted within time limit (2.0) | 0.151 | 68.4 | 0.122 | 44.9 | 23.8 | 13362 |
| Duration to first press of snap-button in each round (2.11) | 0.1 | 31.6 | 0.082 | 21.3 | 22.9 | 13741 |
| Time to complete round (0.1) | 0.143 | 36 | 0.117 | 24.4 | 22.5 | 7670 |
| Time to complete round (2.2) | 0.114 | 39.4 | 0.093 | 26.8 | 22.1 | 13394 |
| Mean time to correctly identify matches (2.0) | 0.141 | 61.6 | 0.116 | 42 | 21.6 | 13768 |
| Maximum digits remembered correctly (0.0) | 0.15 | 38.2 | 0.129 | 28.7 | 15.9 | 7465 |
| Fluid intelligence score (2.0) | 0.256 | 198.7 | 0.232 | 162.6 | 10.3 | 13362 |
| Fluid intelligence score (0.0) | 0.199 | 47.8 | 0.183 | 40.5 | 8.8 | 5266 |
| Touchscreen duration (2.0) | 0.127 | 52.5 | 0.121 | 48 | 4.7 | 14412 |
| Fluid intelligence score (1.0) | 0.18 | 31 | 0.188 | 33.5 | -3.9 | 4116 |
| Number of fluid intelligence questions attempted within time limit (0.0) | 0.098 | 12.4 | 0.103 | 13.4 | -4 | 5266 |
| Fluid intelligence score (0.0) | 0.198 | 72.1 | 0.209 | 80.5 | -5.3 | 8090 |
| Touchscreen duration (1.0) | 0.103 | 10.8 | 0.127 | 16 | -19.2 | 4134 |
| **Health and Medical History Health Outcomes** | | | | | | |
| Age asthma diagnosed (0.0) | 0.122 | 5.6 | 0.051 | 1.5 | 139.4 | 1397 |
| Interpolated Age of participant when non-cancer illness first diagnosed (0.3) | 0.102 | 5.1 | 0.072 | 2.9 | 42.7 | 1794 |
| Interpolated Year when non-cancer illness first diagnosed (0.3) | 0.104 | 5.3 | 0.074 | 3 | 41.5 | 1794 |
| Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones (2.0) | 0.153 | 40.4 | 0.113 | 22.2 | 36.4 | 7507 |
| Treatment/medication code (1140884600 - metformin) | 0.121 | 48.1 | 0.092 | 28.1 | 32 | 14503 |
| Number of treatments/medications taken (0.0) | 0.13 | 55.6 | 0.099 | 32.4 | 31.8 | 14503 |
| Age started wearing glasses or contact lenses (2.0) | 0.158 | 73.8 | 0.124 | 45.8 | 27.3 | 13129 |
| Number of self-reported non-cancer illnesses (0.0) | 0.115 | 43.7 | 0.093 | 28.6 | 24.4 | 14503 |
| Number of treatments/medications taken (1.0) | 0.13 | 16.7 | 0.107 | 11.7 | 21.4 | 4134 |
| Age started wearing glasses or contact lenses (0.0) | 0.149 | 61.6 | 0.124 | 43.2 | 19.8 | 12321 |
| Number of treatments/medications taken (2.0) | 0.188 | 114.4 | 0.157 | 80.2 | 19.4 | 14435 |
| Treatment/medication code (1140879802 - amlodipine) | 0.113 | 41.9 | 0.096 | 30.5 | 17.8 | 14503 |
| Medication for cholesterol, blood pressure or diabetes (0.0) | 0.15 | 35.1 | 0.131 | 26.7 | 15.1 | 6756 |
| Diabetes diagnosed by doctor (2.0) | 0.124 | 49.6 | 0.108 | 38.3 | 14.2 | 14379 |
| Non-cancer illness code, self-reported (1220 - diabetes) | 0.109 | 39.1 | 0.098 | 31.8 | 11.1 | 14503 |
| Diagnoses - secondary ICD10 (I10 - I10 Essential (primary) hypertension) | 0.146 | 69.8 | 0.135 | 59.5 | 8.4 | 14503 |
| Overall health rating (0.0) | 0.113 | 41.6 | 0.104 | 35.7 | 8.1 | 14477 |
| Non-cancer illness code, self-reported (1065 - hypertension) | 0.218 | 155.7 | 0.203 | 133.6 | 7.9 | 14503 |
| Medication for cholesterol, blood pressure or diabetes (2.0) | 0.145 | 33.2 | 0.136 | 29.3 | 6.7 | 6829 |
| Overall health rating (2.0) | 0.11 | 39.5 | 0.113 | 41.7 | -2.8 | 14388 |
| Non-cancer illness code, self-reported (1261 - multiple sclerosis) | 0.111 | 40.2 | 0.118 | 46 | -6.6 | 14503 |

**Table A.2:** Comparison of prediction performance of 158 nIDPs between FLICA (nIC=100) and 5,812 IDPs in the HCP dataset. We excluded an nIDP if both methods have prediction r-value$< 0.1$. The meanings of each variables can be found at HCP wiki: `https://wiki.humanconnectome.org/display/PublicData/HCP+Data+Dictionary+Public-+Updated+for+the+1200+Subject+Release`

| Variable Names | r FLICA | -log10(p) | r IDPs | -log10(p) | Improvement (%) | nSubj | Variable Names | r FLICA | -log10(p) | r IDPs | -log10(p) | Improvement (%) | nSubj |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Odor_AgeAdj | 0.108 | 3.5 | 0.036 | 0.9 | 197.4 | 1001 | ASR_Thot_Pct | 0.216 | 11.6 | 0.206 | 10.6 | 4.9 | 1000 |
| FamHist_Fath_DrgAlc | 0.158 | 6.6 | 0.065 | 1.7 | 144.1 | 999 | PicSeq_AgeAdj | 0.301 | 22 | 0.29 | 20.4 | 3.9 | 1003 |
| NEOFAC_C | 0.176 | 8 | 0.084 | 2.4 | 110 | 1001 | Avg_Weekend_Beer_Wine_Cooler_7days | 0.307 | 22.5 | 0.296 | 20.9 | 3.8 | 985 |
| ASR_Soma_Pct | 0.124 | 4.4 | 0.074 | 2 | 66.3 | 1000 | DSM_Somp_Pct | 0.12 | 4.1 | 0.115 | 3.9 | 3.6 | 1000 |
| ASR_Soma_Raw | 0.162 | 6.9 | 0.101 | 3.2 | 60.7 | 1000 | PicSeq_Unadj | 0.293 | 20.9 | 0.284 | 19.6 | 3.2 | 1003 |
| VSPLOT_CRTE | 0.13 | 4.7 | 0.085 | 2.4 | 53.2 | 1000 | ASR_Oth_Raw | 0.168 | 7.3 | 0.163 | 7 | 2.7 | 1000 |
| ASR_Intn_Raw | 0.101 | 3.2 | 0.067 | 1.8 | 51.4 | 1000 | ASR_Thot_Raw | 0.231 | 13.2 | 0.226 | 12.6 | 2.4 | 1000 |
| DSM_Inat_Raw | 0.161 | 6.8 | 0.109 | 3.5 | 48.3 | 1000 | VSPLOT_OFF | 0.417 | 42.8 | 0.407 | 40.8 | 2.3 | 1000 |
| SSAGA_Alc_D4_Dp_Sx | 0.204 | 10.4 | 0.139 | 5.3 | 46.6 | 1002 | SSAGA_Alc_Hvy_Frq_Drk | 0.244 | 13.9 | 0.239 | 13.4 | 2 | 951 |
| AngHostil_Unadj | 0.103 | 3.2 | 0.072 | 2 | 42 | 1002 | Avg_Weekend_Cigarettes_7days | 0.181 | 8.3 | 0.178 | 8 | 1.9 | 985 |
| ProcSpeed_AgeAdj | 0.24 | 14.2 | 0.175 | 7.9 | 37.2 | 1003 | SSAGA_Alc_Hvy_Max_Drinks | 0.466 | 52 | 0.459 | 50.3 | 1.5 | 951 |
| DSM_Depr_Raw | 0.118 | 4.1 | 0.087 | 2.5 | 36.7 | 1000 | ASR_Extn_T | 0.186 | 8.8 | 0.183 | 8.6 | 1.4 | 1000 |
| Times_Used_Any_Tobacco_Today | 0.157 | 6.5 | 0.117 | 3.9 | 34.8 | 985 | NEOFAC_N | 0.159 | 6.6 | 0.157 | 6.5 | 0.9 | 1001 |
| ProcSpeed_Unadj | 0.268 | 17.6 | 0.201 | 10.2 | 33.5 | 1003 | ASR_Rule_Raw | 0.285 | 19.7 | 0.283 | 19.4 | 0.6 | 1000 |
| ASR_Intr_Raw | 0.189 | 9.1 | 0.143 | 5.5 | 32.3 | 1000 | ER40SAD | 0.117 | 4 | 0.117 | 4 | 0.6 | 1000 |
| SSAGA_Mj_Ab_Dep | 0.127 | 4.5 | 0.097 | 3 | 31.2 | 1002 | SSAGA_Alc_Hvy_Frq_5plus | 0.3 | 20.8 | 0.299 | 20.6 | 0.5 | 951 |
| SSAGA_Alc_D4_Ab_Sx | 0.153 | 6.3 | 0.119 | 4.1 | 29.4 | 1002 | ListSort_Unadj | 0.397 | 38.7 | 0.4 | 39.3 | -0.7 | 1003 |
| NEOFAC_O | 0.253 | 15.6 | 0.195 | 9.6 | 29.2 | 1001 | AngAggr_Unadj | 0.28 | 19 | 0.283 | 19.4 | -1.1 | 1002 |
| Flanker_AgeAdj | 0.181 | 8.4 | 0.142 | 5.5 | 27.5 | 1003 | VSPLOT_TC | 0.361 | 31.8 | 0.367 | 32.9 | -1.7 | 1000 |
| ASR_Intr_Pct | 0.161 | 6.8 | 0.127 | 4.5 | 26.8 | 1000 | FearAffect_Unadj | 0.118 | 4.1 | 0.12 | 4.2 | -1.9 | 1002 |
| DDisc_AUC_40K | 0.341 | 28.3 | 0.277 | 18.7 | 23 | 1000 | Total_Drinks_7days | 0.28 | 18.7 | 0.286 | 19.5 | -2.1 | 985 |
| SSAGA_Alc_12_Drinks_Per_Day | 0.336 | 26 | 0.277 | 17.7 | 21.3 | 950 | NEOFAC_A | 0.223 | 12.4 | 0.229 | 12.9 | -2.3 | 1001 |
| Flanker_Unadj | 0.239 | 14 | 0.197 | 9.8 | 20.9 | 1003 | Total_Cigarettes_7days | 0.174 | 7.7 | 0.178 | 8 | -2.4 | 985 |
| SSAGA_Times_Used_Stimulants | 0.117 | 4 | 0.098 | 3 | 19.9 | 1002 | ListSort_AgeAdj | 0.393 | 38.1 | 0.406 | 40.6 | -3 | 1003 |
| ASR_Witd_Raw | 0.23 | 13 | 0.192 | 9.3 | 19.5 | 1000 | DSM_Hype_Raw | 0.163 | 6.9 | 0.17 | 7.5 | -4 | 1000 |
| PicVocab_Unadj | 0.572 | 87.7 | 0.487 | 60.5 | 17.4 | 1003 | Avg_Weekday_Any_Tobacco_7days | 0.183 | 8.5 | 0.191 | 9.1 | -4.3 | 985 |
| ER40ANG | 0.102 | 3.2 | 0.087 | 2.5 | 16.9 | 1000 | Avg_Weekend_Drinks_7days | 0.265 | 16.8 | 0.278 | 18.5 | -4.8 | 985 |
| Avg_Weekday_Drinks_7days | 0.197 | 9.7 | 0.169 | 7.3 | 16.8 | 985 | SSAGA_Alc_12_Frq | 0.228 | 12.3 | 0.24 | 13.5 | -4.8 | 951 |
| EmotSupp_Unadj | 0.128 | 4.6 | 0.11 | 3.6 | 16.5 | 1002 | ASR_Rule_Pct | 0.185 | 8.7 | 0.194 | 9.5 | -5 | 1000 |
| Dexterity_AgeAdj | 0.316 | 24.3 | 0.271 | 18 | 16.5 | 1003 | ASR_Aggr_Raw | 0.103 | 3.2 | 0.109 | 3.5 | -5.3 | 1000 |
| SSAGA_Alc_Hvy_Drinks_Per_Day | 0.335 | 25.9 | 0.288 | 19.1 | 16.3 | 950 | THC | 0.264 | 17 | 0.281 | 19.2 | -6 | 1003 |
| CardSort_AgeAdj | 0.345 | 28.9 | 0.297 | 21.4 | 16 | 1001 | Total_Any_Tobacco_7days | 0.192 | 9.2 | 0.205 | 10.4 | -6.2 | 985 |
| PicVocab_AgeAdj | 0.57 | 87 | 0.492 | 61.9 | 15.8 | 1003 | Avg_Weekday_Cigarettes_7days | 0.167 | 7.2 | 0.178 | 8.1 | -6.3 | 985 |
| Taste_AgeAdj | 0.186 | 8.8 | 0.161 | 6.8 | 15.6 | 998 | SSAGA_Times_Used_Illicits | 0.147 | 5.8 | 0.158 | 6.6 | -6.5 | 1002 |
| DSM_Antis_Pct | 0.168 | 7.3 | 0.146 | 5.7 | 15.4 | 1000 | PercStress_Unadj | 0.165 | 7.1 | 0.177 | 8.1 | -6.7 | 1002 |
| SSAGA_Times_Used_Hallucinogens | 0.163 | 7 | 0.143 | 5.6 | 14.1 | 1002 | SSAGA_Alc_12_Frq_5plus | 0.324 | 24.3 | 0.349 | 28.1 | -7.1 | 951 |
| DSM_Adh_Pct | 0.219 | 11.9 | 0.192 | 9.3 | 14.1 | 1000 | ASR_Crit_Raw | 0.185 | 8.7 | 0.2 | 10.1 | -7.5 | 1000 |
| Taste_Unadj | 0.188 | 9 | 0.165 | 7.1 | 13.9 | 998 | MMSE_Score | 0.175 | 7.9 | 0.189 | 9.1 | -7.6 | 1003 |
| ASR_Attn_Raw | 0.147 | 5.8 | 0.129 | 4.7 | 13.8 | 1000 | Num_Days_Used_Any_Tobacco_7days | 0.218 | 11.8 | 0.238 | 14 | -8.7 | 1000 |
| Dexterity_Unadj | 0.302 | 22.2 | 0.266 | 17.3 | 13.5 | 1003 | PercHostil_Unadj | 0.107 | 3.4 | 0.117 | 4 | -8.8 | 1002 |
| SSAGA_Alc_12_Max_Drinks | 0.409 | 39.2 | 0.36 | 30.1 | 13.5 | 951 | SCPT_SPEC | 0.185 | 8.7 | 0.204 | 10.5 | -9.7 | 1000 |
| DDisc_AUC_200 | 0.295 | 21.2 | 0.26 | 16.5 | 13.4 | 1000 | ER40_CR | 0.158 | 6.6 | 0.176 | 8 | -10.4 | 1000 |
| DSM_Depr_Pct | 0.119 | 4.1 | 0.105 | 3.4 | 13.4 | 1000 | ASR_Totp_Raw | 0.151 | 6.1 | 0.171 | 7.6 | -11.8 | 1000 |
| SSAGA_ChildhoodConduct | 0.192 | 9.4 | 0.169 | 7.5 | 13.4 | 1002 | SSAGA_TB_Smoking_History | 0.148 | 5.9 | 0.168 | 7.4 | -11.8 | 1002 |
| ASR_Witd_Pct | 0.151 | 6.1 | 0.134 | 4.9 | 13.1 | 1000 | IWRD_TOT | 0.174 | 7.8 | 0.198 | 9.9 | -12.1 | 1000 |
| CardSort_Unadj | 0.36 | 31.6 | 0.319 | 24.7 | 12.9 | 1001 | ASR_TAO_Sum | 0.169 | 7.5 | 0.195 | 9.6 | -13.1 | 1000 |
| SSAGA_Alc_Hvy_Frq | 0.254 | 15.1 | 0.226 | 12.1 | 12.5 | 951 | Num_Days_Drank_7days | 0.183 | 8.6 | 0.211 | 11.1 | -13.3 | 1000 |
| PMAT24_A_CR | 0.512 | 67.4 | 0.456 | 51.9 | 12.3 | 999 | Avg_Weekend_Any_Tobacco_7days | 0.193 | 9.3 | 0.224 | 12.2 | -13.6 | 985 |
| LifeSatisf_Unadj | 0.208 | 10.8 | 0.186 | 8.8 | 11.8 | 1002 | FearSomat_Unadj | 0.106 | 3.4 | 0.125 | 4.4 | -15.1 | 1002 |
| DSM_Somp_Raw | 0.154 | 6.3 | 0.138 | 5.2 | 11.3 | 1000 | ASR_Totp_T | 0.153 | 6.2 | 0.183 | 8.5 | -16.1 | 1000 |
| DSM_Adh_Raw | 0.182 | 8.5 | 0.165 | 7.1 | 10.4 | 1000 | DSM_Anxi_Raw | 0.143 | 5.6 | 0.174 | 7.8 | -17.6 | 1000 |
| Avg_Weekday_Beer_Wine_Cooler_7days | 0.286 | 19.6 | 0.26 | 16.3 | 10 | 985 | SSAGA_TB_Still_Smoking | 0.181 | 8.4 | 0.225 | 12.5 | -19.5 | 1002 |
| PSQI_Score | 0.168 | 7.4 | 0.153 | 6.3 | 9.9 | 1003 | SCPT_SEN | 0.083 | 2.4 | 0.106 | 3.4 | -21.2 | 1000 |
| Total_Beer_Wine_Cooler_7days | 0.36 | 31.1 | 0.329 | 25.9 | 9.5 | 985 | SSAGA_Mj_Times_Used | 0.172 | 7.7 | 0.221 | 12.1 | -21.9 | 1002 |
| ASR_Attn_Pct | 0.188 | 9 | 0.173 | 7.7 | 8.8 | 1000 | MeanPurp_Unadj | 0.085 | 2.4 | 0.11 | 3.6 | -22.8 | 1002 |
| SSAGA_Alc_12_Frq_Drk | 0.275 | 17.5 | 0.257 | 15.3 | 7.1 | 951 | PercReject_Unadj | 0.083 | 2.4 | 0.108 | 3.5 | -23.4 | 1002 |
| ReadEng_AgeAdj | 0.494 | 62.3 | 0.464 | 54.4 | 6.3 | 1003 | NEOFAC_E | 0.102 | 3.2 | 0.139 | 5.3 | -26.8 | 1001 |
| DSM_Antis_Raw | 0.244 | 14.5 | 0.229 | 13 | 6.2 | 1000 | DSM_Anxi_Pct | 0.088 | 2.6 | 0.122 | 4.3 | -27.7 | 1000 |
| ASR_Extn_Raw | 0.203 | 10.3 | 0.191 | 9.3 | 6 | 1000 | Avg_Weekend_Hard_Liquor_7days | 0.074 | 2 | 0.106 | 3.4 | -29.8 | 985 |
| ASR_Aggr_Pct | 0.119 | 4.1 | 0.113 | 3.7 | 5.6 | 1000 | SSAGA_Alc_Age_1st_Use | 0.136 | 4.9 | 0.202 | 9.8 | -32.8 | 951 |
| ReadEng_Unadj | 0.487 | 60.4 | 0.464 | 54.2 | 5 | 1003 | Total_Hard_Liquor_7days | 0.07 | 1.9 | 0.135 | 4.9 | -47.9 | 985 |

**Table A.3:** Three examples of top 10 most significant correlations of BigFLICA modes (left) and IDPs (right) with nIDPs in UKB dataset.

| BigFLICA modes | r-value | p-value | IDP names | r-value | p-value |
|---|---|---|---|---|---|
| **Top 10 modes/IDPs correlate with *fluid intelligence*** | | | | | |
| IC25 | -0.146 | 6.54E-65 | IDP_tfMRI_90th-percentile_BOLD_shapes | -0.074 | 1.05E-15 |
| IC57 | -0.122 | 3.49E-45 | IDP_tfMRI_median_BOLD_shapes | -0.069 | 1.28E-13 |
| IC332 | -0.081 | 7.26E-21 | IDP_tfMRI_90th-percentile_zstat_faces-shapes_amygdala | 0.068 | 1.48E-13 |
| IC484 | -0.072 | 9.87E-17 | rfMRI amplitudes (ICA25 node 6) | 0.066 | 2.75E-13 |
| IC4 | -0.069 | 1.18E-15 | rfMRI connectivity (ICA100: IC13-IC32) | -0.065 | 3.82E-13 |
| IC27 | 0.064 | 1.98E-13 | IDP_tfMRI_90th-percentile_zstat_shapes | -0.063 | 7.95E-12 |
| IC188 | -0.058 | 2.80E-11 | IDP_tfMRI_median_zstat_faces-shapes | 0.063 | 1.20E-11 |
| IC708 | -0.055 | 1.51E-10 | IDP_tfMRI_median_zstat_faces-shapes_amygdala | 0.062 | 1.51E-11 |
| IC164 | 0.055 | 1.68E-10 | rfMRI connectivity (ICA100: IC11-IC19) | 0.056 | 3.94E-10 |
| IC47 | -0.054 | 3.40E-10 | IDP_tfMRI_median_BOLD_faces-shapes | 0.057 | 5.16E-10 |
| **Top 10 modes/IDPs correlate with *Age started wearing glasses or contact lenses*** | | | | | |
| IC164 | 0.101 | 2.74E-31 | IDP_tfMRI_90th-percentile_BOLD_faces-shapes | 0.081 | 5.38E-18 |
| IC25 | 0.067 | 1.63E-14 | IDP_tfMRI_median_BOLD_faces-shapes | 0.064 | 7.16E-12 |
| IC249 | 0.057 | 5.65E-11 | IDP_tfMRI_median_zstat_faces-shapes | 0.064 | 1.01E-11 |
| IC13 | 0.054 | 7.68E-10 | rfMRI connectivity (ICA100: IC4-IC40) | -0.061 | 1.81E-11 |
| IC138 | 0.041 | 2.33E-06 | rfMRI connectivity (ICA100: IC17-IC42) | -0.060 | 4.34E-11 |
| IC190 | 0.038 | 1.65E-05 | IDP_tfMRI_90th-percentile_zstat_faces-shapes | 0.061 | 9.88E-11 |
| IC563 | -0.037 | 1.84E-05 | rfMRI amplitudes (ICA25 node 19) | 0.056 | 8.97E-10 |
| IC656 | 0.037 | 1.98E-05 | rfMRI amplitudes (ICA100 node 16) | 0.054 | 3.08E-09 |
| IC297 | -0.037 | 2.32E-05 | rfMRI connectivity (ICA100: IC15-IC43) | -0.050 | 3.14E-08 |
| IC580 | 0.037 | 2.53E-05 | rfMRI connectivity (ICA100: IC21-IC28) | 0.049 | 6.90E-08 |
| **Top 10 modes/IDPs correlate with *hypertension*** | | | | | |
| IC259 | 0.121 | 2.95E-48 | IDP_dMRI_TBSS_MD_External_capsule_L | 0.135 | 4.17E-53 |
| IC38 | -0.111 | 3.09E-41 | IDP_dMRI_TBSS_MD_External_capsule_R | 0.132 | 6.44E-51 |
| IC319 | 0.094 | 5.98E-30 | IDP_dMRI_TBSS_L1_External_capsule_L | 0.130 | 2.29E-49 |
| IC29 | 0.091 | 5.82E-28 | IDP_dMRI_TBSS_L3_External_capsule_R | 0.126 | 6.51E-47 |
| IC40 | 0.088 | 2.82E-26 | IDP_dMRI_TBSS_L3_External_capsule_L | 0.126 | 1.70E-46 |
| IC1 | -0.083 | 1.57E-23 | IDP_dMRI_TBSS_L3_Anterior_limb_of_internal_capsule_L | 0.122 | 4.64E-44 |
| IC171 | -0.073 | 1.00E-18 | IDP_dMRI_TBSS_L1_External_capsule_R | 0.122 | 5.37E-44 |
| IC26 | -0.067 | 6.73E-16 | IDP_dMRI_TBSS_ISOVF_External_capsule_L | 0.122 | 7.49E-44 |
| IC176 | -0.062 | 6.75E-14 | IDP_dMRI_TBSS_L2_Anterior_limb_of_internal_capsule_L | 0.120 | 2.60E-42 |
| IC84 | 0.057 | 9.14E-12 | IDP_dMRI_TBSS_MD_Anterior_limb_of_internal_capsule_L | 0.118 | 2.91E-41 |
| **Top 10 modes/IDPs correlate with *handedness*** | | | | | |
| IC235 | -0.226 | 5.71E-168 | rfMRI connectivity (ICA100: IC29-IC34) | 0.115 | 1.30E-40 |
| IC408 | -0.079 | 1.76E-21 | rfMRI connectivity (ICA25: IC1-IC6) | 0.095 | 6.93E-28 |
| IC569 | 0.066 | 1.90E-15 | rfMRI connectivity (ICA100: IC10-IC34) | 0.095 | 7.70E-28 |
| IC382 | 0.051 | 8.40E-10 | rfMRI connectivity (ICA25: IC14-IC22) | 0.085 | 7.43E-23 |
| IC251 | 0.047 | 1.13E-08 | rfMRI connectivity (ICA100: IC3-IC19) | 0.085 | 8.01E-23 |
| IC232 | 0.043 | 2.12E-07 | rfMRI connectivity (ICA100: IC14-IC34) | 0.081 | 7.28E-21 |
| IC742 | 0.042 | 4.41E-07 | rfMRI connectivity (ICA25: IC1-IC22) | 0.080 | 2.93E-20 |
| IC643 | -0.039 | 2.39E-06 | rfMRI connectivity (ICA100: IC30-IC34) | -0.074 | 1.46E-17 |
| IC419 | 0.036 | 1.66E-05 | rfMRI connectivity (ICA100: IC27-IC52) | 0.073 | 3.07E-17 |
| IC332 | 0.036 | 1.76E-05 | rfMRI connectivity (ICA100: IC6-IC13) | 0.071 | 3.39E-16 |

**Table A.4: Percent of shared variance (%) of BigFLICA decomposition across a range of dimensionalities in the UKB data.** Upper triangle: the explained variance of a lower-dimensional decomposition by a higher-dimensional decomposition. Lower triangle: the explained variance of a higher-dimensional decomposition by a lower-dimensional decomposition.

| | **IC25** | **IC100** | **IC250** | **IC500** | **IC750** |
|---|---|---|---|---|---|
| **IC25** | 100.00 | 99.98 | 99.99 | 99.99 | 99.99 |
| **IC100** | 88.57 | 100.00 | 99.97 | 99.98 | 99.99 |
| **IC250** | 86.38 | 96.99 | 100.00 | 99.98 | 99.99 |
| **IC500** | 87.77 | 95.45 | 96.79 | 100.00 | 99.97 |
| **IC750** | 85.33 | 95.31 | 96.91 | 99.65 | 100.00 |

**Table A.5:** 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 |

## A.3 Supplementary materials for Chapter 3

**Table A.6:** 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 |

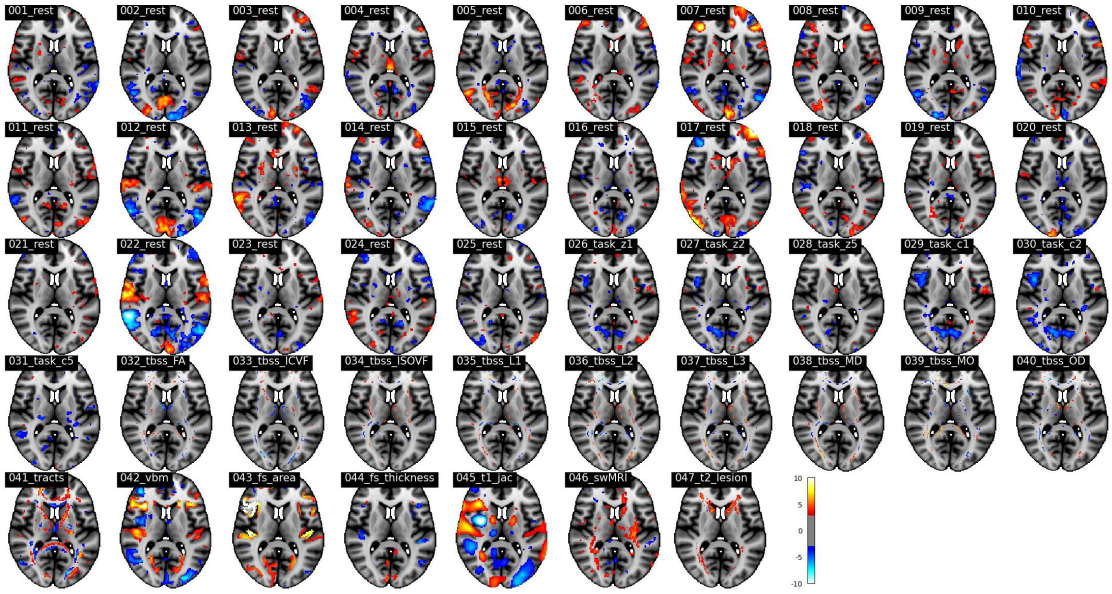



**Figure A.15:** The *Z*-score normalized spatial maps of the strongest modes that contributing to the prediction of *hypertension* in a 1,000-dimensional SuperBigFLICA, with all 17,485 nIDPs as supervision targets (MNI152 coordinate z=10).

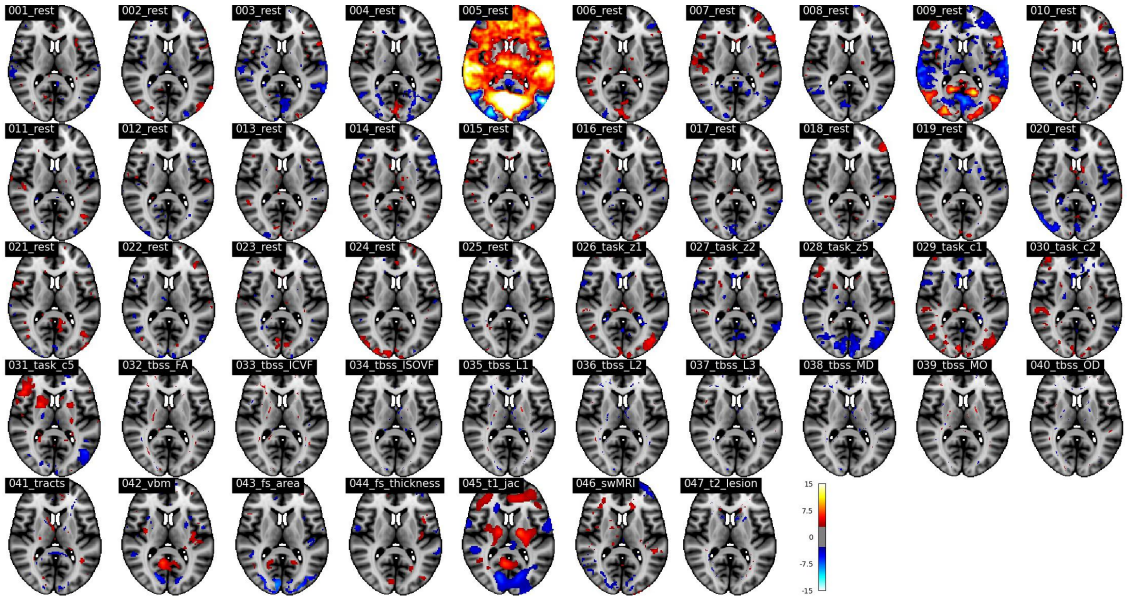



**Figure A.16:** The *Z*-score normalized spatial maps of the strongest modes that contributing to the prediction of *age started wearing glasses or contact lenses* in a 1,000-dimensional SuperBigFLICA, with all 17,485 nIDPs as supervision targets (MNI152 coordinate z=10).

# Bibliography

[Abrol et al., 2020] 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 and Berrada, 2018] 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 et al., 2018] 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 et al., 2021] 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.

[Allen et al., 2016] Allen, B., Muldoon, M. F., Gianaros, P. J., and Jennings, J. R. (2016). Higher blood pressure partially links greater adiposity to reduced brain white matter integrity. *American journal of hypertension*, 29(9):1029–1037.

[Allen et al., 2014] Allen, G. I., Grosenick, L., and Taylor, J. (2014). A generalized least-square matrix decomposition. *Journal of the American Statistical Association*, 109(505):145–159.

[Alnæs et al., 2018] Alnæs, D., Kaufmann, T., Doan, N. T., Córdova-Palomera, A., Wang, Y., Bettella, F., Moberget, T., Andreassen, O. A., and Westlye, L. T. (2018). Association of heritable cognitive ability and psychopathology with white matter properties in children and adolescents. *JAMA psychiatry*, 75(3):287–295.

[Amodio and Frith, 2006] 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.

[Ashburner and Friston, 2000] Ashburner, J. and Friston, K. J. (2000). Voxel-based morphometry—the methods. *Neuroimage*, 11(6):805–821.

[Ball et al., 2017] Ball, G., Aljabar, P., Nongena, P., Kennea, N., Gonzalez-Cinca, N., Falconer, S., Chew, A. T., Harper, N., Wurie, J., Rutherford, M. A., et al. (2017). Multimodal image analysis of clinical influences on preterm brain development. *Annals of neurology*, 82(2):233–246.

[Ball et al., 2018] Ball, G., Malpas, C. B., Genc, S., Efron, D., Sciberras, E., Anderson, V., Nicholson, J. M., and Silk, T. J. (2018). Multimodal structural neuroimaging markers of brain development and adhd symptoms. *American Journal of Psychiatry*, 176(1):57–66.

[Ball et al., 2019] 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.

[Baydin et al., 2018] Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018). Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18.

[Beckmann and Smith, 2004] 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.

[Beckmann and Smith, 2005] Beckmann, C. F. and Smith, S. M. (2005). Tensorial extensions of independent component analysis for multisubject fMRI analysis. *Neuroimage*, 25(1):294–311.

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

[Blei et al., 2017] Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

[Calhoun et al., 2006] Calhoun, V. D., Adali, T., Giuliani, N., Pekar, J., Kiehl, K., and Pearlson, G. (2006). Method for multimodal analysis of independent source differences in schizophrenia: combining gray matter structural and auditory oddball functional data. *Human brain mapping*, 27(1):47–62.

[Calhoun and Sui, 2016] 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 et al., 2020] 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.

[Choudrey, 2002] Choudrey, R. A. (2002). *Variational methods for Bayesian independent component analysis*. PhD thesis, University of Oxford Oxford, UK.

[Çiçek et al., 2016] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer.

[Cui and Gong, 2018] 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.

[Da Silva et al., 2021] Da Silva, M., Sudre, C. H., Garcia, K., Bass, C., Cardoso, M. J., and Robinson, E. C. (2021). Distinguishing healthy ageing from dementia: a biomechanical simulation of brain atrophy using deep networks. *arXiv preprint arXiv:2108.08214*.

[Dadi et al., 2020] 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 et al., 2013] 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 et al., 2005] 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.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

[Douaud et al., 2014] 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.

[Douaud et al., 2021] Douaud, G., Lee, S., Alfaro-Almagro, F., Arthofer, C., Wang, C., Lange, F., Andersson, J. L., Griffanti, L., Duff, E., Jbabdi, S., et al. (2021). Brain imaging before and after covid-19 in uk biobank. *medRxiv*.

[Efron et al., 2004] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of statistics*, 32(2):407–499.

[Eickhoff et al., 2018] 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 et al., 2018] 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.

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

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

[Friedman et al., 2010] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

[Gafuroğlu et al., 2018] Gafuroğlu, C., Rekik, I., et al. (2018). Joint prediction and classification of brain image evolution trajectories from baseline brain image with application to early dementia. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 437–445. Springer.

[Ganin et al., 2016] Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

[Ghribi et al., 2021] Ghribi, O., Li, G., Lin, W., Shen, D., and Rekik, I. (2021). Multi-regression based supervised sample selection for predicting baby connectome evolution trajectory from neonatal timepoint. *Medical Image Analysis*, 68:101853.

[Gilpin et al., 2018] 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.

[Glasser et al., 2013] Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., et al. (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124.

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

[Gong et al., 2021a] Gong, W., Bai, S., Zheng, Y.-Q., Smith, S., and Beckmann, C. (2021a). Supervised phenotype discovery from multimodal brain imaging. *bioRxiv*.

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

[Gong et al., 2021c] Gong, W., Beckmann, C. F., Vedaldi, A., Smith, S. M., and Peng, H. (2021c). Optimising a simple fully convolutional network for accurate brain age prediction in the pac 2019 challenge. *Frontiers in Psychiatry*, 12.

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

[Gordon et al., 2017] Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., et al. (2017). Precision functional mapping of individual human brains. *Neuron*, 95(4):791–807.

[Griffanti et al., 2014] Griffanti, L., Salimi-Khorshidi, G., Beckmann, C. F., Auerbach, E. J., Douaud, G., Sexton, C. E., Zsoldos, E., Ebmeier, K. P., Filippini, N., Mackay, C. E., et al. (2014). ICA-based artefact removal and accelerated fMRI acquisition for improved resting state network imaging. *Neuroimage*, 95:232–247.

[Griffanti et al., 2016] 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 et al., 2011] 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.

[Groves et al., 2012] Groves, A. R., Smith, S. M., Fjell, A. M., Tamnes, C. K., Walhovd, K. B., Douaud, G., Woolrich, M. W., and Westlye, L. T. (2012). Benefits of multi-modal fusion analysis on a large-scale dataset: life-span patterns of inter-subject variability in cortical morphometry and white matter microstructure. *Neuroimage*, 63(1):365–380.

[Hannawi et al., 2018] 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.

[Hariri et al., 2002] Hariri, A. R., Tessitore, A., Mattay, V. S., Fera, F., and Weinberger, D. R. (2002). The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage*, 17(1):317–323.

[He et al., 2020] 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.

[Hibar et al., 2018] Hibar, D., Westlye, L. T., Doan, N. T., Jahanshad, N., Cheung, J., Ching, C. R., Versace, A., Bilderbeck, A., Uhlmann, A., Mwangi, B., et al. (2018). Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. *Molecular psychiatry*, 23(4):932.

[Hoffman et al., 2013] Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(5).

[Hong et al., 2019] Hong, Y., Kim, J., Chen, G., Lin, W., Yap, P.-T., and Shen, D. (2019). Longitudinal prediction of infant diffusion mri data via graph convolutional adversarial networks. *IEEE transactions on medical imaging*, 38(12):2717–2725.

[Hoyos-Idrobo et al., 2019] Hoyos-Idrobo, A., Varoquaux, G., Kahn, J., and Thirion, B. (2019). Recursive nearest agglomeration (ReNA): fast clustering for approximation of structured signals. *IEEE transactions on pattern analysis and machine intelligence*, 41(3):669–681.

[Hyvarinen, 1999] Hyvarinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634.

[Hyvärinen and Oja, 2000] Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

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

[Isensee et al., 2021] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H. (2021). nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211.

[Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134.

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

[Jernigan et al., 2018] Jernigan, T. L., Brown, S. A., and Dowling, G. J. (2018). The adolescent brain cognitive development study.

[Jollans et al., 2019] 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*.

[Kariya and Kurata, 2004] Kariya, T. and Kurata, H. (2004). *Generalized least squares*. John Wiley & Sons.

[Kelly et al., 2017] Kelly, S., Jahanshad, N., Zalesky, A., Kochunov, P., Agartz, I., Alloza, C., Andreassen, O., Arango, C., Banaj, N., Bouix, S., et al. (2017). Widespread white matter microstructural differences in schizophrenia across 4322 individuals: results from the ENIGMA Schizophrenia DTI Working Group. *Molecular psychiatry*.

[Kendall et al., 2018] 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.

[Kettenring, 1971] Kettenring, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451.

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

[Kingma and Welling, 2013] Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Klami et al., 2015] Klami, A., Virtanen, S., Leppäaho, E., and Kaski, S. (2015). Group factor analysis. *IEEE transactions on neural networks and learning systems*, 26(9):2136–2147.

[Kong et al., 2018] Kong, R., Li, J., Orban, C., Sabuncu, M. R., Liu, H., Schaefer, A., Sun, N., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., et al. (2018). Spatial topography of individual-specific cortical networks predicts human cognition, personality, and emotion. *Cerebral Cortex*, 29(6):2533–2551.

[Lawrence, 2005] 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 et al., 2011] 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.

[Lebreton et al., 2019] Lebreton, M., Bavard, S., Daunizeau, J., and Palminteri, S. (2019). Assessing inter-individual differences with task-related functional neuroimaging. *Nature Human Behaviour*, 3(9):897–905.

[Lee et al., 2019] 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 et al., 2019] Li, H., Smith, S. M., Gruber, S., Lukas, S. E., Silveri, M. M., Hill, K. P., Killgore, W. D., and Nickerson, L. D. (2019). Denoising scanner effects from multimodal mri data using linked independent component analysis. *NeuroImage*, page 116388.

[Li et al., 2015] 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.

[Li et al., 2012] Li, Y.-O., Yang, F. G., Nguyen, C. T., Cooper, S. R., LaHue, S. C., Venugopal, S., and Mukherjee, P. (2012). Independent component analysis of DTI reveals multivariate microstructural correlations of white matter in the human brain. *Human brain mapping*, 33(6):1431–1451.

[Liégeois et al., 2019] Liégeois, R., Li, J., Kong, R., Orban, C., Van De Ville, D., Ge, T., Sabuncu, M. R., and Yeo, B. T. (2019). Resting brain dynamics at different timescales capture distinct aspects of human behavior. *Nature communications*, 10(1):2317.

[Liu and Nocedal, 1989] 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 et al., 2009] Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N. I., and Calhoun, V. (2009). Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ica. *Human brain mapping*, 30(1):241–255.

[Liu et al., 2020] 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 et al., 2018] 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.

[Mairal et al., 2010] Mairal, J., Bach, F., Ponce, J., and Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11(Jan):19–60.

[Marquand et al., 2014] 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 et al., 2013] 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 et al., 2016] 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.

[Mills and Tamnes, 2014] Mills, K. L. and Tamnes, C. K. (2014). Methods and considerations for longitudinal structural brain imaging analysis across development. *Developmental cognitive neuroscience*, 9:172–190.

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

[Moon et al., 2005] Moon, W.-J., Na, D. G., Kim, S. S., Ryoo, J. W., and Chung, E. C. (2005). Diffusion abnormality of deep gray matter in external capsular hemorrhage. *American journal of neuroradiology*, 26(2):229–235.

[Nebli et al., 2020] Nebli, A., Kaplan, U. A., and Rekik, I. (2020). Deep evographnet architecture for time-dependent brain graph data synthesis from a single timepoint. In *International Workshop on PRedictive Intelligence In MEdicine*, pages 144–155. Springer.

[Ngo et al., 2021] Ngo, G., Khosla, M., Jamison, K., Kuceyeski, A., and Sabuncu, M. R. (2021). Predicting individual task contrasts from resting-state functional connectivity using a surface-based convolutional network. *bioRxiv*.

[Nguyen et al., 2020] 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.

[Oktay et al., 2018] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*.

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

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

[Parra, 2018] Parra, L. C. (2018). Multi-set canonical correlation analysis simply explained. *arXiv preprint arXiv:1802.03759*.

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

[Peng et al., 2021] Peng, H., Gong, W., Beckmann, C. F., Vedaldi, A., and Smith, S. M. (2021). Accurate brain age prediction with lightweight deep neural networks. *Medical image analysis*, 68:101871.

[Penny et al., 2011] 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 et al., 2017] 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 et al., 2018] 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.

[Rachmadi et al., 2020] Rachmadi, M. F., Valdés-Hernández, M. d. C., Makin, S., Wardlaw, J., and Komura, T. (2020). Automatic spatial estimation of white matter hyperintensities evolution in brain mri using disease evolution predictor deep neural networks. *Medical image analysis*, 63:101712.

[Rahim et al., 2017] 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.

[Ranganath et al., 2014] Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR.

[Roberts and Choudrey, 2004] Roberts, S. and Choudrey, R. (2004). Bayesian independent component analysis with prior constraints: An application in biosignal analysis. In *International Workshop on Deterministic and Statistical Methods in Machine Learning*, pages 159–179. Springer.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

[Santarnecchi et al., 2017] 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.

[Schaefer et al., 2017] Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., and Yeo, B. T. (2017). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114.

[Schulz et al., 2020] 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 et al., 2020] 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 et al., 2014] Smith, S. M., Hyvärinen, A., Varoquaux, G., Miller, K. L., and Beckmann, C. F. (2014). Group-PCA for very large fMRI datasets. *Neuroimage*, 101:738–749.

[Smith et al., 2006] 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 et al., 2004] 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.

[Smith and Nichols, 2018] Smith, S. M. and Nichols, T. E. (2018). Statistical challenges in "big data" human neuroimaging. *Neuron*, 97(2):263–268.

[Smith et al., 2015] Smith, S. M., Nichols, T. E., Vidaurre, D., Winkler, A. M., Behrens, T. E., Glasser, M. F., Ugurbil, K., Barch, D. M., Van Essen, D. C., and Miller, K. L. (2015). A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature neuroscience*, 18(11):1565.

[Srivastava et al., 2014] 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 et al., 2012] Sui, J., Adali, T., Yu, Q., Chen, J., and Calhoun, V. D. (2012). A review of multivariate methods for multimodal fusion of brain imaging data. *Journal of neuroscience methods*, 204(1):68–81.

[Sui et al., 2018] 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 et al., 2013] 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.

[Tavor et al., 2016] Tavor, I., Jones, O. P., Mars, R. B., Smith, S., Behrens, T., and Jbabdi, S. (2016). Task-free mri predicts individual differences in brain activity during task performance. *Science*, 352(6282):216–220.

[Tieleman and Hinton, 2012] 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 and Roebroeck, 2014] Uludağ, K. and Roebroeck, A. (2014). General overview on the merits of multimodal neuroimaging data fusion. *Neuroimage*, 102:3–10.

[Van Essen et al., 2013] Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., Consortium, W.-M. H., et al. (2013). The WU-Minn human connectome project: an overview. *Neuroimage*, 80:62–79.

[Van Rooij et al., 2017] Van Rooij, D., Anagnostou, E., Arango, C., Auzias, G., Behrmann, M., Busatto, G. F., Calderoni, S., Daly, E., Deruelle, C., Di Martino, A., et al. (2017). Cortical and subcortical brain morphometry differences between patients with autism spectrum disorder and healthy individuals across the lifespan: results from the ENIGMA ASD working group. *American Journal of Psychiatry*, 175(4):359–369.

[Varoquaux et al., 2011] 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.

[Wainwright and Jordan, 2008] Wainwright, M. J. and Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

[Wang et al., 2017] 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.

[Wang et al., 2020] Wang, Z., Zou, N., Shen, D., and Ji, S. (2020). Non-local u-nets for biomedical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6315–6322.

[Wiberg et al., 2019] Wiberg, A., Ng, M., Al Omran, Y., Alfaro Almagro, F., McCarthy, P., Marchini, J., Bennett, D., Smith, S., Douaud, G., and Furniss, D. (2019). Handedness, language areas and neuropsychiatric diseases: insights from brain imaging and genetics. *Brain*.

[Wipf and Nagarajan, 2008] 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.

[Xia et al., 2021] Xia, T., Chartsias, A., Wang, C., and Tsaftaris, S. A. (2021). Learning to synthesise the ageing brain without longitudinal data. *Medical Image Analysis*, page 102169.

[Zeighami et al., 2015] Zeighami, Y., Ulla, M., Iturria-Medina, Y., Dadar, M., Zhang, Y., Larcher, K. M.-H., Fonov, V., Evans, A. C., Collins, D. L., and Dagher, A. (2015). Network structure of brain atrophy in de novo parkinson's disease. *Elife*, 4:e08440.

[Zhang et al., 2012a] 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 et al., 2012b] 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 and Yang, 2017] Zhang, Y. and Yang, Q. (2017). A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*.

[Zheng et al., 2021] Zheng, Y.-Q., Farahibozorg, S.-R., Gong, W., Rafipoor, H., Jbabdi, S., and Smith, S. (2021). Accurate predictions of individual differences in task-evoked brain activity from resting-state fmri using a sparse ensemble learner. *bioRxiv*.

[Zhou et al., 2020] 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.

[Zhou et al., 2018] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., and Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 3–11. Springer.

[Zhu et al., 2021] Zhu, Y., Kim, M., Zhu, X., Kaufer, D., Wu, G., Initiative, A. D. N., et al. (2021). Long range early diagnosis of alzheimer's disease using longitudinal mr imaging data. *Medical image analysis*, 67:101825.

[Zou and Hastie, 2005] 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.