Predicting sex, age, general cognition and mental health with machine learning on brain structural connectomes

Hon Wah Yeung[1], Aleks Stolicyn[1], Colin R. Buchanan[2,9,10], Elliot M. Tucker-Drob[3,4], Mark E. Bastin[5,9,10], Saturnino Luz[6], Andrew M. McIntosh[1,7], Heather C. Whalley[1,*], Simon R. Cox[2,9,10,*], and Keith Smith[8,*]

[1]Department of Psychiatry, University of Edinburgh, Edinburgh, United Kingdom
[2]Department of Psychology, University of Edinburgh, Edinburgh, United Kingdom
[3]Department of Psychology, University of Texas, Austin, TX, USA
[4]Population Research Center and Center on Aging and Population Sciences, University of Texas at Austin, TX, USA
[5]Centre for Clinical Brain Science, University of Edinburgh, Edinburgh, United Kingdom
[6]Usher Institute, Edinburgh Medical School, The University of Edinburgh, Edinburgh, United Kingdom
[7]Centre for Genomic and Experimental Medicine, Institute of Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK
[8]Department of Physics and Mathematics, Nottingham Trent University, Nottingham, United Kingdom
[9]Lothian Birth Cohorts, University of Edinburgh, Edinburgh, United Kingdom
[10]Scottish Imaging Network, A Platform for Scientific Excellence Collaboration (SINAPSE), Edinburgh, United Kingdom
[*]These authors share joint senior authorship

February 28, 2022

Abstract

There is increasing expectation that advanced, computationally expensive machine learning techniques, when applied to large population-wide neuroimaging datasets, will help to uncover key differences in the human brain in health and disease. We take a comprehensive approach to explore how multiple aspects of brain structural connectivity can predict sex, age, general cognitive function and general psychopathology, testing different machine learning algorithms from deep learning model (BrainNetCNN) to classical machine learning methods. We modelled $N = 8,183$ structural connectomes from UK Biobank using six different structural network weightings obtained from diffusion MRI. Streamline count generally provided highest prediction accuracies in all prediction tasks. Deep learning did not improve on prediction accuracies from simpler linear models. Further, high correlations between model coefficients from deep learning and linear models suggested similar decision mechanisms were used. This highlights that model complexity is unlikely to improve detection of associations between structural connectomes and complex phenotypes.

Introduction

Recent advances in neuroimaging have made it possible to acquire high quality structural brain scans in large samples, which provides opportunities for the development and application of novel machine learning (ML) techniques geared specifically for analysing brain structural architecture. Machine learning methods have the ability to uncover latent features from high-dimensional imaging data which are not apparent when using more conventional statistical methods. This is thought to be particularly promising for better understanding important biological and complex behavioural information.

The structural connectome is a comprehensive description of the connectivity between different brain regions, defined by the structure of white-matter fibre tracts (Bullmore and Sporns, 2009). The connectome is a relatively recent development which holds promise for discovery. One might expect deep learning methods to be well suited for the analysis of connnectomes, given their complexity and high-dimensionality. The structural connectome is based on information from diffusion magnetic

resonance imaging (dMRI), which measures the directional diffusion of water molecules in the brain to identify connections between distal brain regions and enables estimation microstructural properties in brain white matter. The measures acquired through dMRI can be used to reconstruct individual participant connectomes in the form of adjacency matrices, where row and column indices specify to each brain region and the matrix entries are the connectivity strengths (which can be measured using various diffusion parameters) between any two regions of interest (ROI). The weighted networks constructed from different connectivity modalities then represent different interpretations of connection strength (Agosta et al., 2014; Buchanan et al., 2020; Collin et al., 2014; Hagmann et al., 2008; Jiang et al., 2020b; Robinson et al., 2010; Rutland et al., 2019).

The most commonly used connectivity weightings are based on streamline count (SC), fractional anisotropy (FA) and mean diffusivity (MD). SC provides both intra-regional as well as inter-regional streamline densities (Hagmann et al., 2008), FA measures the degree of directional dependence of the water molecular diffusion (Basser and Pierpaoli, 2011; Robinson et al., 2010), and MD measures the average magnitude of diffusion of water molecules in all directions (Agosta et al., 2014; Alexander et al., 2007; Collin et al., 2014). In addition, three newer network weightings are available thanks to the development of Neurite Orientation Dispersion and Density Imaging (NODDI) (Zhang et al., 2012). NODDI estimates include neurite density (intra-cellular volume fraction; ICVF), extra-cellular water diffusion (isotropic volume fraction; ISOVF) and tract fanning/complexity (orientation dispersion; OD).

There is accruing evidence that differences in connectomic properties between people, particularly for global network measures, are associated with basic demographic variables namely age and sex, both of which are important predictors of brain health (Buchanan et al., 2020; Madole et al., 2021; Ritchie et al., 2018). Connectomic differences are also associated with psychiatric and neurological brain disorders, and that some connectome alterations are shared across multiple brain disorders (de Lange et al., 2019; Korgaonkar et al., 2014; Ma et al., 2020). Conversely, some studies do not find significant psychiatric disorder-related connectomic abnormalities (Qin et al., 2014; Zhu et al., 2016). While univariate statistical methods allow assessing relationships between measures of interest and brain disorders, machine learning (ML) approaches allow combining information from different features for diagnostic classification, prediction of symptom severity levels and prediction of treatment response. More recently, researchers have been exploring ways of applying ML to connectome data. Some examples include the development of the Connectome-based Predictive Modelling approach and application of the Support Vector Machines for predicting cognitive ability, behavioral measures and clinical outcomes (Finn et al., 2015; Gong and He, 2015; Griffa et al., 2013; Jiang et al., 2020a; Payabvash et al., 2019).

Although an individual connectome provides a lower-dimensional representation of the brain when compared to a voxel-wise MRI image, there are still thousands of features - connections between brain regions - which may lead to over-fitting. To reduce dimensionality, global graph-theoretic metrics are commonly derived from connectomes and analysed as coarse representations of the participants' brain networks (de Lange et al., 2019; Sun et al., 2017; Suo et al., 2018). A popular alternative technique is subnetwork analysis, where brain regions in the connectome are limited to those previously shown as relevant to the clinical traits or conditions of interest (Chen et al., 2016; Zheng et al., 2019). It is also common to employ feature selection to reduce dimensionality (Shen et al., 2017). Some of the common prediction models used with high-dimensional connectomic data are LASSO regression (Madole et al., 2021; Sripada et al., 2019; Wager et al., 2013), ridge regression (Gao et al., 2019; Siegel et al., 2016), elastic-net (Rahim et al., 2017), multiple kernel learning Support Vector Machine (MKL-SVM) (Xu et al., 2020), as well as relevance vector regression (Gong et al., 2014). Most of these models only consider linear interactions between ROIs

and treat each connection independently. More recent studies, however, have shown that brain regions interact nonlinearly, and thus these approaches may not be optimal (Breakspear, 2017; Ocker et al., 2017; Wang et al., 2019).

Measures of cognitive ability and psychiatric symptoms, like imaging data, can be high dimensional and have often been analysed as general factors to reduce dimensionality (Cronbach and Meehl, 1955). A general intelligence factor (g-factor) can be derived from a sufficiently broad domain of cognitive tasks, and g-factors from different sets of cognitive tasks have been shown to be highly correlated (Johnson et al., 2004, 2008). Likewise, researchers have explored the possibility of a single dimension (p-factor) to measure one's mental health (Caspi and Moffitt, 2018). Empirical evidence indicates a continuum of symptoms across multiple mental illnesses and an overlap of symptoms across mental disorders (Kotov et al., 2017), with some disorders sharing the same set of risk factors and biomarkers (Goodkind et al., 2015; Pinto et al., 2017). While molecular levels of analysis support the g-factor as a unitary construct, less molecular evidence provides support for the unitary nature of the p-factor (de la Fuente et al., 2021; Grotzinger et al., 2021, 2020). Rather, the p-factor may represent an emergent phenomenon, arising from the aggregation of varied mechanisms present across subsets of psychiatric conditions. Studies have found that both the g-factor and the p-factor are highly heritable in adulthood (Grotzinger et al., 2019; Harden et al., 2020), and associated with variation in the human brain connectome (Allegrini et al., 2020; Lund et al., 2020; Madole et al., 2021).

Rapid growth in availability of parallel computing in the recent years has enabled application of deep learning (DL). DL is very effective in extracting latent features and non-linear patterns from complex data (Vieira et al., 2017). A 3D convolutional neural network (CNN) approach has been used to investigate automated brain disease classification based on or voxel-based morphometric (VBM) features derived from structural MRI scans (Cole et al., 2017; Hu et al., 2020; Zou et al., 2017). Additionally, different graph-based neural network (GNN) models have been developed for predictive modelling based on connectivity matrices from diffusion MRI (dMRI) and resting state functional MRI (rs-fMRI) (Kawahara et al., 2017; Li and Duncan, 2020). In a previous study (Yeung et al., 2020), we have shown that the BrainNetCNN neural network architecture proposed by Kawahara et al. (Kawahara et al., 2017) was more appropriate for sex classification based on brain connectome adjacency matrices compared with a naive image-based CNN architecture. Although there are several theoretical merits in the application of DL methods for understanding the neurobiological correlates of important between-person differences, the use of promising cutting-edge DL methods and its quantitative benefits beyond more conventional statistical methods remain moot. He et al. reported that Kernel Ridge Regression achieved comparable results with DL models (He et al., 2020). Schulz et al. also showed that non-linear models did not outperform simple linear models in predicting common phenotypes from brain scans (Schulz et al., 2020). In contrast, Abrol et al. found in their study that DL models substantially outperformed standard ML methods (Abrol et al., 2021). Large sample sizes and appropriate image processing methods constitute a highly valuable test-bed in which to directly address the challenge of assessing the performance of DL methods in comparison to other methods.

In this study, we directly compare the BrainNetCNN model against more classical ML methods when applied to connectomes based on 6 different structural connectivity modalities, namely SC, FA, MD, OD, ICVF and ISOVF in the UK Biobank (UKB). Our main aims were i) to compare the performances on prediction of sex and age as benchmark tasks, as well as general cognition (g-factor) and mental health (p-factor) among the 6 different connectivity modalities with the BrainNetCNN models, ii) to investigate the how the BrainNetCNN model used the brain features for prediction through the use of Gradient Attribution Map, iii) to compare the BrainNetCNN DL model's performance and model's decisions with other classical

ML methods, using prediction of sex and age as benchmark and then extending to prediction of more complex phenotypes (cognition and psychopathology).

Results

Prediction performances for different tasks

We first applied the BrainNetCNN classification models to predict sex. Table 1a shows the prediction accuracies for sex for the different structural connectivity weightings. We found that SC attained the best performance (Accuracy = 86.91%) of all modalities, on average 4.1% clear of the next best, OD. We then applied the BrainNetCNN regression models to predict age. Table 1b shows the mean absolute error (MAE) and correlation between the raw predicted age and the true values for different connectivity weightings. Here the best performance was again obtained with SC (MAE = 4.245 years).

We also assessed prediction of the g-factor and the p-factor. Table 1c shows the results for g-factor. Cognition and psychopathology being complex phenotypes are more subtly reflected in brain phenotypes compared to age, and hence we expected much larger prediction errors for these variables. This was indeed the case with correlation between the raw predicted g-factor and the true values being $60 - 75\%$ lower compared to those observed for age prediction. Again, SC was the modality with best performance (g-factor: MAE = 0.780, Correlation = 0.201, p-factor: MAE = 0.790, Correlation = 0.143). Table 1d shows the results for p-factor, where the correlations were $80 - 85\%$ lower compared to those observed for age prediction.

We then investigated whether addition of age and sex covariates to the DL model improved predictive performance. Age and sex were both significantly associated with the g-factor and p-factor, and we considered that if these covariates were combined with the brain connectivity matrix, this would improve predictive performance of the DL model. Indeed, the DL model with additional covariates was able to achieve better performance for g-factor where all modalities achieved correlations around 0.238 - 0.251 (Table 2a). The additional covariates also improved prediction of the p-factor, with all modalities achieving correlations around 0.202 - 0.245 (please see Table 2b). Moreover, when trained with additional covariates, models based on all modalities achieved comparable performances on predicting g-factor and p-factor.

The training progress plots for the BrainNetCNN models are presented in section B.3.4 of the Supplementary Materials.

Mapping predictive streamline count and fractional anisotropy features

Since SC and FA are the most common weights used in previous studies, the remainder of the analysis focuses on the results obtained from prediction models based on the SC and FA inputs. We computed the gradient attribution maps for the DL models to investigate how the deep neural networks made their predictive decisions, and to estimate ROI predictive powers, see Methods. Figure 1 and 2 show the circos plots of gradient maps for all the prediction tasks based respectively on the SC and FA connectomes.

Sex classification gradient maps

According to the SC gradient attribution map for sex classification (SexGradient), many connections from right medial orbital-frontal, right inferior temporal regions and right hippocampus are predictive of being female (Figure 1a). Conversely,

we saw that many connections from left superior temporal, right accumbens areas and left ventral diencephalon regions were chosen as being predictive of being male. In general, the precuneus, right superior temporal region , right insula, right putamen and right thalamus were considered by the model as being hubs of important connections for sex classification.

According to SexGradient based on FA, many connections from the frontal pole, right inferior temporal and left occipital regions were highlighted as being predictive of being female (Figure 2a). On the other hand, we saw that many connections from right accumbens, right putamen and left entorhinal were predictive of being male. In general, right thalamus, right caudate, right putamen, left isthmus cingulate and left superior frontal regions were considered by the model as being hubs of important connections for sex classification.

Age prediction gradient maps

According to the gradient attribution map for age prediction (AgeGradient) based on SC, many connections from pallidum, temporal pole and medial orbital frontal regions are predictive of being older, while connections from hippocampus, right ventral diencephalon and right transverse temporal regions are predictive of being younger (Figure 1b). In general, the right thalamus, right accumbens areas, right ventral diencephalon, right medial orbital frontal and right precuneus were considered by the model as being hubs of important connections for age prediction.

According to AgeGradient based on FA, many connections from accumbens, isthmus cingulate and right insula regions are predictive of being older, while connections from pars opercularis, pars traingularis, superior frontal and left lateral orbital frontal were predictive of being younger (Figure 2b). In general, the thalamus, putamen, right caudate, left ventral diencephalon and left isthmus cingulate regions were considered by the model as being hubs of important connections for age prediction.

g-factor prediction gradient maps

It was found that adding covariates to the DL model helped to reduce confounding effects as seen in classical methods. The correlation between gradient attribution map for g-factor prediction without covariates (GGradient) and AgeGradient was -0.2722, while correlation between gradient attribution map for g-factor prediction with covariates (GCovGradient) and AgeGradient was -0.0996, averaged across cross-validation folds. We hypothesised that the addition of age and sex covariates to the model would have accounted for the effects of these covariates on the g-factor and we here therefore report the g-factor gradients for the DL model which includes these covariates. We note that according to GCovGradient based on SC, many connections from thalamus, left superior temporal and right parahippocampal regions were chosen as being predictive of higher g-factor, while most connections from medial orbital frontal, right amygdala and left lateral orbital frontal regions were predictive of lower g-factor (Figure 1c). In general, right thalamus, left putamen, left hippocampus, right superior parietal and right precuneus regions were considered by the model as being hubs of important connections for predicting cognitive g-factor.

Adding covariates had similar effects for the models with FA as what we saw for SC. The correlation between GGradient and AgeGradient was -0.2581 while correlation between GCovGradient and AgeGradient was -0.0852, averaged across folds. According to GCovGradient based on FA, many connections from the left lingual, left rostral anterior cingulate and right entorhinal regions are predictive of higher cognitive g-factor, while most connections from left temporal pole, right caudal anterior cingulate and left pallidum regions are predictive of lower cognitive g-factor (Figure 2c). In general, ventral

diencephalon, right insula, right inferior temporal, right caudate and right isthumus cingulate regions were considered by the model as being hubs of important connections for predicting cognitive g-factor.

p-factor prediction gradient maps

According to gradient attribution map for p-factor prediction with covariates (PCovGradient) based on SC, many connections from right middle temporal, left entorhinal and left superior frontal regions were chosen as being predictive of higher p-factor score (higher p-factor representing higher psychopathology, while most connections from left pericalcarine, left superior temporal and right caudate regions were predictive of lower p-factor score (Figure 1d). In general, superior temporal, left precuneus, left putamen, right superior parietal and left lingual regions were considered by the model as being hubs of important connections for predicting p-factor score.

According to PCovGradient based on FA, many connections from left amygdala, left lateral orbital frontal and left pars triangularis regions were chosen as being predictive of higher p-factor score, while most connections from left insula, right pallidum and right amygdala regions were predictive of lower p-factor score (Figure 2d). In general, insula, left ventral diencephalon, left posterior cingulate, left middle temporal and right inferior temporal regions were considered by the model as being hubs of important connections for predicting p-factor score.

Interpretation of edge importance for each prediction task

In order to verify if small subsets of edges were predominantly responsible for performance at each prediction task, kurtosis measures of the edge-gradient distributions were examined to evaluate whether the probability mass was concentrated around the mean. This way we assessed the gradients for age, g-factor and p-factor prediction. It was found that the kurtoses of gradients for age, g-factor and p-factor predictions based on SC were respectively 43.67, 10.93, and 11.50 (Figure 3a), and were respectively 19.24, 4.36 and 5.16 based on FA (Figure 3b). This implies that there were more edges having gradients close to the mean (zero) for age prediction compared to the other two prediction tasks. This indicates that the model mainly relied on a smaller subset of edges for age prediction, but on wider ranges of edges for g-factor and p-factor predictions. Figure 3 shows the histograms of gradients for each prediction task based on SC and FA modalities.

Comparing prediction performance with classical machine learning approaches

In addition to BainNetCNN, we also applied Ridge Regression, LASSO Regression, linear SVM and Kernel Ridge Regression (KRR) to predict sex, age, g-factor and p-factor. Without covariates, all four alternative prediction algorithms (Ridge Regression, LASSO Regression, linear SVM and KRR) performed the best with SC data on most prediction tasks with some exceptions (Ridge Regression and KRR performed best on ISOVF weights for age prediction). When predicting g-factor and p-factor with additional covariates, linear methods (Ridge Regression, LASSO Regression, linear SVM) performed better with FA and OD rather than SC data, while KRR performed the best with SC data. The full results are included in section B.2 of the Supplementary Materials. Overall, we found performance of the classical ML methods to be comparable to that of the BrainNetCNN. Tables 3a - 3d show the predictive performance of the five models respectively on sex classification, age prediction, as well as g-factor and p-factor predictions (both with age and sex covariates added), based on streamline counts.

For Ridge Regression, LASSO Regression and linear SVM models we additionally extracted feature beta weights in order to compare them with the BrainNetCNN gradient attribution maps. This was not possible for the KRR model due to non-linear transformation of the feature space.

Consistency of the Gradients/Betas across folds as well as across different models

Beta weights in classical ML models were generally more stable across folds compared to gradients from the DL models, likely due to the linear nature of the classical approaches. The betasfor the ML models from the three classical methods achieved correlations of > 0.68 across folds, for all prediction tasks and for each of the network weightings. The betas were generally more stable for sex classification and age prediction compared to prediction of g-factor or p-factor. For DL gradients, we saw that the gradients for age and sex prediction were consistent across folds (correlations > 0.74). The gradients were however not as consistent for g-factor prediction (correlations of 0.480.6). Details of across-fold gradient and beta weight correlations are presented in section B.3.1 of the Supplementary Materials.

Regarding the consistency of the DL gradients with classical ML betas, the betas from linear ridge regression were the most consistent with the gradient attribution maps from BrainNetCNN across different prediction tasks. Spearman correlations between the betas and the gradient attribution maps based on SC and FA were $0.66 - 0.77$ and $0.49 - 0.85$ respectively for different prediction tasks, see section B.3.2 of the Supplementary Materials.

The top brain regions predictive of g-factor and p-factor within BrainNetCNN and within linear Ridge Regression had a high percentage ($\geq 60\%$) of overlap, within both SC and FA modalities (see section B.3.3 in the Supplementary Materials). This implies that the deep learning models and the linear ridge regression models rated features in highly similar ways.

We additionally found that inclusion of covariates in linear models had similar effects on the betas as on the gradients in BrainNetCNN, when both models were based on SC matrices. Without the included covariates, the betas for g-factor prediction (GBetas) and the betas for age prediction (AgeBetas) had average correlation of -0.2592. With covariates, the betas for g-factor (GCovBetas) and AgeBetas had average correlation of -0.0527.

Discussion

In the current study, we applied the BrainNetCNN DL model to connectivity matrices derived from six different structural connectivity modalities, MD, FA, SC and three other novel measures OD, ISOVF and ICVF, to predict sex, age, cognitive g-factor and mental health p-factor using one of the largest brain structural connectome datasets currently available, the UKB. All connectome processing and network construction steps were performed locally on diffusion and structural MRI data. It was found that the best predictions were generally achieved with SC matrices across all models, though these are likely more susceptible to the confounding effects of age and sex, probably via head size. Adding age and sex as external regressors to the BrainNetCNN model helped improve the prediction of both g-factor and p-factor. This is consistent with the previous studies, which have shown that adding relevant covariates to convolutional neural network models may lead to better predictive performance (Yamamoto et al., 2020). In our study we found that addition of external regressors (i.e. the covariates) had similar effects on the performance of the DL model, as it had on the performance of the classical linear regression models. A recommendation which follows from our results is that researchers should consider adding clinical and

demographic covariates to DL models to improve prediction results.

When comparing BrainNetCNN to the classical ML methods, we found similar performance across all prediction tasks. The KRR model did not outperform the other three linear predictive models. This implies that models which simply consider ROI-to-ROI connections individually and linearly can match state-of-the-art machine learning models when applied to structural connectomes. We further found high correlations between gradient maps from DL models and beta maps from linear models, which indicates that BrainNetCNN estimated the connectivity graph node and edge importance weights in similar ways as linear prediction models. Moreover, inclusion of age and sex covariates had similar effects in both BrainNetCNN and linear models. This potentially implies that there is a lack of exploitable non-linearity in brain scans for phenotype prediction, as previously suggested in Schulz et al. (Schulz et al., 2020). Taken together, these results indicate that most of the variation between phenotypes in weighted structural connectomes is linear rather than non-linear in nature.

In the current study we found that the models based on SC measures had the best predictive performance compared to other network weightings. This is in line with the results of the previous studies which showed that use of SC measures led to better performance when predicting Autism Spectrum Disorders and paediatric Traumatic Brain Injury (Payabvash et al., 2019; Raji et al., 2020). SC measures might overall be more sensitive to phenotype and psychopathology-related differences in brain connectivity (e.g. (Oestreich et al., 2019)). We note that the entries in SC matrices followed log normal distributions, while entries in other types of connectivity matrices followed normal distributions, which might have been one reason for superior predictive performance with SC matrices. Another possibility, as suggested in Buchanan et al. (Buchanan et al., 2020), could be that the confounding effect of age and sex via head size on SC weightings was larger than on other types of network weightings.

As expected, we found prediction of g-factor ($r = 0.249$) and the p-factor ($r = 0.234$) to be less accurate than prediction of age and sex. Predicting complex cognitive phenotypes from structural neuroimaging is a difficult task. Existing studies conducted on large community samples found that structural neuroimaging measures tend to account for only modest proportions of variance in intelligence (Cox et al., 2019b; Deary et al., 2021). Prediction of case-control status for specific mental illnesses or modelling of mental health continuum has been a similarly difficult task. While some recent studies have had some success in predicting anxiety (Greening and Mitchell, 2015; Wang et al., 2021) and depression symptom severity (Yu et al., 2021) with functional and structural connectomes (achieving correlations > 0.24 between the predicted score and the actual score), they generally had relatively small sample sizes with replication yet to be achieved in larger cohorts. Our study extends these results and suggests that brain regions and connections are differentially important for prediction of behavioural and mental health phenotypes such as the g-factor and p-factor (Cox et al., 2019b; Dubois et al., 2018).

Our analysis of SC gradient maps revealed the putamen, right precuneus, right thalamus and left superior temporal regions as hubs of important connections for all prediction tasks. For FA gradient maps, we found that the right isthmus cingulate, right ventral diencephalon and right insula were selected as hubs of important connections for all prediction tasks. The relative importance of the regions shown here in the context of prediction modelling should not be interpreted from a functional perspective. In other words, the estimated relative region importance for g-factor prediction must not be taken to indicate importance of that region for e.g. the cognitive processes underpinning the g-factor. This is because the collinearity among regional brain properties results in the selection of some predictors that carry global information, with other predictors being used for fine tuning of the prediction model. The low-to-medium correlations between the beta coefficients (from ML

models) and gradient maps (from DL models) lends some credence to such hypothesis.

Some limitations of the current study should be mentioned. First, it is possible that the representation of structural connectivity in the form of adjacency matrices might have limited the deep learning model's potential to utilise the latent non-linear properties of the brain networks, thus leading to suboptimal performance on prediction of the g-factor and p-factor. Another possibility, as described in both He et al. and Schulz et al., is that a much larger sample size may be needed for DL models to show superiority over classical ML methods (He et al., 2020; Schulz et al., 2020). BrainNetCNN has novel layers designed for preserving the topology of adjacency matrices. However, the setup of the E2E and E2N layers only allowed aggregating values in the 1-hop neighbourhood (primary neighbours) and this might have limited its ability to explore long-range connections.This could have been one of the reasons for similar decision mechanisms between the DL and the linear methods. A simple adjustment could be to alter the network architecture by stacking a number (e.g. k) of E2E layers, which would enable the model to explore brain region connections to k-hop rather than 1-hop neighbours. In this case, however, the simpler and shorter-range connections (i.e. 1-hop to (k1)-hop neighbourhood) may be ignored and not used for prediction in the top layers. To overcome these limitations, future studies could modify the BrainNetCNN architecture according to the idea of residual connections in deep neural networks (He et al., 2016) – which would enable the model to take into account both direct and indirect longer-range connections between brain regions.

The relatively poor performance of the models on predicting the p-factor may have been in part due to the overall validity of the p-factor as a construct to capture general psychopathology. The p-factor is not well supported as unitary dimension at molecular levels of analysis (Grotzinger et al., 2020). Moreover, some studies concluded that a single model does not provide the best fit to capture general psychopathology and that a 3-factor model (Vollebergh et al., 2001) or a bifactor model (Caspi et al., 2014; Lahey et al., 2012) may be better. Some studies suggest that a robust p-factor is yet to be defined (Cervin et al., 2020). The challenges in defining the p-factor might arise from the absence of a positive manifold in mental health scores (van Bork et al., 2017). Nonetheless, other studies do support existence of a general dimension to capture psychopathology (Allegrini et al., 2020; Lund et al., 2020), and hence we evaluated prediction of the p-factor alongside the g-factor in the current study.

A further, more general limitation of our study is that there is not yet an agreed methodology for constructing structural brain networks from dMRI data (Qi et al., 2015). Connectomic measures are known to be sensitive to the network construction methodology, and different predictive modelling results might be achieved with network construction approaches other than those applied here. dMRI itself only enables noisy and indirect measurement of water molecule diffusion, and thus faithful reconstruction of brain connectivity remains challenging (Buchanan et al., 2020; Jones et al., 2013). These limitations will be addressed in the future with further advances in both structural imaging and tractography methods.

To summarize, the BrainNetCNN DL model did not perform substantially better compared to classical ML approaches when applied to predict age, sex, g-factor or p-factor from connectomic data.Comparison of gradient maps from DL models and beta maps from linear ML models suggested that these methods used similar predictive decision mechanisms. Results also showed that the BrainNetCNN model treated the additional covariates (age and sex) in a similar way as did the classical linear regression models. Overall, results of our study imply that additional model complexity may not improve prediction of complex phenotypes from structural connectomic data.

Materials and Methods

Materials

Participants were recruited and brain imaging was completed as part of the UKB study. The UKB field IDs for the cognitive tasks and mental health questionnaire (MHQ) items used in this study are presented in section A.1 of the Supplementary Material. Details of the UKB data can be found at `https://biobank.ndph.ox.ac.uk/showcase/label.cgi?id=100000`. As a brief summary, four cognitive tasks were chosen for deriving the g-factor and 14 MHQ items were chosen for derivation of the p-factor.

Participants

A subset of the UKB participants underwent brain MRI at the UKB imaging centre in Cheadle, Manchester, UK. $N = 9,858$ participants with compatible T1-weighted and dMRI data were collected from the UKB for which we derived connectomes locally, as described below.

MRI Acquisition

All imaging data were acquired using a single Siemens Skyra 3T scanner. 3D T1-weighted volumes were acquired using a MP RAGE sequence at $1 \times 1 \times 1$ mm resolution with $208 \times 256 \times 256$ field of view (FOV). The dMRI data were acquired using a spin-echo EPI sequence (50 $b = 1000s/mm^2$, 50 $b = 2000s/mm^2$ and 10 $b = 0s/mm^2$) resulting in 100 distinct diffusion-encoding directions, FOV $= 104 \times 104$ mm, imaging matrix $= 52 \times 52$, 72 slices, slice thickness $= 2$ mm. $N = 831$ ($\sim 8.4\%$ of total N) participants with missing dMRI data or processing failure were excluded from the analysis. Details of the MRI protocol and imaging data processing can be found in (Alfaro-Almagro et al., 2018; Miller et al., 2016).

Network Construction

The network data and methods of network construction used in this study have been published previously (Buchanan et al., 2020; Madole et al., 2021) and are outlined below. All network construction steps were performed locally on MRI data. Each T1-weighted image was segmented into 85 distinct neuroanatomical Regions-Of-Interest (ROI) using volumetric segmentation and cortical reconstruction (FreeSurfer v5.3.0), 34 cortical structures per hemisphere were identified using the Desikan-Killany atlas (Desikan et al., 2006). Brain stem, accumbens area, amygdala, caudate nucleus, hippocampus, pallidum, putamen, thalamus and ventral diencephalon were also extracted with FreeSurfer. Outputs were visually inspected and those which failed to meet quality control (QC) standards were removed (Cox et al., 2019a). $N = 842$ ($\sim 8.5\%$ of total N) participants with incomplete FreeSurfer output or processing failure were excluded from the analysis. A cross-modal nonlinear registration method was used to align ROIs from T1-weighted volume to diffusion space (skull stripping (Smith, 2002), initial alignment by affine transformation with 12 degrees of freedom (FLIRT; (Jenkinson and Smith, 2001)) followed by a nonlinear deformation method (FNIRT; (Andersson et al., 2007))).

Networks were constructed by identifying connections between all ROI pairs. The endpoint of a streamline was recorded as the first ROI encountered when tracking from the seed location. Successful connections were recorded in an 85×85 adjacency matrix. In total, six network weightings, FA, MD, SC, OD, ISOVF and ICVF, were computed. For each weighting,

an adjacency matrix was computed with each element, a_{ij}, recording the mean value of the diffusion parameter in voxels identified along all interconnecting streamlines between nodes i and j. All matrices were made symmetric since afferent and efferent connections are indistinguishable for tractography. Self-connections were removed, setting diagonal entries to zero (Buchanan et al., 2020). $N = 2$ participants with disconnected connectivity matrices were excluded.

In total, $8,183$ participants (45.1–78.5 years of age, 3,869 male) remained after participants were excluded at QC stage or due to failure in processing. Although it was reported in Madole et al. that $N = 157$ participants ($< 2\%$) might have dementias and neurological syndromes, for example stroke and brain injuries, they found that excluding these participants from the sample did not result in significant differences in their primary outcome measures (Madole et al., 2021). We therefore kept the full sample for the following analysis. On average, 6.01 million streamlines were seeded per subject of which 1.49 million (24.9%) were found to successfully connect between nodes following the tracking procedure and removal of self-connections. For each participant, networks were produced from the same set of streamlines, where the range of values for MD is 0 - 0.003 $\times 10^{-3} mm^2/s$, for FA is 0 - 0.9, for SC is 0 - 4.03 $\times 10^4$, for OD is 0 - 0.9, for ISOVF is 0 - 1 and for ICVF is 0 - 1. Before any thresholding was introduced, the mean value of network density (percentage of non-zero entries in an adjacency matrix) across subjects was 68.1% (SD = 3.1). Proportional-thresholding was used to keep only connections present in at least 2/3 of subjects, which resulted in connection density of $\sim$60% after thresholding.

Cognitive Tasks and g-factor

The cognitive tasks included the Verbal Numerical Reasoning (VNR), Reaction Time (RT, log-transformed), Pairs Matching (Pairs Match, log (x+1) transformed) and Prospective Memory (Fawns-Ritchie and Deary, 2020; Lyall et al., 2016) tests. Table 4 shows participants characteristics, demographic information and cognitive task scores. Description of the cognitive tasks' scores is presented in section A.1.1 of the Supplementary Materials.

Principal component analysis (PCA) implemented in *pcacov* function in MATLAB 2020a was applied to the 4 cognitive tasks as an intermediate step to derive the cognitive g-factor value for each participant. The principal component score, g-factor, was computed as the sum of z-normalised cognitive task scores multiplied by the respective first unrotated component coefficients from *pcacov*. The g-factor explained 34.84% of the variance. To avoid data leakage, the principal component coefficients were computed with exclusion of the test data (see below). Excluding the missing values and test data in the cognitive tasks, 6558 samples were available for computing the principal component coefficient. The g-factor was z-normalised based on the training data. The resulting g-factor had a range [-4.3218, 2.7983]. The standardised loadings of scores for each task on the g-factor are shown in Table 5.

p-factor

p-factor derivation followed the same principles as described in (Lund et al., 2020). Lund et al. applied principal component analysis (PCA) to the 14 mental health questionnaire (MHQ) items from the UKB which measured common forms of depression, anxiety, psychotic experiences and substance abuse, and took the first unrotated principal component score as the p-factor. In this study, the same set of MHQ items was used to derive the p-factor. Out of 8,183 participants, 6,247 participants had fewer than 3 missing values on these 14 MHQ items. The missing values were imputed in R with the missMDA package (Josse et al., 2016).

Since all MHQ items were categorical variables, the polychoric correlation was more appropriate than Pearson's correlation for application of PCA. We applied PCA on polychoric correlation of the 14 items and took the first unrotated principal component score as the p-factor. A comparison of the two PCA methods (using polychoric and Pearson correlation) is presented in section A.2.1 of the Supplementary Materials. We checked the variance explained, factor congruence between the two methods, and assessed correlations between the principal component scores derived using the two methods. This motivated selection of polychoric correlation for the PCA. The derived p-factor explained 49.60% for the variance. Figure 4 shows the loadings of the individual MHQ items on the principal components. We saw strong positive loadings of depressive/anxiety items on the first PC and positive loadings of psychotic experiences items on the second PC, which is similar to the results reported in (Lund et al., 2020). It was also found that the p-factor was not significantly associated with the g-factor (Pearson's Correlation = 0.0025, p-value = 0.8483).

Methods

BrainNetCNN Components

Kawahara et. al. provide clear formulations of the special layers for the BrainNetCNN (Kawahara et al., 2017). Briefly, the layers user in the network are following:

1. Edge-to-Edge (E2E) layer: This is a convolution layer with a cross-shaped filter where the (i, j) entry output is given by the weighted sum of the i-th row and weighted sum of the j-th column. This can be written as:

$$(A^{k,n})_{ij} = \sum_{f=1}^{F_{k-1}} A_{i,:}^{k-1} \cdot w_r^{k-1,f,n} + A_{:,j}^{k-1} \cdot w_c^{k-1,f,n} \tag{0.1}$$

   where $A^{k,n}$ is the filtered adjacency matrix, $A_{i,:}$ is the i-th row of A and $A_{:,j}$ is the j-th column of A, and $w_r^{k-1,f,n}$ and $w_c^{k-1,f,n}$ are the learnt row and column weights for the n-th filter at the $k-1$-th layer, respectively. F_{k-1} is the number of feature maps at the $k-1$-th layer.

2. Edge-to-Node (E2N) layer: An E2N layer is equivalent to adding a spatial 1D convolutional row filter of the adjacency matrix to a transposed spatial 1D convolutional column filter of the adjacency matrix. Kawahara et al. (Kawahara et al., 2017), showed that combining the row and column filters does not result in improvement of model performance when compared to using only the row filter, and hence only the row filter was applied in the current study.

Network Architecture

Architecture without covariates The architecture of the deep neural network is presented in Figure 5a. This is a simpler version of the original architecture described in (Kawahara et al., 2017). We were able to significantly reduce the number of layers and filters while maintaining prediction performance. Moreover, either increasing the number of filters per layer or the number of hidden layers made the overfitting more severe.

Architecture with covariates The g-factor and p-factor are known to be significantly associated with age and sex, and hence consideration of these factors in the predictive model can improve its performance. To take into account age and sex

covariates, the model was update with additional layers. Specifically, a fully connected layer with 30 nodes was added after the E2N layer, followed by a concatenation layer to include the covariates. Figure 5b shows the architecture of the updated model.

Network Input

The SC matrices were first made symmetric by addition of the transpose and division by 2. We then performed maximum value normalisation on the symmetric SC matrices. The maximum value normalisation was performed on all modalities as we previously found this to be the most effective normalisation (Yeung et al., 2020). Maximum value normalisation is formulated as follows:

$$A_{max} = \frac{A}{\max\limits_{1 \leq i,j \leq 85} A_{ij}} \tag{0.2}$$

Since it is known that the head size has great influence on the streamline counts , we computed the correlation between the intracranial volume (ICV) and the total sum of entries in the SC matrices before and after normalisation. We found that the correlation was 0.7405 ($\beta = 0.7988$, $p < 1 \times 10^{-324}$) before normalisation and -0.0701 ($\beta = -1.023 \times 10^{-5}$, $p = 3.784 \times 10^{-11}$) after normalisation. Therefore, we saw that the maximum value normalisation was able to largely remove the effect of ICV. Figure 6 shows the mean connectivity matrices computed across all participants for different weightings and histograms of edge weights pooled across participants. The SC weights were shown in log-scale. The five network weightings except for SC had similar distribution shapes. MD, FA, OD and ICVF had skewness in the range [-0.3, 0.3]. ISOVF had a skewness of 1.36 and SC has a skewness of 8.01.

Gradient Attribution Map

Gradient Attribution Map were computed to check which connectivity matrix entries influenced prediction the most (Simonyan et al., 2013). This is similar to computing the partial derivative of the class score with respect to each entry of the A^n matrix, where the magnitudes of the derivatives represent the degrees of influence on the final class score, y^n, for some participant n.

Due to the stochastic nature of the neural network training process, the gradient maps may be different across different training runs and across different parameter initialisations. In our study the BrainNetCNN models were therefore trained 20 times for each cross-validation (CV) partition with randomised parameter initialisations. Since we used 5-fold cross validation, there were 100 BrainNetCNN models trained in total. We then took the average of the gradient maps from the 100 models for each modality and each prediction task.

Since A^n is symmetric, we expected that $\frac{\partial y^n}{\partial A^n_{ij}} = \frac{\partial y^n}{\partial A^n_{ji}}$. However, in reality, since the column weights and row weights within the filters were trained independently, it was found that not only A^n_{ij} and A^n_{ji} had different magnitudes, but some of the pairs even had opposing signs. To solve this issue, we summed each gradient map with its transpose, thus getting a single gradient estimate for each ROI-to-ROI connection.

Predictive power of each node (ROI) was defined as the sum of the gradients of the node's edges. This enabled determining whether the node was generally predictive of higher or lower outcome values.Importance of each node was defined as the sum of the absolute values of the node edges' gradients This was used to determine whether a node was a hub of important

connections for prediction. We report the top regions relevant for each prediction task based on these two measures (predictive power and importance) in section B.3.3 of the supplementary material. We should note that our assessment of node importance was purely based on the magnitudes and directions of the gradient map entries. We thus do not claim that the regions defined as important for prediction in this study were the only ones related to (or responsible for) the evaluated phenotypes.

Model Hyperparameters

For both models we used Adam optimiser (Kingma and Ba, 2014) for weight updates, with the same set of hyperparameters used in training: learning rate = 0.001, gradient decay factor = 0.9, squared gradient decay factor = 0.9, mini batch size = 128, L2-norm regularization = 0.0001, validation frequency = 50, training epoch = 200. The training stopping points corresponded to the best validation accuracy.

Experimental Set-up

The participants were first sorted according to their participant ID in ascending order. The test set consisted of the first 15% ($\sim$ 1230 participants for prediction of sex, age and cognitive functioning, $\sim$ 940 participants for p-factor) of the total sample. 5-fold CV was performed on the rest of the data. The same set of CV fold splits was used across the imaging modalities to enable direct comparisons. Each of the models was run for 200 epochs and the weights at each epoch were recorded. The weights that yielded the best validation accuracy were taken as the optimal weights for the fold and used for performance evaluation on the test data. Test accuracies were averaged across the folds to get performance estimates.

Classical Machine Learning Approaches

We compared the DL models with linear versions of three common classical methods: Ridge Regression, LASSO Regression, and linear SVM. We also applied Kernel Ridge Regression (KRR). Multiple studies suggest that KRR may have comparable performances with DL methods (He et al., 2020; Mihalik et al., 2019; Schulz et al., 2020). Input features to these models were the individual ROI-to-ROI connections specified in the normalised adjacency matrices (i.e. the upper/lower triangular entries of the matrices).

In addition to the predictive analyses based on the connectomic data we also tested using only age and sex to predict the g-factor and p-factor. The models were underfit and we therefore did not examine them further.

Ridge Regression Ridge regression is very similar to ordinary least squares regression except for the fact that it has an extra regularisation term:

$$Y = X\beta + e \tag{0.3}$$

where Y is the response variable, X is the input feature vectors, β are the learnt feature weights, subjected to the following cost function,

$$\min_{\beta} \text{square loss} + \lambda\beta^T\beta$$

where square loss $= (Y - X\beta)^T(Y - X\beta)$. The L_2 regularisation term $(\lambda\beta^T\beta)$ is added to prevent overfitting. For logistic regression, the equation is the following,

$$\ln \frac{p}{1-p} = X\beta + e$$

where $p = P(Y = 1)$, and square loss is replaced by logistic loss:

$$\text{logistic loss} = -\sum_{i=1}^{N} \left[-\ln\left(1 + e^{X_{i,:}\beta}\right) + Y_i(X_{i,:}\beta)\right]$$

where N is the number of observation, $X_{i,:}$ is the feature vector for observation i, Y_i is the response for observation i.

LASSO Regression LASSO regression is similar to Ridge regression except that the L2 regularisation term is replaced by L1 regularisation term, with the same regression equation, but subject to a different constraint:

$$\min_{\beta} \text{square loss} + \lambda \|\beta\|_1$$

Similarly to logistic Ridge regression, square loss is replaced by logistic loss in logistic LASSO regression.

Support Vector Machine SVM is one of the most popular ML techniques applied in the neuroimaging literature. For classification, SVM constructs hyperplanes between two classes such that the distance between them is as large as possible. The larger the hyperplane margin between classes, the lower the out-of-sample error. SVM regression (SVM-R) is similar to the "soft margin" concept in SVM classification. More details can be found in (Boser et al., 1992) for support vector classification and (Drucker et al., 1997; Ho and Lin, 2012; Vapnik et al., 1997) for support vector regression. Linear kernel was used in this study. The epsilon parameter for linear SVM-R was estimated using the interquartile range of response variable Y.

Kernel Ridge Regression Kernel regression is a non-parametric ML algorithm (Murphy, 2012). Instead of taking the original features as predictors, it takes the subject pairwise-similarities as the predictors for the ML model. Let $X_{train} = \{x_1, x_2, \cdots, x_n\}$ be the set of feature vectors and $Y_{train} = \{y_1, y_2, \cdots, y_n\}$ be the set of responses in the training set. For a test subject x', y' is estimated through the following equation:

$$y' = \sum_{i=1}^{n} \beta_i K(x_i, x') y_i$$

where the $K(x_i, x')$ is the similarity between x' and each feature vector x_i in the training set. β_i are the learnt feature weights, subjected to the cost function similar to the one stated in the section of *Ridge Regression*. With this setup, the dimension is basically equal to the number of observations in the training set in each fold. Similar to He et al., the similarity $K(\cdot, \cdot)$ was chosen to be the Pearson's correlation between the upper triangular entries of matrices (He et al., 2020).

Finding optimial λ Regularisation is needed for such high dimensional data. In order to find the optimal λ, we tested different values ranging from 1e-6 to 1e-1. The λ that optimized the validation accuracies/correlation was chosen, so the λs

are different for each fold, each model and each connectivity weight.

Acknowledgements

This study was supported by Wellcome Trust awards (References 104036/Z/14/Z; 220857/Z/20/Z), and was also supported by National Institutes of Health (NIH) research grant R01AG054628 which supported CRB, EMT, MEB and SRC. The research was conducted using the UK Biobank resource, with approved project number 10279. Structural brain imaging data from UK Biobank was processed using facilities within the Lothian Birth Cohort group at the University of Edinburgh, which is supported by Age UK (as The Disconnected Mind project), the Medical Research Council (MR/R024065/1), and the University of Edinburgh. This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). The Population Research Center (PRC) and Center on Aging and Population Sciences (CAPS) at The University of Texas at Austin are supported by National Institutes of Health (NIH) grants P2CHD042849 and P30AG066614, respectively. KMS was supported by Health Data Research UK, an initiative funded by UK Research and Innovation Councils, NIH Research (England) and the UK devolved administrations, and leading medical research charities. SRC was also supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (221890/Z/20/Z). AMM and HCW are additionally supported by a UKRI award (Reference MC_PC_17209).

References

Abrol, A., Fu, Z., Salman, M., Silva, R., Du, Y., Plis, S., and Calhoun, V. (2021). Deep learning encodes robust discriminative neuroimaging representations to outperform standard machine learning. *Nature communications*, 12(1):1–17.

Agosta, F., Galantucci, S., Riva, N., Chiò, A., Messina, S., Iannaccone, S., Calvo, A., Silani, V., Copetti, M., Falini, A., et al. (2014). Intrahemispheric and interhemispheric structural network abnormalities in PLS and ALS. *Human brain mapping*, 35(4):1710–1722.

Alexander, A. L., Lee, J. E., Lazar, M., and Field, A. S. (2007). Diffusion tensor imaging of the brain. *Neurotherapeutics*, 4(3):316–329.

Alfaro-Almagro, F. et al. (2018). Image processing and quality control for the first 10,000 brain imaging datasets from UK Biobank. *Neuroimage*, 166:400–424.

Allegrini, A. G., Cheesman, R., Rimfeld, K., Selzam, S., Pingault, J.-B., Eley, T. C., and Plomin, R. (2020). The p factor: genetic analyses support a general dimension of psychopathology in childhood and adolescence. *Journal of child psychology and psychiatry*, 61(1):30–39.

Andersson, J. L. et al. (2007). Non-linear registration aka Spatial normalisation FMRIB Technial Report TR07JA2. *FMRIB Analysis Group of the University of Oxford*.

Basser, P. J. and Pierpaoli, C. (2011). Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *Journal of magnetic resonance*, 213(2):560–570.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Breakspear, M. (2017). Dynamic models of large-scale brain activity. *Nature neuroscience*, 20(3):340–352.

Buchanan, C. R. et al. (2020). The effect of network thresholding and weighting on structural brain networks in the UK Biobank. *NeuroImage*, page 116443.

Bullmore, E. and Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews neuroscience*, 10(3):186–198.

Caspi, A., Houts, R. M., Belsky, D. W., Goldman-Mellor, S. J., Harrington, H., Israel, S., Meier, M. H., Ramrakha, S., Shalev, I., Poulton, R., et al. (2014). The p factor: one general psychopathology factor in the structure of psychiatric disorders? *Clinical psychological science*, 2(2):119–137.

Caspi, A. and Moffitt, T. E. (2018). All for one and one for all: Mental disorders in one dimension. *American Journal of Psychiatry*, 175(9):831–844.

Cervin, M., Norris, L. A., Ginsburg, G., Gosch, E. A., Compton, S. N., Piacentini, J., Albano, A. M., Sakolsky, D., Birmaher, B., Keeton, C., et al. (2020). The p Factor Consistently Predicts Long-Term Psychiatric and Functional Outcomes in Anxiety-Disordered Youth. *Journal of the American Academy of Child & Adolescent Psychiatry*.

Chen, V. C.-H., Shen, C.-Y., Liang, S. H.-Y., Li, Z.-H., Tyan, Y.-S., Liao, Y.-T., Huang, Y.-C., Lee, Y., McIntyre, R. S., and Weng, J.-C. (2016). Assessment of abnormal brain structures and networks in major depressive disorder using morphometric and connectome analyses. *Journal of Affective Disorders*, 205:103–111.

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

Collin, G., Sporns, O., Mandl, R. C., and Van Den Heuvel, M. P. (2014). Structural and functional aspects relating to cost and benefit of rich club organization in the human cerebral cortex. *Cerebral cortex*, 24(9):2258–2267.

Cox, S. R., Lyall, D. M., Ritchie, S. J., Bastin, M. E., Harris, M. A., Buchanan, C. R., Fawns-Ritchie, C., Barbu, M. C., De Nooij, L., Reus, L. M., et al. (2019a). Associations between vascular risk factors and brain MRI indices in UK Biobank. *European heart journal*, 40(28):2290–2300.

Cox, S. R., Ritchie, S. J., Fawns-Ritchie, C., Tucker-Drob, E. M., and Deary, I. J. (2019b). Structural brain imaging correlates of general intelligence in uk biobank. *Intelligence*, 76:101376.

Cronbach, L. J. and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

de la Fuente, J., Davies, G., Grotzinger, A. D., Tucker-Drob, E. M., and Deary, I. J. (2021). A general dimension of genetic sharing across diverse cognitive traits inferred from molecular data. *Nature Human Behaviour*, 5(1):49–58.

de Lange, S. C., Scholtens, L. H., van den Berg, L. H., Boks, M. P., Bozzali, M., Cahn, W., Dannlowski, U., Durston, S., Geuze, E., van Haren, N. E., et al. (2019). Shared vulnerability for connectome alterations across psychiatric and neurological brain disorders. *Nature human behaviour*, 3(9):988–998.

Deary, I. J., Cox, S. R., and Hill, W. D. (2021). Genetic variation, brain, and intelligence differences. *Molecular Psychiatry*, pages 1–19.

Desikan, R. S. et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., Vapnik, V., et al. (1997). Support vector regression machines. *Advances in neural information processing systems*, 9:155–161.

Dubois, J., Galdi, P., Paul, L. K., and Adolphs, R. (2018). A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1756):20170284.

Fawns-Ritchie, C. and Deary, I. J. (2020). Reliability and validity of the UK Biobank cognitive tests. *PloS one*, 15(4):e0231627.

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.

Gao, S., Greene, A. S., Constable, R. T., and Scheinost, D. (2019). Combining multiple connectomes improves predictive modeling of phenotypic measures. *Neuroimage*, 201:116038.

Gong, Q. and He, Y. (2015). Depression, neuroimaging and connectomics: a selective overview. *Biological psychiatry*, 77(3):223–235.

Gong, Q., Li, L., Du, M., Pettersson-Yeo, W., Crossley, N., Yang, X., Li, J., Huang, X., and Mechelli, A. (2014). Quantitative prediction of individual psychopathology in trauma survivors using resting-state FMRI. *Neuropsychopharmacology*, 39(3):681–687.

Goodkind, M., Eickhoff, S. B., Oathes, D. J., Jiang, Y., Chang, A., Jones-Hagata, L. B., Ortega, B. N., Zaiko, Y. V., Roach, E. L., Korgaonkar, M. S., et al. (2015). Identification of a common neurobiological substrate for mental illness. *JAMA psychiatry*, 72(4):305–315.

Greening, S. G. and Mitchell, D. G. (2015). A network of amygdala connections predict individual differences in trait anxiety. *Human Brain Mapping*, 36(12):4819–4830.

Griffa, A., Baumann, P. S., Thiran, J.-P., and Hagmann, P. (2013). Structural connectomics in brain diseases. *Neuroimage*, 80:515–526.

Grotzinger, A. D., Cheung, A. K., Patterson, M. W., Harden, K. P., and Tucker-Drob, E. M. (2019). Genetic and environmental links between general factors of psychopathology and cognitive ability in early childhood. *Clinical Psychological Science*, 7(3):430–444.

Grotzinger, A. D., de la Fuente, J., Davies, G., Nivard, M. G., and Tucker-Drob, E. M. (2021). Transcriptome-wide and Stratified Genomic Structural Equation Modeling Identify Neurobiological Pathways Underlying General and Specific Cognitive Functions. *medRxiv*.

Grotzinger, A. D., Mallard, T. T., Akingbuwa, W. A., Ip, H. F., Adams, M. J., Lewis, C. M., McIntosh, A. M., Grove, J., Dalsgaard, S., Lesch, K.-P., et al. (2020). Genetic architecture of 11 major psychiatric disorders at biobehavioral, functional genomic, and molecular genetic levels of analysis. *medRxiv*.

Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., and Sporns, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biol*, 6(7):e159.

Harden, K. P., Engelhardt, L. E., Mann, F. D., Patterson, M. W., Grotzinger, A. D., Savicki, S. L., Thibodeaux, M. L., Freis, S. M., Tackett, J. L., Church, J. A., et al. (2020). Genetic associations between executive functions and a general factor of psychopathology. *Journal of the American Academy of Child & Adolescent Psychiatry*, 59(6):749–758.

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

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.

Ho, C.-H. and Lin, C.-J. (2012). Large-scale linear support vector regression. *The Journal of Machine Learning Research*, 13(1):3323–3348.

Hu, M., Sim, K., Zhou, J. H., Jiang, X., and Guan, C. (2020). Brain MRI-based 3D Convolutional Neural Networks for Classification of Schizophrenia and Controls*. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 1742–1745.

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

Jiang, R., Calhoun, V. D., Fan, L., Zuo, N., Jung, R., Qi, S., Lin, D., Li, J., Zhuo, C., Song, M., et al. (2020a). Gender differences in connectome-based predictions of individualized intelligence quotient and sub-domain scores. *Cerebral Cortex*, 30(3):888–900.

Jiang, S., Li, X., Li, Z., Chang, X., Chen, Y., Huang, Y., Zhang, Y., Wang, H., Zuo, X., Li, X., et al. (2020b). Cerebello-cerebral connectivity in idiopathic generalized epilepsy. *European radiology*, 30(7):3924–3933.

Johnson, W., Bouchard Jr, T. J., Krueger, R. F., McGue, M., and Gottesman, I. I. (2004). Just one g: Consistent results from three test batteries. *Intelligence*, 32(1):95–107.

Johnson, W., te Nijenhuis, J., and Bouchard Jr, T. J. (2008). Still just 1 g: Consistent results from five test batteries. *Intelligence*, 36(1):81–95.

Jones, D. K., Knösche, T. R., and Turner, R. (2013). White matter integrity, fiber count, and other fallacies: the do's and don'ts of diffusion MRI. *Neuroimage*, 73:239–254.

Josse, J., Husson, F., et al. (2016). missMDA: a package for handling missing values in multivariate data analysis. *Journal of Statistical Software*, 70(1):1–31.

Kawahara, J. et al. (2017). BrainNetCNN: convolutional neural networks for brain networks; towards predicting neurodevelopment. *NeuroImage*, 146:1038–1049.

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

Korgaonkar, M. S., Fornito, A., Williams, L. M., and Grieve, S. M. (2014). Abnormal structural networks characterize major depressive disorder: a connectome analysis. *Biological psychiatry*, 76(7):567–574.

Kotov, R., Krueger, R. F., Watson, D., Achenbach, T. M., Althoff, R. R., Bagby, R. M., Brown, T. A., Carpenter, W. T., Caspi, A., Clark, L. A., et al. (2017). The hierarchical taxonomy of psychopathology (hitop): A dimensional alternative to traditional nosologies. *Journal of abnormal psychology*, 126(4):454.

Lahey, B. B., Applegate, B., Hakes, J. K., Zald, D. H., Hariri, A. R., and Rathouz, P. J. (2012). Is there a general factor of prevalent psychopathology during adulthood? *Journal of abnormal psychology*, 121(4):971.

Li, X. and Duncan, J. (2020). Braingnn: Interpretable brain graph neural network for fmri analysis. *bioRxiv*.

Lund, M. J., Alnæs, D., Schwab, S., van der Meer, D., Andreassen, O. A., Westlye, L. T., and Kaufmann, T. (2020). Differences in directed functional brain connectivity related to age, sex and mental health. *Human brain mapping*, 41(15):4173–4186.

Lyall, D. M., Cullen, B., Allerhand, M., Smith, D. J., Mackay, D., Evans, J., Anderson, J., Fawns-Ritchie, C., McIntosh, A. M., Deary, I. J., et al. (2016). Cognitive test scores in UK Biobank: data reduction in 480,416 participants and longitudinal stability in 20,346 participants. *PloS one*, 11(4):e0154222.

Ma, Q., Tang, Y., Wang, F., Liao, X., Jiang, X., Wei, S., Mechelli, A., He, Y., and Xia, M. (2020). Transdiagnostic dysfunctions in brain modules across patients with schizophrenia, bipolar disorder, and major depressive disorder: a connectome-based study. *Schizophrenia bulletin*, 46(3):699–712.

Madole, J. W., Ritchie, S. J., Cox, S. R., Buchanan, C. R., Hernández, M. V., Maniega, S. M., Wardlaw, J. M., Harris, M. A., Bastin, M. E., Deary, I. J., et al. (2021). Aging-Sensitive Networks Within the Human Structural Connectome Are Implicated in Late-Life Cognitive Declines. *Biological psychiatry*, 89(8):795–806.

Mihalik, A., Brudfors, M., Robu, M., Ferreira, F. S., Lin, H., Rau, A., Wu, T., Blumberg, S. B., Kanber, B., Tariq, M., et al. (2019). ABCD Neurocognitive Prediction Challenge 2019: predicting individual fluid intelligence scores from structural MRI using probabilistic segmentation and kernel ridge regression. In *Challenge in Adolescent Brain Cognitive Development Neurocognitive Prediction*, pages 133–142. Springer.

Miller, K. L. et al. (2016). Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nature neuroscience*, 19(11):1523.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Ocker, G. K., Josić, K., Shea-Brown, E., and Buice, M. A. (2017). Linking structure and activity in nonlinear spiking networks. *PLoS computational biology*, 13(6):e1005583.

Oestreich, L. K., Randeniya, R., and Garrido, M. I. (2019). White matter connectivity reductions in the pre-clinical continuum of psychosis: A connectome study. *Human brain mapping*, 40(2):529–537.

Payabvash, S., Palacios, E. M., Owen, J. P., Wang, M. B., Tavassoli, T., Gerdes, M., Brandes-Aitken, A., Cuneo, D., Marco, E. J., and Mukherjee, P. (2019). White matter connectome edge density in children with autism spectrum disorders: potential imaging biomarkers using machine-learning models. *Brain connectivity*, 9(2):209–220.

Pinto, J. V., Moulin, T. C., and Amaral, O. B. (2017). On the transdiagnostic nature of peripheral biomarkers in major psychiatric disorders: a systematic review. *Neuroscience & Biobehavioral Reviews*, 83:97–108.

Qi, S., Meesters, S., Nicolay, K., ter Haar Romeny, B. M., and Ossenblok, P. (2015). The influence of construction methodology on structural brain network measures: A review. *Journal of neuroscience methods*, 253:170–182.

Qin, J., Wei, M., Liu, H., Yan, R., Luo, G., Yao, Z., and Lu, Q. (2014). Abnormal brain anatomical topological organization of the cognitive-emotional and the frontoparietal circuitry in major depressive disorder. *Magnetic resonance in medicine*, 72(5):1397–1407.

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.

Raji, C. A., Wang, M. B., Nguyen, N., Owen, J. P., Palacios, E. M., Yuh, E. L., and Mukherjee, P. (2020). Connectome mapping with edge density imaging differentiates pediatric mild traumatic brain injury from typically developing controls: proof of concept. *Pediatric radiology*, 50(11):1594–1601.

Ritchie, S. J., Cox, S. R., Shen, X., Lombardo, M. V., Reus, L. M., Alloza, C., Harris, M. A., Alderson, H. L., Hunter, S., Neilson, E., et al. (2018). Sex differences in the adult human brain: evidence from 5216 UK biobank participants. *Cerebral cortex*, 28(8):2959–2975.

Robinson, E. C., Hammers, A., Ericsson, A., Edwards, A. D., and Rueckert, D. (2010). Identifying population differences in whole-brain structural networks: a machine learning approach. *NeuroImage*, 50(3):910–919.

Rutland, J. W., Brown, S., Verma, G., Feldman, R. E., Sharma, H., Markowitz, M., Schneider, M., Delman, B. N., Murrough, J., and Balchandani, P. (2019). Hippocampal subfield-specific connectivity findings in major depressive disorder: A 7 Tesla diffusion MRI study. *Journal of psychiatric research*, 111:186–192.

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.

Shen, X., Finn, E. S., Scheinost, D., Rosenberg, M. D., Chun, M. M., Papademetris, X., and Constable, R. T. (2017). Using connectome-based predictive modeling to predict individual behavior from brain connectivity. *nature protocols*, 12(3):506–518.

Siegel, J. S., Ramsey, L. E., Snyder, A. Z., Metcalf, N. V., Chacko, R. V., Weinberger, K., Baldassarre, A., Hacker, C. D., Shulman, G. L., and Corbetta, M. (2016). Disruptions of network connectivity predict impairment in multiple behavioral domains after stroke. *Proceedings of the National Academy of Sciences*, 113(30):E4367–E4376.

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

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

Sripada, C., Angstadt, M., Rutherford, S., and Taxali, A. (2019). Predictive Modeling Achieves High Test-Retest Reliability with Resting State Functional Connectivity. *bioRxiv*, page 796714.

Sun, Y., Dai, Z., Li, J., Collinson, S. L., and Sim, K. (2017). Modular-level alterations of structure–function coupling in schizophrenia connectome. *Human brain mapping*, 38(4):2008–2025.

Suo, X., Du Lei, L. L., Li, W., Dai, J., Wang, S., He, M., Zhu, H., Kemp, G. J., and Gong, Q. (2018). Psychoradiological patterns of small-world properties and a systematic review of connectome studies of patients with 6 major psychiatric disorders. *Journal of psychiatry & neuroscience: JPN*, 43(6):416.

van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., and van der Maas, H. L. (2017). What is the p-factor of psychopathology? some risks of general factor modeling. *Theory & Psychology*, 27(6):759–773.

Vapnik, V., Golowich, S. E., Smola, A., et al. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in neural information processing systems*, pages 281–287.

Vieira, S., Pinaya, W. H., and Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. *Neuroscience & Biobehavioral Reviews*, 74:58–75.

Vollebergh, W. A., Iedema, J., Bijl, R. V., de Graaf, R., Smit, F., and Ormel, J. (2001). The structure and stability of common mental disorders: the nemesis study. *Archives of general psychiatry*, 58(6):597–603.

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., and Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine*, 368(15):1388–1397.

Wang, P., Kong, R., Kong, X., Liégeois, R., Orban, C., Deco, G., Van Den Heuvel, M. P., and Yeo, B. T. (2019). Inversion of a large-scale circuit model reveals a cortical hierarchy in the dynamic resting human brain. *Science advances*, 5(1):eaat7854.

Wang, Z., Goerlich, K. S., Ai, H., Aleman, A., Luo, Y.-j., and Xu, P. (2021). Connectome-based predictive modeling of individual anxiety. *Cerebral Cortex*, 31(6):3006–3020.

Xu, X., Li, W., Mei, J., Tao, M., Wang, X., Zhao, Q., Liang, X., Wu, W., Ding, D., and Wang, P. (2020). Feature selection and combination of information in the functional brain connectome for discrimination of mild cognitive impairment and analyses of altered brain patterns. *Frontiers in aging neuroscience*, 12:28.

Yamamoto, N., Sukegawa, S., Kitamura, A., Goto, R., Noda, T., Nakano, K., Takabatake, K., Kawai, H., Nagatsuka, H., Kawasaki, K., et al. (2020). Deep Learning for Osteoporosis Classification Using Hip Radiographs and Patient Clinical Covariates. *Biomolecules*, 10(11):1534.

Yeung, H. W., Luz, S., Cox, S. R., Buchanan, C. R., Whalley, H. C., and Smith, K. M. (2020). Pipeline comparisons of convolutional neural networks for structural connectomes: predicting sex across 3,152 participants. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1692–1695. IEEE.

Yu, J., Rawtaer, I., Feng, L., Kua, E.-H., and Mahendran, R. (2021). The functional and structural connectomes associated with geriatric depression and anxiety symptoms in mild cognitive impairment: Cross-syndrome overlap and generalization. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 110:110329.

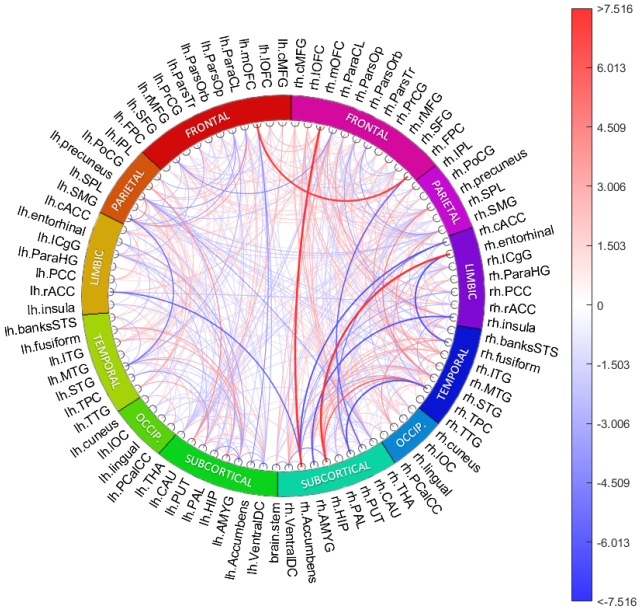
Zhang, H., Schneider, T., Wheeler-Kingshott, C. A., and Alexander, D. C. (2012). NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. *Neuroimage*, 61(4):1000–1016.

Zheng, K., Wang, H., Li, J., Yan, B., Liu, J., Xi, Y., Zhang, X., Yin, H., Tan, Q., Lu, H., et al. (2019). Structural networks analysis for depression combined with graph theory and the properties of fiber tracts via diffusion tensor imaging. *Neuroscience letters*, 694:34–40.

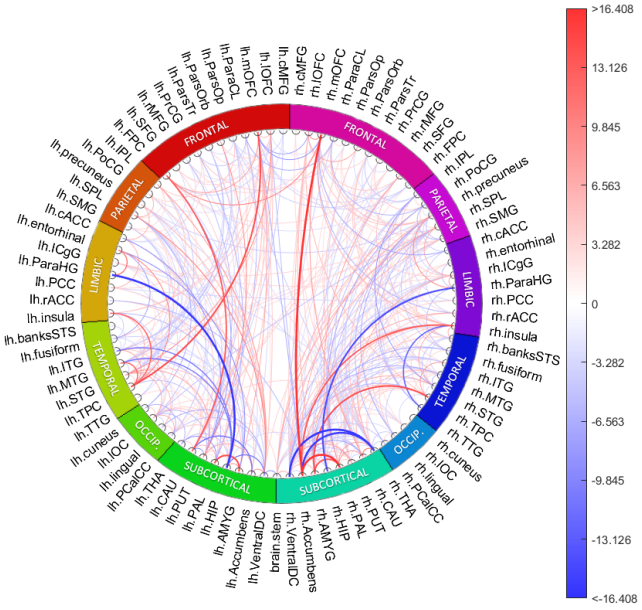
Zhu, J., Wang, C., Liu, F., Qin, W., Li, J., and Zhuo, C. (2016). Alterations of functional and structural networks in schizophrenia patients with auditory verbal hallucinations. *Frontiers in human neuroscience*, 10:114.

Zou, L., Zheng, J., Miao, C., Mckeown, M. J., and Wang, Z. J. (2017). 3d cnn based automatic diagnosis of attention deficit hyperactivity disorder using functional and structural mri. *IEEE Access*, 5:23626–23636.

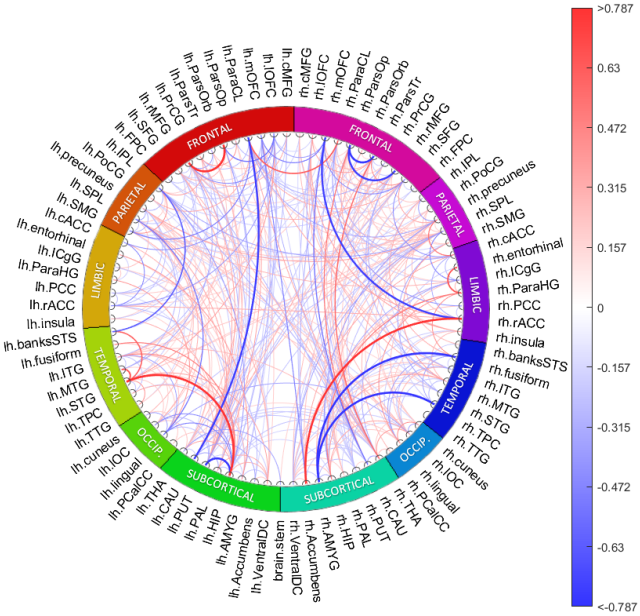
Figures



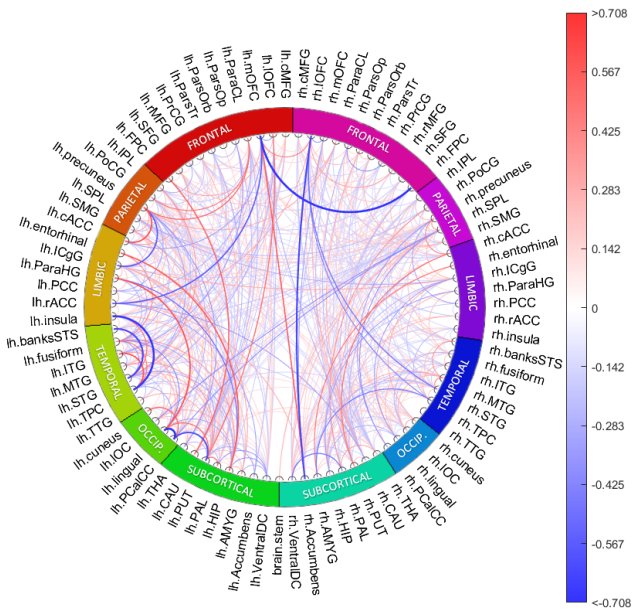
(a) Sex



(b) Age

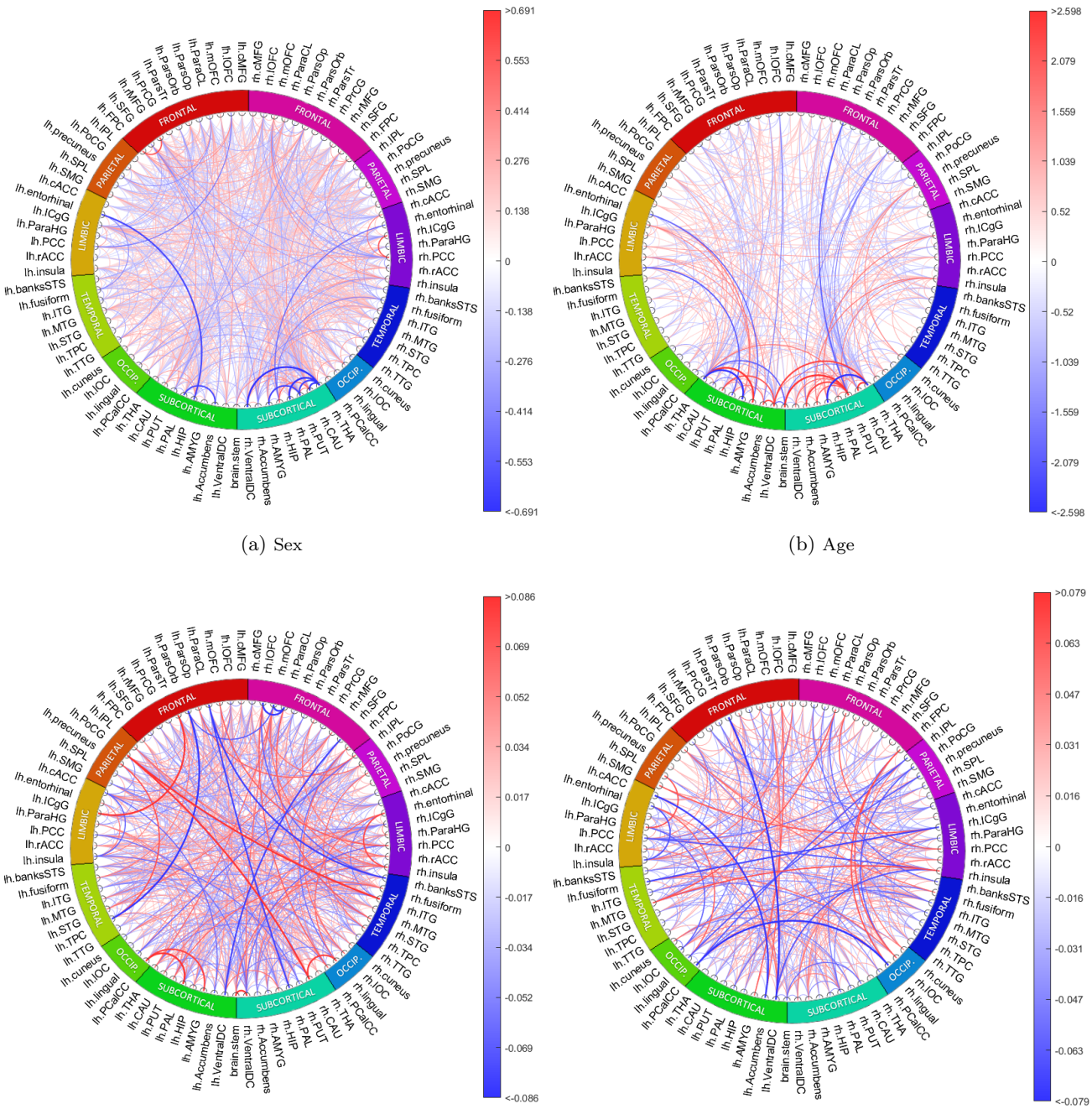


(c) g-factor



(d) p-factor

Figure 1: The saliency maps based on SC connectomes for all prediction tasks. SC = streamline count. Blue represents negative gradients and red represents positive gradients. The colour bar has interval of 0.1 to 99.9 percentile of the gradients



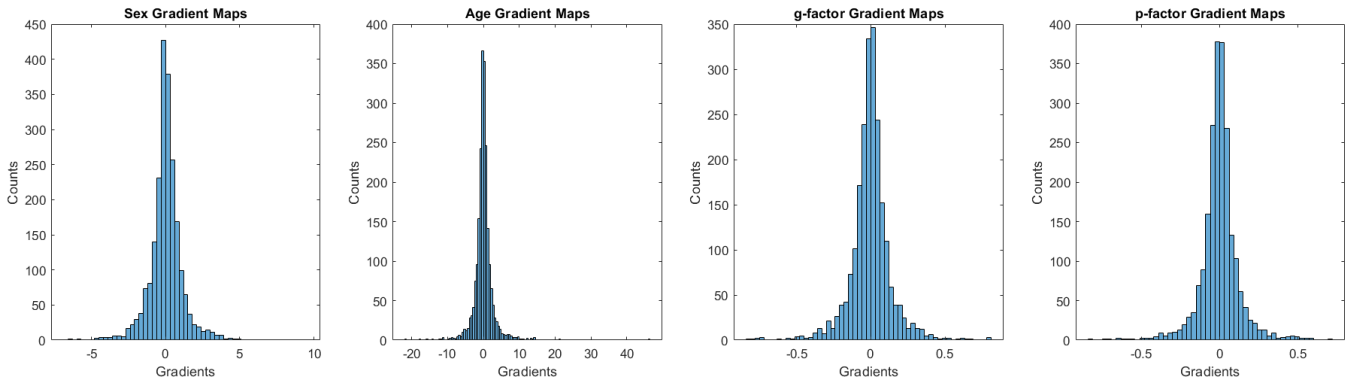
(a) Sex

(b) Age

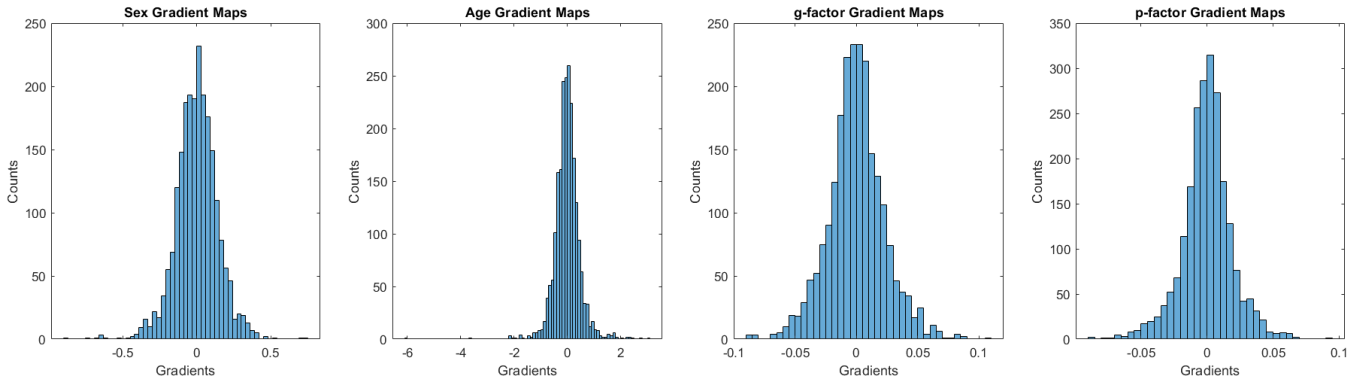
(c) g-factor

(d) p-factor

Figure 2: The saliency maps based on FA connectomes for all prediction tasks. FA = fractional anisotropy. Blue represents negative gradients and red represents positive gradients. The colour bar has interval of 0.1 to 99.9 percentile of the gradients



(a) SC



(b) FA

Figure 3: The histograms of gradients based on SC and FA for each prediction task. FA = fraction anisotropy; SC = streamline count. The kurtosis measures of gradient distributions for age, g-factor and p-factor predictions based on SC were 43.67, 10.93, and 11.50 respectively, and were 19.24, 4.36 and 5.16 based on FA. This implies that there were more edges having gradients close to the mean (zero) for age predictions than in the other 2 prediction tasks. This indicates that the model relied on a smaller subset of edges for age prediction, but relied on a wide range of edges for g-factor and p-factor predictions.

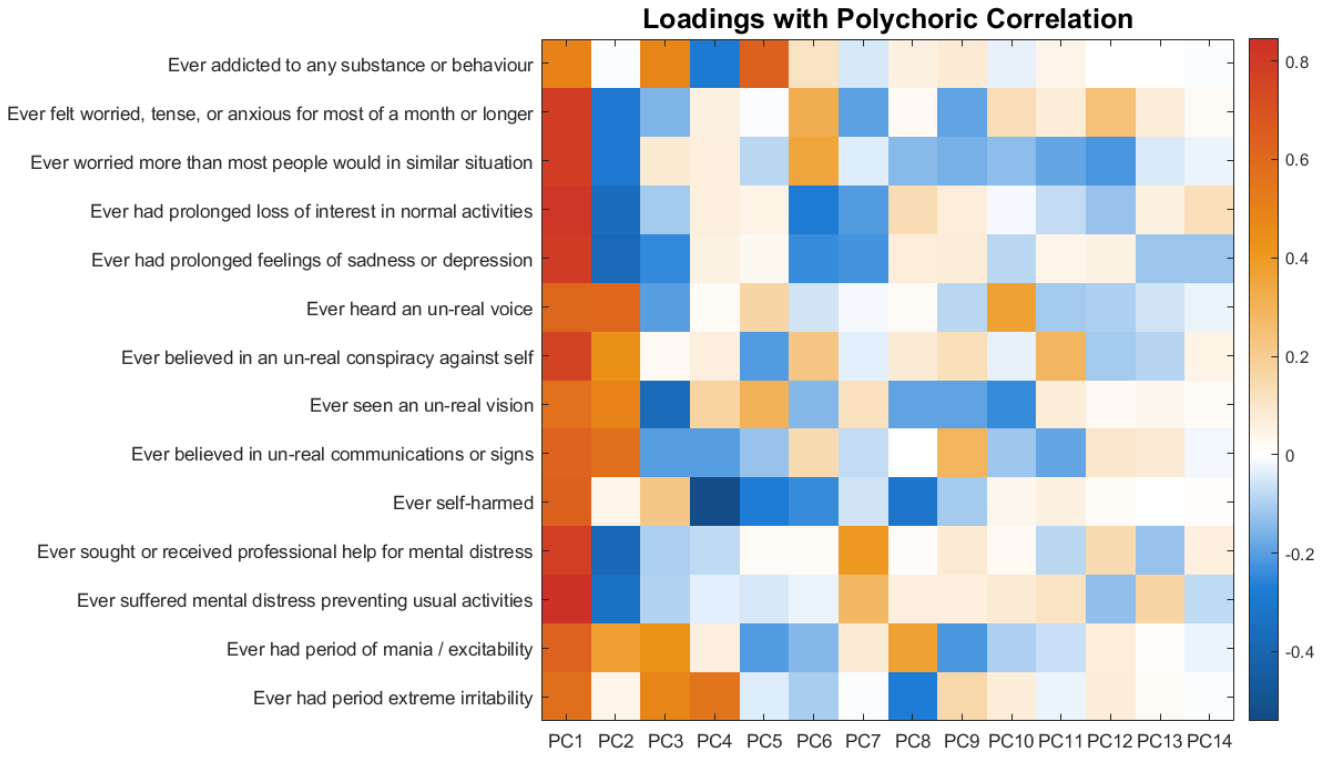
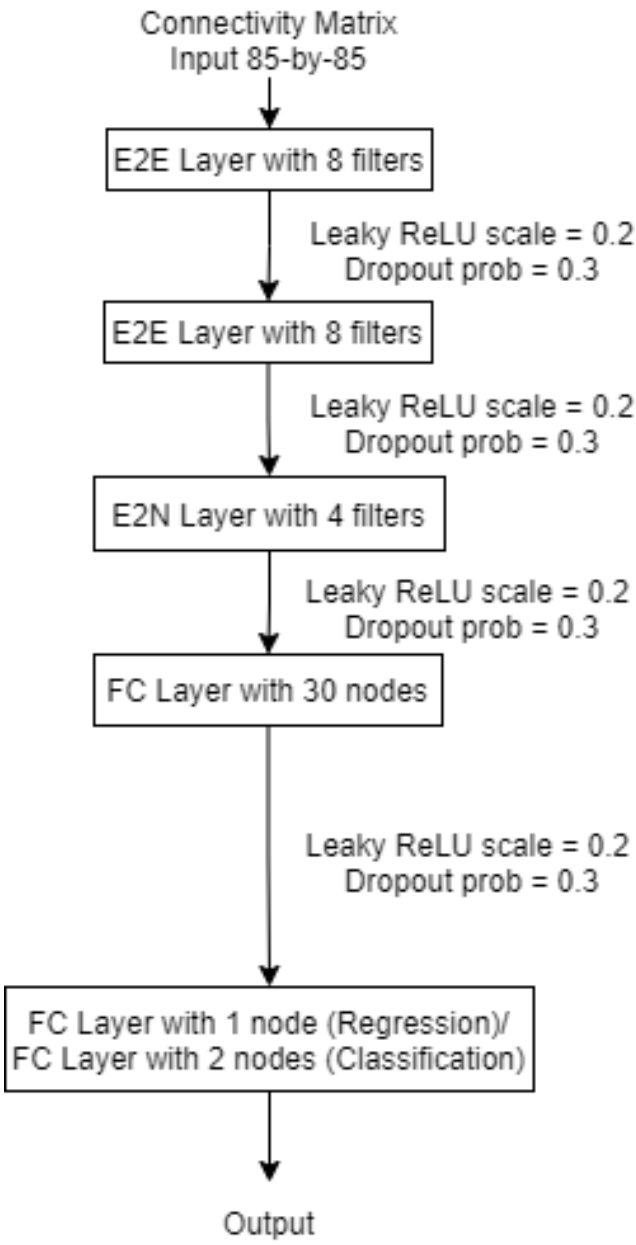
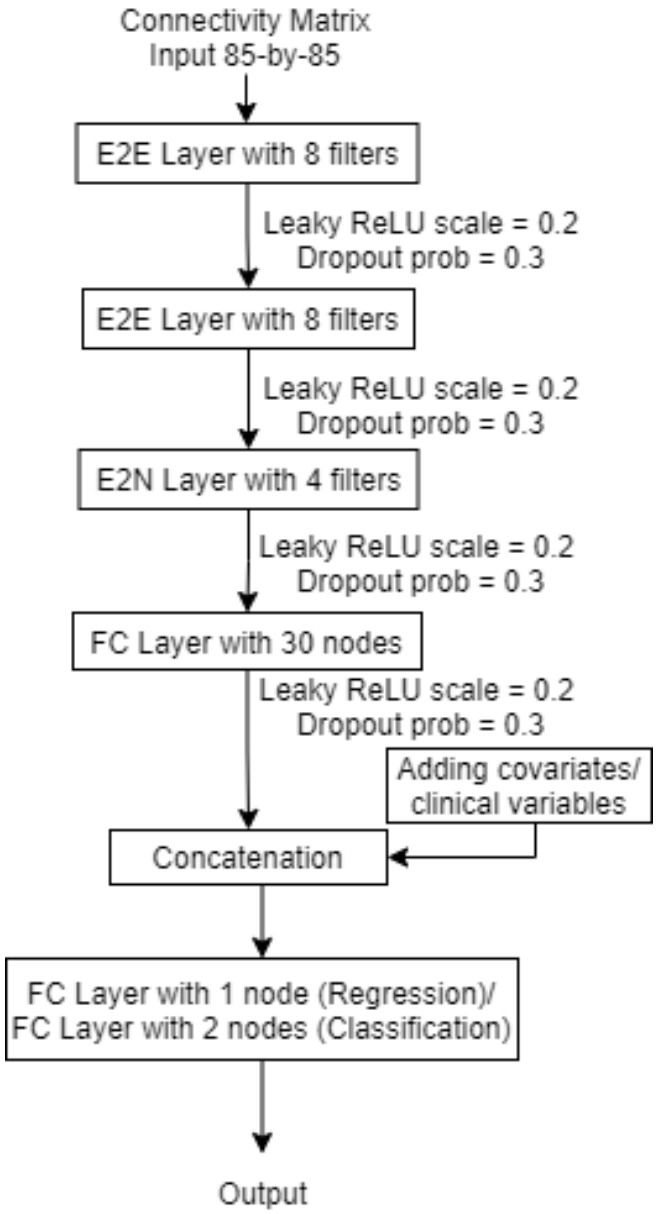


Figure 4: Loadings of the 14 MHQ items in UKB on the principal components derived with PCA. We saw strong positive loadings of depression/anxiety items on the first PC and positive loadings of psychotic experience items on the second PC. Please see Supplementary Table 1 in section A.2.2 for exact values of standardised item loadings on principal components.



(a) The BrainNetCNN Architecture



(b) The BrainNetCNN Architecture with Covariates

Figure 5: Two BrainNetCNN Architectures. E2E: Edge-to-Edge, E2N: Edge-to-Node, FC: Fully connected, Leaky ReLU: Leaky Rectified Linear Unit, prob: probability of dropout

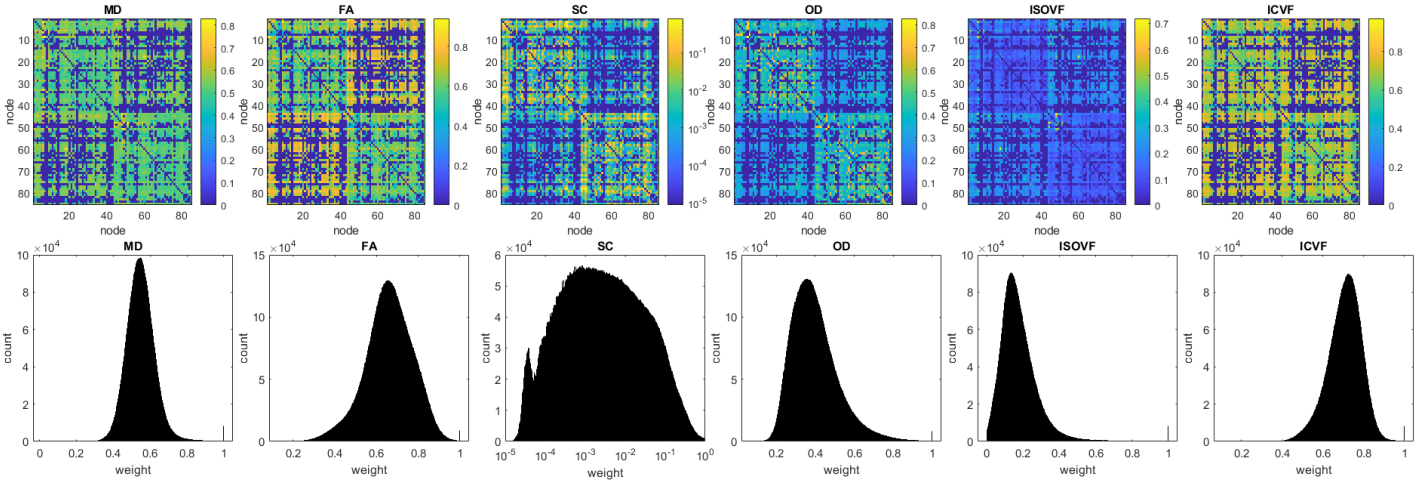


Figure 6: Top: 85 *times* 85 mean connectivity matrices of inter-region connection weights averaged across all participants (N = 8,183) for six network weightings after performing maximum value normalisation. The two blocks along the diagonal in each case correspond to the left and right hemispheres. Bottom: the corresponding histograms of nonzero edge weights pooled across all participants for each weighting (SC is log-scaled). MD = mean diffusivity; FA = fraction anisotropy; SC = streamline count; OD = orientation dispersion; ISOVF = Isotropic volume fraction; ICVF = Intracellular volume fraction. The SC weights were shown in log-scale. Apart from the SC weights, the other five network weightings had similar distribution shape. MD, FA, OD and ICVF had skewness in the range [-0.3, 0.3]. ISOVF had a skewness of 1.36 and SC has a skewness of 8.01.

Tables

Weights	MD	FA	SC	OD	ISOVF	ICVF
Validation	78.07 (0.54)	80.76 (0.70)	89.18 (0.60)	83.53 (0.51)	82.07 (0.75)	80.45 (0.89)
Training	87.17 (2.35)	89.44 (2.40)	94.10 (1.28)	91.60 (1.76)	91.49 (2.70)	91.58 (3.27)
Test	78.15 (0.86)	79.74 (0.82)	86.91 (0.72)	82.88 (0.95)	81.82 (0.22)	78.34 (1.03)

(a) Sex prediction accuracies (mean percentage with standard deviation in brackets) with BrainNetCNN for different connectivity weightings.

Weights	MD	FA	SC	OD	ISOVF	ICVF
Validation	4.696 (0.640)	4.636 (0.645)	4.153 (0.727)	4.374 (0.691)	4.256 (0.703)	4.731 (0.625)
Training	3.797 (0.785)	3.669 (0.796)	3.634 (0.795)	3.550 (0.804)	3.538 (0.806)	3.941 (0.765)
Test	4.640 (0.644)	4.674 (0.636)	4.245 (0.706)	4.571 (0.655)	4.363 (0.678)	4.692 (0.626)

(b) Age prediction performance (mean absolute errors with correlations in brackets) with BrainNetCNN for different connectivity weightings

Weights	MD	FA	SC	OD	ISOVF	ICVF
Validation	0.151 (0.770)	0.168 (0.770)	0.227 (0.764)	0.185 (0.768)	0.168 (0.780)	0.169 (0.769)
Training	0.232 (0.764)	0.275 (0.761)	0.314 (0.745)	0.315 (0.753)	0.286 (0.758)	0.264 (0.763)
Test	0.138 (0.782)	0.168 (0.780)	0.201 (0.780)	0.132 (0.787)	0.155 (0.786)	0.160 (0.782)

(c) g-factor prediction performance (correlations with mean absolute error in brackets) with BrainNetCNN for different connectivity weightings

Weights	MD	FA	SC	OD	ISOVF	ICVF
Validation	0.100 (0.803)	0.089 (0.780)	0.117 (0.785)	0.118 (0.794)	0.095 (0.795)	0.073 (0.794)
Training	0.238 (0.797)	0.252 (0.768)	0.323 (0.751)	0.218 (0.785)	0.403 (0.785)	0.199 (0.787)
Test	0.116 (0.817)	0.112 (0.792)	0.143 (0.790)	0.130 (0.804)	0.112 (0.804)	0.103 (0.805)

(d) p-factor prediction performance (correlations with mean absolute error in brackets) with BrainNetCNN for different connectivity weightings

Table 1: Prediction performances with BrainNetCNN of four different prediction tasks based on different connectivity weightings. MD = mean diffusivity; FA = fraction anisotropy; SC = streamline count; OD = orientation dispersion; ISOVF = isotropic volume fraction; ICVF = intracellular volume fraction. Use of connectomes based on streamline counts generally led to the best predictive performance results.

Weights	MD	FA	SC	OD	ISOVF	ICVF
Validation	0.247 (0.758)	0.251 (0.760)	0.264 (0.756)	0.255 (0.752)	0.242 (0.766)	0.247 (0.751)
Training	0.311 (0.746)	0.314 (0.744)	0.354 (0.731)	0.325 (0.737)	0.303 (0.748)	0.307 (0.738)
Test	0.247 (0.762)	0.251 (0.764)	0.244 (0.769)	0.242 (0.760)	0.238 (0.768)	0.249 (0.759)

(a) g-factor prediction performance (correlations with mean absolute error in brackets) using BrainNetCNN model with age and sex covariates for different connectivity weightings

Weights	MD	FA	SC	OD	ISOVF	ICVF
Validation	0.228 (0.773)	0.223 (0.757)	0.204 (0.758)	0.228 (0.758)	0.183 (0.768)	0.215 (0.768)
Training	0.254 (0.767)	0.269 (0.749)	0.335 (0.738)	0.274 (0.751)	0.376 (0.739)	0.262 (0.759)
Test	0.243 (0.778)	0.239 (0.763)	0.231 (0.763)	0.245 (0.764)	0.202 (0.772)	0.238 (0.771)

(b) p-factor prediction performance (correlations with mean absolute error in brackets) using BrainNetCNN model with age and sex covariates for different connectivity weightings

Table 2: g-factor and p-factor prediction performance using BrainNetCNN model with age and sex covariates for different connectivity weightings. MD = mean diffusivity; FA = fraction anisotropy; SC = streamline count; OD = orientation dispersion; ISOVF = Isotropic volume fraction; ICVF = Intracellular volume fraction. After adding covariates as external regressors to the BrainNetCNN model, all network weightings had comparable performances on both g-factor and p-factor prediction.

Model	BrainNetCNN	Ridge Logistic Regression	LASSO Logistic Regression	SVM	KRR
Validation	89.18 (0.70)	89.67 (1.01)	88.22 (1.42)	89.77 (0.91)	86.44 (1.30)
Training	94.10 (1.28)	93.75 (1.40)	90,41 (0.29)	93.90 (0.89)	88.09 (0.36)
Test	86.91 (0.72)	87.54 (0.48)	85.90 (0.68)	87.27 (0.56)	84.28 (0.36)

(a) Performance of the five tested algorithms on sex classification (mean percentage with standard deviation in brackets) with connectivity measures based on streamline counts.

Model	BrainNetCNN	Ridge Regression	LASSO Regression	SVM-R	KRR
Validation	4.153 (0.727)	4.221 (0.713)	4.238 (0.71)	4.318 (0.698)	4.402 (0.685)
Training	3.634 (0.795)	3.773 (0.779)	3.897 (0.763)	3.975 (0.738)	4.271 (0.707)
Test	4.245 (0.706)	4.258 (0.698)	4.282 (0.695)	4.357 (0.684)	4.434 (0.669)

(b) Performance of the five tested algorithms on age prediction (mean absolute error with correlations in brackets) with connectivity measures based on streamline counts.

Model	BrainNetCNN	Ridge Regression	LASSO Regression	SVM-R	KRR
Validation	0.264 (0.756)	0.270 (0.748)	0.268 (0.749)	0.270 (0.745)	0.245 (0.759)
Training	0.354 (0.731)	0.330 (0.735)	0.335 (0.733)	0.316 (0.730)	0.354 (0.728)
Test	0.244 (0.769)	0.248 (0.764)	0.248 (0.765)	0.246 (0.763)	0.257 (0.768)

(c) Performance of the five tested algorithms on cognitive g-factor prediction (correlation with mean absolute error in brackets) with connectivity measures and common covariates based on streamline counts

Model	BrainNetCNN	Ridge Regression	LASSO Regression	SVM-R	KRR
Validation	0.204 (0.758)	0.223 (0.767)	0.226 (0.768)	0.224 (0.794)	0.207 (0.774)
Training	0.335 (0.738)	0.281 (0.753)	0.279 (0.755)	0.286 (0.750)	0.269 (0.758)
Test	0.231 (0.763)	0.244 (0.772)	0.237 (0.776)	0.230 (0.794)	0.229 (0.780)

(d) Performance of the five tested algorithms on mental health p-factor prediction (correlation with mean absolute error in brackets) with connectivity measures and common covariates based on streamline counts

Table 3: Performances of the five tested algorithms on four different prediction tasks with connectivity measures based on streamline counts. The linear ML models have comparable performance with the non-linear model (KRR) and the deep learning model (BrainNetCNN) on all four evaluated prediction tasks.

UK Biobank characteristics	Mean(SD)	Data Available (N)
Sex (F:M)	4314:3869	8183
Age (in years)	62.43(7.445)	8183
VNR	6.8721(2.103)	7806
RT (in $\log x$)	6.3581(0.1681)	8090
Pairs Match (in $\log(x+1)$)	1.2854(0.6480)	8127
Prospective Memory(1:0)	7257:855	8112

Table 4: Summary characteristics of the UK Biobank participants included in the study. VNR: Verbal Numerical Reasoning. RT: Reaction Time. Pairs Match: Pairs Matching. For the prospective memory test, **1** means recall at the first attempt and **0** otherwise.

Cognitive Tasks	Loadings
VNR	0.6942
RT	-0.5503
Pairs Match	-0.5344
Prospective Memory	0.5685
PropVar	0.3484

Table 5: Standardised loadings of individual cognitive test scores on the cognitive g-factor. VNR: Verbal Numerical Reasoning. RT: Reaction Time. Pairs Match: Pairs Matching. PropVar: Proportion of variance explained by the g-factor.