NeuroImage 291 (2024) 120600

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Computational limits to the legibility of the imaged human brain

James K. Ruffle [a,*], Robert J Gray [a], Samia Mohinta [a], Guilherme Pombo [a], Chaitanya Kaul [b], Harpreet Hyare [a], Geraint Rees [a], Parashkev Nachev [a,*]

[a] Queen Square Institute of Neurology, University College London, London, United Kingdom
[b] School of Computing Science, University of Glasgow, Glasgow, United Kingdom

ARTICLE INFO

Keywords:
Deep learning
Neuroimaging
Magnetic resonance imaging
High performance computing
Artificial intelligence
Brain structure
Brain function
Functional connectivity
Machine learning

ABSTRACT

Our knowledge of the organisation of the human brain at the *population-level* is yet to translate into power to predict functional differences at the *individual-level,* limiting clinical applications and casting doubt on the generalisability of inferred mechanisms. It remains unknown whether the difficulty arises from the absence of individuating biological patterns within the brain, or from limited power to access them with the models and compute at our disposal.

Here we comprehensively investigate the resolvability of such patterns with data and compute at unprecedented scale. Across 23 810 unique participants from UK Biobank, we systematically evaluate the predictability of 25 individual biological characteristics, from all available combinations of structural and functional neuroimaging data. Over 4526 GPU*hours of computation, we train, optimize, and evaluate out-of-sample 700 individual predictive models, including fully-connected feed-forward neural networks of demographic, psychological, serological, chronic disease, and functional connectivity characteristics, and both uni- and multimodal 3D convolutional neural network models of macro- and micro-structural brain imaging.

We find a marked discrepancy between the high predictability of sex (balanced accuracy 99.7%), age (mean absolute error 2.048 years, R^2 0.859), and weight (mean absolute error 2.609Kg, R^2 0.625), for which we set new state-of-the-art performance, and the surprisingly low predictability of other characteristics. Neither structural nor functional imaging predicted an individual's psychology better than the coincidence of common chronic disease ($p < 0.05$). Serology predicted chronic disease ($p < 0.05$) and was best predicted by it ($p < 0.001$), followed by structural neuroimaging ($p < 0.05$).

Our findings suggest either more informative imaging or more powerful models will be needed to decipher individual level characteristics from the human brain. We make our models and code openly available.

1. Introduction

That the brain exhibits a finely wrought functional-anatomical organisation is no longer in doubt. Macro- and micro-structural features, task-specific and resting state neural activity, focal disruptive and lesion-related neural dependence, all show richly structured, replicable variation across the population (Littlejohns et al., 2020; Bethlehem et al., 2022; Elliott et al., 2018; Wang et al., 2022; Bazinet et al., 2023; Hansen et al., 2022; Suárez et al., 2020; Honey et al., 2009; Fischl et al., 2008; Thomas Yeo, 2011; Hansen et al., 2021). But whether these now familiar patterns can explain *individual-level* differences remains an open question (Marek et al., 2022; Finn et al., 2015; Bzdok et al., 2020; Wu et al., 2023). Its answer is crucially important for two reasons: first, because the clinical applications of our knowledge of the brain are necessarily addressed not to populations but to individual patients, and second, because the fewer the individuals to which any model generalises, the weaker the grounds it provides for mechanistic inference, no matter how well supported its parameters. It is also a far harder question to address, for individual-level models must inevitably capture the many complex interactions between multiple features on which individual functions may jointly depend. Model architectures of the requisite expressivity (LeCun et al., 2015; Goodfellow et al., 2017; Richards et al.,

Abbreviations: CNN, convolutional neural network; DWI, diffusion weighted imaging; FDR, false discovery rate; FLAIR, fluid-attenuated inversion recovery; FNN, feed-forward neural network; Hb, haemoglobin; HbA1c, glycated haemoglobin; HDL, high density lipoprotein; LDL, low density lipoprotein; MAE, mean absolute error; MRI, magnetic resonance imaging; rsfMRI, resting-state functional magnetic resonance imaging.

* Corresponding authors.

E-mail addresses: j.ruffle@ucl.ac.uk (J.K. Ruffle), p.nachev@ucl.ac.uk (P. Nachev).

https://doi.org/10.1016/j.neuroimage.2024.120600
Received 19 December 2023; Received in revised form 8 March 2024; Accepted 31 March 2024
Available online 2 April 2024

2019)—whose least upper bound is unknown—plausibly require data of far greater scale and inclusivity than is usual in the field (Szucs and Ioannidis, 2020), and computational resources rare in neuroscience.

This places us in a Catch 22. Failure to find individually discriminating patterns may be a consequence not of their absence, but of inadequacies of the data and the computational regime (Marek et al., 2022; Schulz et al., 2020). Yet calibrating the regime to the demands of the brain's complexity cannot be done with small samples, so distinguishing between the two possibilities is impossible without the resource the distinction is needed to justify in the first place.

How do we break out of this? Models of the necessary complexity here can always be improved upon, so no limit on theoretically achievable fidelity can be definitively set. But we can conduct a *comparative* analysis of a set of biological characteristics, at the current *practical* limit of model expressivity and the compute it requires. If such a state-of-the-art analysis reveals a marked *contrast* of predictability—very high for some characteristics, very low for others—then the conclusion that the unpredictable characteristics are *practically* inaccessible with current data and models is corroborated, and a wholesale change in our approach—data, models, and compute—is motivated (Fig. 1). If performance is uniformly poor, then our test may have been inadequate; if it is uniformly excellent, then no change to current practice is indicated.

Here we conduct such an analysis in a sample of 23 810 unique participants from UK Biobank (Littlejohns et al., 2020; Alfaro-Almagro et al., 2018), systematically evaluating the predictability of a wide set of individual characteristics from all possible combinations of available neuroimaging data, spanning structural and functional domains. We build and evaluate a suite of 700 discriminative models of different combinations of brain imaging—uni- and multi-modal, macro- and micro-structural, and resting state functional—with biological characteristics ranging across psychology, serology, and disease comorbidity. Over 4526 GPU*hours of computation, including extensive hyper-parameter optimisation, we comprehensively evaluate the individual-level predictability of common biological and pathological characteristics from current brain imaging in this general population cohort. Our analysis sets a new state-of-the-art benchmark for age regression and sex classification (for which we make all model weights open source), demonstrating the felicity of our modelling approach, and reveals a marked heterogeneity of individual predictability that argues for a radical change in the current brain modelling regime.

2. Materials and methods

2.1. Data

Data was retrieved from the UK Biobank repository (https://www.ukbiobank.ac.uk) (Littlejohns et al., 2020; Alfaro-Almagro et al., 2018; Sudlow et al., 2015). From here, we retrieved an unselected fully inclusive representation of the cohort, a sample of 502 505 individuals with 3581 individual variables detailing them. Data missingness of the parameters modelled never exceeded 20% for any variable in our study. We imputed missing variables from this full set using multivariate iterative imputation (Pedregosa et al., 2011), with hyperparameters of 10,000 maximum iterations, a default stopping condition tolerance of 0.001 and an initial strategy of median imputation.

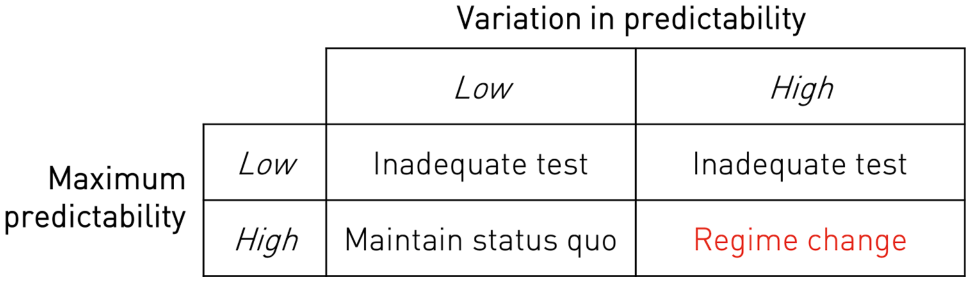
We selected 31 participant variables as predictive modelling targets that engendered a range of domains of individuality. We chose this number as a reasonable balance between high dimensionality of individuating factors, and the volume of models required to be trained for each individual proposed target and anticipating a training time of several months even on cutting-edge computational hardware. We applied an exclusion criterion after variable selection, where if a feature was categorical/binary in nature, it was excluded if an imbalance between majority and minority class were 10:1 or greater. The rationale was to minimize the impact of imbalance on our evaluations of predictability. This led to 3 variables being removed (namely pathological classes of previous stroke, myocardial infarction, and the presence of type 2 diabetes), leaving 28 variables for possible targets.

Next, we computed the pairwise Maximal Information Coefficient (Reshef et al., 2011) between all 28 targets. The purpose of this was to identify features that were highly collinear to one another, limiting the interpretability of their individual prediction where they are jointly modelled. This led to the removal of 3 further variables, namely neuroticism (closely related to several other psychological factors, haematocrit (closely related to haemoglobin concentration), and body mass index (closely related to weight). In total, this process yielded 25 unique and minimally inter-related target features (Fig. 2, Supplementary Figure 1, Supplementary Table 1). The choice to not select further variables was entirely driven by available computing resource and the substantial GPU-hours required for the large number of possible models with different input features to predict each target.

We then organised these 25 targets into their respective domains, as follows: i) Constitutional, comprising sex, age, weight and handedness; ii) Psychology, comprising feelings of guilt, loneliness, worry, feeling tense, anxious, nervous, fed-up, sensitive, irritable, miserable, with mood-swings, and participant reaction time (ms); iii) Disease, comprising body fat (obesity), hypertension, asthma, atopy, smoking (addiction); and iv) Serological, comprising concentrations of haemoglobin (g/dl), HbA1c (mmol/mol), HDL (mmol/L) and LDL (mmol/L) (Fig. 2, Supplementary Figure 1, Supplementary Table 1).

2.2. Participant selection

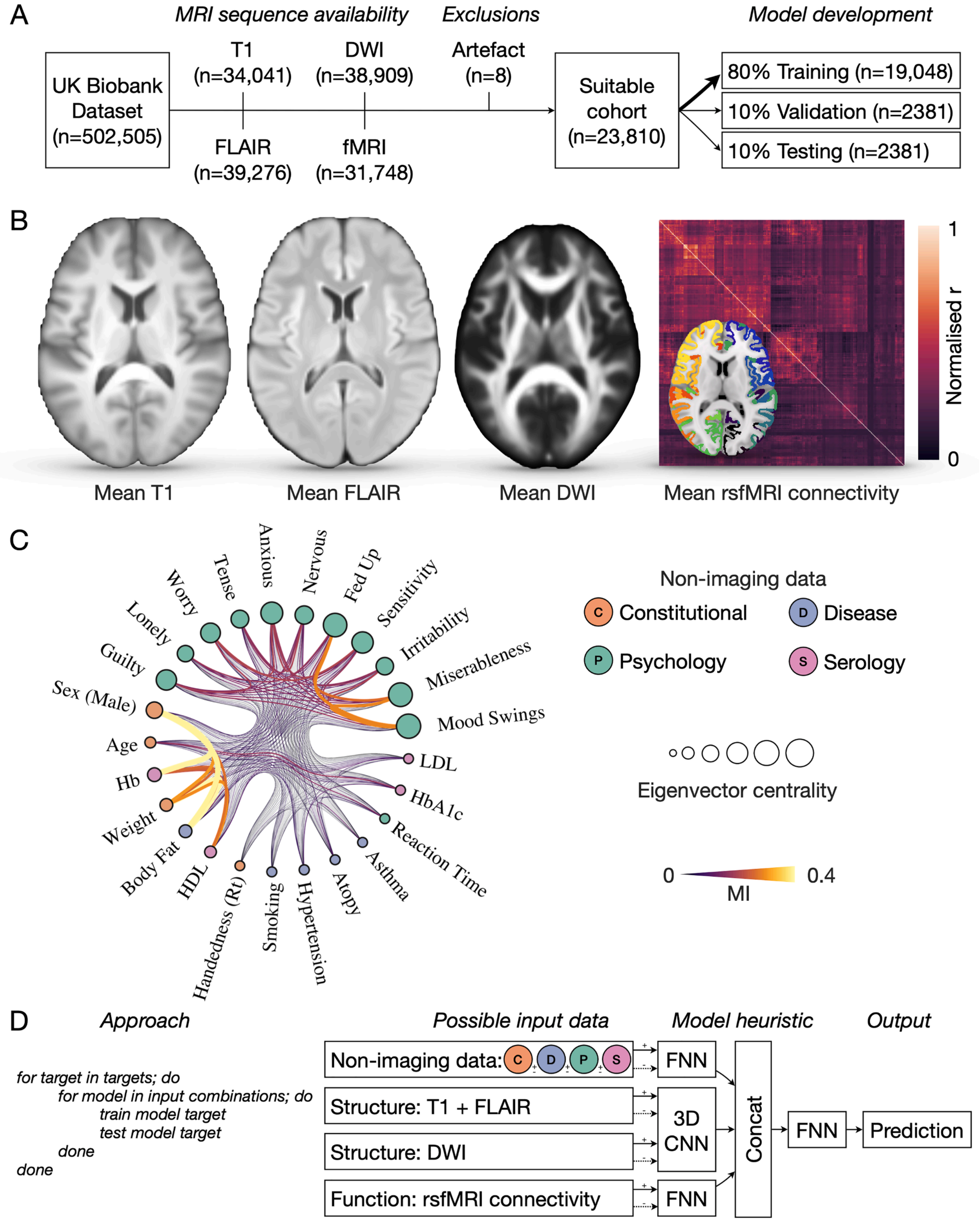
Our next task was to delineate the participants who had undergone comprehensive neuroimaging investigations inclusive of T1-weighted (T1), fluid-attenuated and inversion recovery (FLAIR), diffusion-weighted imaging (DWI), and functional magnetic resonance imaging (fMRI) sequences. Brain MRI data was available in our server for the following: FLAIR ($n = 39\ 276$), T1 ($n = 34\ 041$), DWI ($n = 38\ 909$), and



Variation in predictability

Maximum predictability		*Low*	*High*
	Low	Inadequate test	Inadequate test
	High	Maintain status quo	Regime change

Fig. 1. A 2 × 2 factorial relation between the maximum predictability of a set of characteristics and its variation across the set. A failure to achieve high fidelity for any characteristic suggests a general inadequacy of the modelling framework that casts doubt on the quality of the test. Achieving excellent fidelity across all characteristics suggests current approaches are satisfactory. Achieving high fidelity for some features but not for others, suggests a change in the modelling regime is indicated: any or all of data, model, and compute.



```
for target in targets; do
    for model in input combinations; do
        train model target
        test model target
    done
done
```

Fig. 2. Approach. A) Data selection and partitioning. B) Mean T1-weighted, FLAIR, DWI images, and rsfMRI connectivity matrix across the full cohort of 23 810 participants. C) Layered, nested, generative stochastic block model of modelling targets, with edges depicting the strength of interconnection by mutual information (MI). Node size is proportional to eigenvector centrality, a measure of node 'influence' across its network. D) Algorithmic approach for exploring the model target-feature space to distinguish targets that can be reliably predicted from those that cannot, across all possible data inputs. Shown here is also a schematic of the possible data to train with, ranging from non-imaging data across the constitutional (C) – orange, disease (D) – blue, psychology (P) – green, and serology (S) – pink feature domains; and T1/FLAIR volumetric structural imaging; DWI volumetric imaging; and rsfMRI connectivity. These data are passed to individual trainable model blocks: a fully-connected feed-forward network (FNN) for both non-imaging data, and rsfMRI connectivity, and a 3D convolutional neural network (CNN) for T1/FLAIR and/or DWI. Model block dense layers are then concatenated and passed to a final FNN for output prediction.

fMRI ($n = 31\ 748$). For participants with multiple imaging attendances, we utilised only the first MRI study to prevent an information leak. We removed 8 participants with artefact degraded MRI sequences and subsampled the cohort to include those with all four imaging dimensions of MRI data (Fig. 2). This yielded a suitable sample of 23 810 unique participants (11 141 male, 12 669 female, mean age ± standard deviation (SD) 54.775 ± 7.44 years). A breakdown of this data is provided as Table 1.

Pre-processed brain images were used from the pipeline as curated by the brain imaging leads to the UK Biobank, as described elsewhere (Alfaro-Almagro et al., 2018). All data were held in compressed NIFTI format. Formulations of these imaging sequences included the original raw data and imaging registered to Montreal Neurological Institute (MNI) template space. This pipeline relied on the validated toolkit developed by the FMRIB Oxford team and the FMRIB software library (FSL) (Smith et al., 2004). This included brain extraction (Smith, 2002), and image registration (Jenkinson et al., 2002) to the MNI152 nonlinear sixth generation standard space T1 brain template (Grabner et al., 2006). We utilised the Biobank-release pre-processed 3-dimensional volumetric T1-weighted and FLAIR structural acquisitions, the tract based spatial statistics (TBSS) pre-processed fractional anisotropy (FA) maps, and the pre-processed 4-dimensional (4D) volumetric blood oxygen level dependent (BOLD) fMRI acquisition (Alfaro-Almagro et al., 2018; Smith et al., 2006). From the pre-processed 4D time-series fMRI, we derived z-scored standardized functional connectivity matrices (Abraham et al., 2014) in accordance with the Glasser cortical parcellation scheme, comprising 360 regions of interest (180 per hemisphere) (M.F. Glasser et al., 2016). With pairwise connectivity this yields 64 620 unique edges. We did not use thresholding on the edges, for the downstream deep learning architectures would undertake feature selection intrinsically.

The rationale to reduce dimensionality of 4D time-series fMRI data to a connectome representation was twofold: firstly, since connectivity analyses now form a mainstay of neuroimaging analysis in this scientific domain (Bullmore and Sporns, 2009; Zalesky et al., 2010; Ruffle et al., 2021); and secondly, to reduce computational demand that would otherwise be infeasibly large with direct modelling of the raw fMRI time-series in what would otherwise demand a 4D convolutional neural network, infeasible for this study in scale of sample size, algorithm (GPU size), and the number of models to train. We deliberately chose not to use the mean BOLD 3D image across the 4D timeseries (using instead the aforementioned connectivity approach), as the single mean 3D image would not capture variances in haemodynamic response, and therefore was felt likely to be reductive (Logothetis et al., 2001). Moreover, it seemed prudent to use data formulations like that which is used across the bulk of neuroimaging research, which could form helpful benchmarks in model fidelity. Similarly, we deliberately chose not to pass other measures of structural connectivity, such as gray matter similarity (Raamana et al., 2015) or tractography (Ruffle et al., 2021; Thiebaut de Schotten et al., 2020; M.F. Glasser et al., 2016), as it was felt likely these data would at least in part be leveraged from the 3D T1/FLAIR and DWI images, respectively.

2.3. *Sample partitioning*

We partitioned the suitable cohort of 23 810 unique participants into training, evaluation and testing sets, using 80% of samples for training ($n = 19\ 048$), 10% for model validation ($n = 2381$) and reserving the remaining 10% ($n = 2381$) for model testing on completely unseen data (Fig. 2). This data partition was performed prior to any modelling, with the precise partition maintained for *every* experiment undertaken to ensure their comparability.

We statistically compared all modelling targets across the training, validation, and testing partitions. For continuous targets, these were compared with one-way analysis of variance (ANOVA), and for categorical targets, Chi-squared. P values were corrected for multiple

Table 1
Cohort statistics across all modelled targets within training, validation, and testing data partitions. The domain of each target is demarcated by the bracketed letter: C, constitutional; P, psychology; D, disease; S, serology. Categorical features are shown with the number of entries to each category, whilst continuous are shown with mean and 95% confidence interval. Statistical testing across training, validation, and testing sets shows there are no significant cohort differences between each partition. All p values are False-Discovery-Rate corrected.

Target (Domain)	Training ($n = 19\ 048$)	Validation ($n = 2381$)	Testing ($n = 2381$)	FDR-P value	Statistical test
Sex (C)	Female ($n = 10{,}163$), Male ($n = 8885$)	Female ($n = 1267$), Male ($n = 1114$)	Female ($n = 1239$), Male ($n = 1142$)	0.64	Chi-square
Age (C)	54.74 (54.64 - 54.85)	54.59 (54.29 - 54.89)	55.01 (54.71 - 55.3)	0.45	ANOVA
Weight (C)	76.5 (76.3 - 76.71)	76.45 (75.86 - 77.03)	77.22 (76.62 - 77.81)	0.35	ANOVA
Handedness (C)	Right ($n = 16{,}983$), Left ($n = 2065$)	Right ($n = 2101$), Left ($n = 280$)	Right ($n = 2116$), Left ($n = 265$)	0.57	Chi-square
Mood Swings (P)	No ($n = 11{,}252$), Yes ($n = 7796$)	No ($n = 1415$), Yes ($n = 966$)	No ($n = 1370$), Yes ($n = 1011$)	0.55	Chi-square
Miserableness (P)	No ($n = 11{,}306$), Yes ($n = 7742$)	No ($n = 1415$), Yes ($n = 966$)	No ($n = 1354$), Yes ($n = 1027$)	0.35	Chi-square
Irritability (P)	No ($n = 13{,}804$), Yes ($n = 5244$)	No ($n = 1760$), Yes ($n = 621$)	No ($n = 1681$), Yes ($n = 700$)	0.34	Chi-square
Sensitivity (P)	No ($n = 8998$), Yes ($n = 10{,}050$)	No ($n = 1155$), Yes ($n = 1226$)	No ($n = 1094$), Yes ($n = 1287$)	0.48	Chi-square
Fed Up (P)	No ($n = 12{,}287$), Yes ($n = 6761$)	No ($n = 1504$), Yes ($n = 877$)	No ($n = 1479$), Yes ($n = 902$)	0.34	Chi-square
Nervous (P)	No ($n = 15{,}288$), Yes ($n = 3760$)	No ($n = 1917$), Yes ($n = 464$)	No ($n = 1913$), Yes ($n = 468$)	0.99	Chi-square
Anxious (P)	No ($n = 8953$), Yes ($n = 10{,}095$)	No ($n = 1116$), Yes ($n = 1265$)	No ($n = 1120$), Yes ($n = 1261$)	0.99	Chi-square
Tense (P)	No ($n = 16{,}320$), Yes ($n = 2728$)	No ($n = 2013$), Yes ($n = 368$)	No ($n = 2022$), Yes ($n = 359$)	0.48	Chi-square
Worry (P)	No ($n = 9860$), Yes ($n = 9188$)	No ($n = 1225$), Yes ($n = 1156$)	No ($n = 1207$), Yes ($n = 1174$)	0.76	Chi-square
Lonely (P)	No ($n = 16{,}245$), Yes ($n = 2803$)	No ($n = 2033$), Yes ($n = 348$)	No ($n = 2027$), Yes ($n = 354$)	0.99	Chi-square
Guilty (P)	No ($n = 13{,}820$), Yes ($n = 5228$)	No ($n = 1701$), Yes ($n = 680$)	No ($n = 1714$), Yes ($n = 667$)	0.64	Chi-square
Reaction Time (ms) (P)	537.17 (535.78 - 538.57)	534.12 (530.34 - 537.91)	536.45 (532.54 - 540.35)	0.56	ANOVA
Smoking (D)	No ($n = 11{,}670$), Yes ($n = 7378$)	No ($n = 1482$), Yes ($n = 899$)	No ($n = 1456$), Yes ($n = 925$)	0.76	Chi-square

(*continued on next page*)

Table 1 (*continued*)

Target (Domain)	Training ($n = 19$ 048)	Validation ($n = 2381$)	Testing ($n = 2381$)	FDR-P value	Statistical test
Hypertension (D)	No ($n = 15,612$), Yes ($n = 3436$)	No ($n = 1943$), Yes ($n = 438$)	No ($n = 1940$), Yes ($n = 441$)	0.89	Chi-square
Atopy (D)	No ($n = 14,736$), Yes ($n = 4312$)	No ($n = 1807$), Yes ($n = 574$)	No ($n = 1812$), Yes ($n = 569$)	0.45	Chi-square
Asthma (D)	No ($n = 17,210$), Yes ($n = 1838$)	No ($n = 2135$), Yes ($n = 246$)	No ($n = 2131$), Yes ($n = 250$)	0.52	Chi-square
Body Fat (%) (D)	30.1 (29.99 - 30.22)	30.03 (29.71 - 30.36)	30.24 (29.91 - 30.57)	0.76	ANOVA
Hb (g/dl) (S)	14.16 (14.15 - 14.18)	14.17 (14.12 - 14.22)	14.21 (14.16 - 14.26)	0.48	ANOVA
HbA1c (mmol/mol) (S)	34.92 (34.86 - 34.99)	35.03 (34.85 - 35.22)	35.06 (34.88 - 35.23)	0.48	ANOVA
HDL (mmol/L) (S)	1.48 (1.47 - 1.48)	1.48 (1.46 - 1.49)	1.46 (1.45 - 1.47)	0.34	ANOVA
LDL (mmol/L) (S)	3.59 (3.58 - 3.6)	3.57 (3.54 - 3.61)	3.61 (3.57 - 3.64)	0.56	ANOVA

comparisons by False Discovery Rate using the Benjamini-Hochberg method (Table 1) (Benjamini and Yekutieli, 2005).

2.4. Bayesian graph representations of interactions amongst non-imaging features

To investigate the pairwise relations between the 25 variables we employed graph analysis of the training sample non-imaging data. Given the variety of variable types—continuous or categorical—we used mutual information as the primary index of similarity. This was constructed into the format of an undirected graph, wherein nodes were the targets of study, and the edges were mutual information between pairwise features – creating a graph of 25 nodes and 300 edges. The weighted eigenvector centrality of nodes was also calculated. From these data, we fitted a layered nested stochastic block model, a generative model of the community structure of graphs (Peixoto, 2015; Cipolotti et al., 2022), passing the association direction as a layer property. We further equilibrate the stochastic block model fit with Markov-Chain Monte Carlo (MCMC) simulated annealing to optimise the partition in accordance with minimizing the description length entropy criterion, as detailed elsewhere (T.P. Peixoto, 2014; Peixoto, 2012; J.K. Ruffle et al., 2023) (Fig. 2).

3. Approach

3.1. Multi-modal modelling of volume brain imaging data

We modelled T1 and FLAIR sequences for macrostructure, DWI FA for microstructure, and functional connectivity matrices derived from participant BOLD timeseries for resting state function. The non-target (for each experiment) non-imaging data were organised across the domains of constitutional, psychology, serology, and disease. It is over these seven (three imaging, four non-imaging data) domains that we could evaluate multimodal performance.

3.2. Combinatorial analysis of the imaging and target feature space

Having identified a set of unique targets and predictors from a large population, we set out to perform a systematic combinatorial analysis to determine what targets could—and just as importantly could not—be predicted by machine models, from predictors taken alone or in combination.

First, we constructed models to predict targets within the constitutional domain, i.e., participant sex, age, handedness, and weight. We examined all combinations of the imaging modalities for this case out of i) T1 + FLAIR, ii) DWI, iii) rsfMRI connectivity, and all combinations of the former, i.e., the largest model feature set would therefore be T1 + FLAIR + DWI + rsfMRI connectivity. This yielded 7 different models to fit for each of the 4 constitutional targets. These models would also serve as a benchmark to quantify the felicity of architectural choices in comparison with the extant literature.

We applied a similar approach to the targets across the psychology, serology, and disease domains, fitting models with the same imaging combinations. Constitutional data is typically collected as a standard part of a research experiment (including in neuroimaging), and used for either predictive modelling, nuisance covariates, or even as the variable of interest. To that end, we decided the available inclusion of these constitutional features to all other models of these non-constitutional targets was also reasonable. Similarly, we quantified the benefit of providing further non-imaging data from domains different to the current target, which further supplemented the number of possible imaging and non-imaging input feature sets for a given target significantly (i.e., 32 unique combinations for each target).

Overall, this process yielded 28 unique models to be trained across the constitutional targets (7 combinations * 4 targets = 28 models), and 672 unique models across the psychology, serology, and disease targets (32 combinations * 21 targets = 672 models). This yielded a total of 700 models to be trained in this experimental design.

4. Algorithm

4.1. A role for complex models

A typical volume brain image is a 182×218×182 matrix of voxels: *more than 7 million variables*. This is for a *single* – unimodal - imaging sequence. Our task is to capture complex biological and pathological traits about individuals from high-dimensional data, a task necessarily best solved by models of sufficient complexity to capture this richness. 3D convolutional neural networks (CNNs) offer a potential solution to this problem and have become state-of-the-art for modelling brain imaging data across numerous tasks (Bakas et al., 2018; Isensee et al., 2021; Jonsson et al., 2019; J.K. Ruffle et al., 2023).

4.2. Remediating target class imbalances

Class imbalance was handled by randomly sub-sampling the majority class, performed at the beginning of each epoch. No under- or over-sampling was applied to the validation of test partitions, but performance metrics were always balanced to accommodate class imbalances.

4.3. Data pre-processing and augmentation

For each model target, continuous variables were clamped between the 0.5th and 99.5th percentile, z-scored, and normalized to the range −1 to 1. The targets this applied to were age, weight, haemoglobin, reaction time, body fat, HBA1c, HDL, and LDL. Categorical targets were one-hot encoded. The reasoning behind re-scaling continuous targets into a −1 to 1 range space (for example, as opposed to modelling age in years), was so that all models across different continuous targets with different native ranges were more optimally comparable in both loss function and evaluation metric.

For associated non-imaging data (e.g., any combination of constitutional, psychology, serology, and disease features), the selected combination was first passed to the code, with any unused non-imaging data (including that of the same block to the model target) zeroed. For example, in training a model to predict anxiety, using constitutional and serological non-imaging data, all other psychological data would be zeroed since anxiety were within this target domain, but disease data would also be zeroed as it was not selected to be passed to the model. The reasoning behind this was to 1) ensure the prevention of an information leak between similar feature domains (e.g., it seemed probable one could fit a function of anxiety from a selection of other measures of psychology), but 2) would maintain precisely the same modelling architectural complexity, only where some features were encoded as zero.

For structural neuroimaging, we developed a comprehensive MRI augmentation pipeline using MONAI (Consortium, 2020; Pinaya et al., 2023). This pipeline included the following: image resizing; ii) clamping along the 0.5th and 99.5th percentile; iii) intensity normalisation; intensity scaling to the range of −1 to 1; iv) random histogram shifting; v) random intensity scaling; vi) random affine transformations; vii) random 3D elastic deformations; vii) image re-normalization and viii) image re-scaling to the range of −1 to 1. All random transformations were with a probability of application of 0.1.

To evaluate the discrepancy (if any) in model fidelity at differential imaging resolutions, we considered two image resolutions to resize data to. We firstly fitted the 700 models with resizing images from native 182×218×182, isotropic and volumetric, 1 mm (Elliott et al., 2018) voxel dimensions to a smaller 64×64×64 isotropic resolution. Doing so would enable models to train significantly faster and illuminate where performant signal for a given target could be extracted from the structural 3D sequences. This enabled training one model on four Tesla P100 16Gb GPUs within a DGX environment with a batch size of 64. Having trained these models, we identified the best sets of input combinations for each of the 25 targets and retrained each (i.e., the best model, per target), with resizing images from native 182×218×182, isotropic and volumetric, 1 mm (Elliott et al., 2018) voxel dimensions to a 128×128×128 isotropic resolution. This enabled training one model on eight Tesla P100 16Gb GPUs within a DGX environment with a batch size of 32.

Like the associated non-imaging data pipeline, where a given MRI sequence was not chosen as a data input to a given target, it was zeroed. For example, when training a model to predict sex with T1+FLAIR, the DWI channel was zeroed. The rationale for this was to maintain the same model size/architecture regardless of the data passed to it.

For functional connectivity, we used the bilateral Glasser parcellation of 360 regions (180 per hemisphere) to extract a given regions BOLD time-series signal, and cross-correlate to a symmetrical adjacency matrix of shape 360×360, using nilearn (Abraham et al., 2014). Data were standardised by z-score transformation. For each participant, the upper triangle of this symmetrical rsfMRI connectivity matrix was extracted and flattened to a 1D array, with 64 620 functional connections between each pairwise set of regions. Like the remaining pipeline, where connectivity was not selected as an input to a given model, it was zeroed. We opted for the Glasser parcellation scheme (M.F. Glasser et al., 2016) since it is one of the most widely used and cited (3965 tracked Google Scholar citations as of 01/03/2024). We did consider the use of other functional templates to boost the analysis further, however since the computing requirements for all modelling was already substantial, we did not pursue it further. This could however be explored in future research.

5. Architecture

All deep learning aspects of the study were undertaken using PyTorch (Paszke et al., 2019) and MONAI (Consortium, 2020), with model architectures as listed below.

5.1. Feed-forward neural network for non-imaging data

For modelling with non-imaging data, we constructed a feed-forward neural network (FNN), which took an input dimension of 24, the number of non-imaging data features minus 1 (the target), with sequential dense layers of 128, 64 and 32 with sequential batch normalisation (Ioffe et al., 2015), Gaussian error linear unit (GELU) activation (Hendrycks and Gimpel, 2016), and dropout (rate 0.1) (Srivastava et al., 2014). The non-imaging data FNN comprised 13 984 parameters.

5.2. Three-dimensional convolutional neural networks for volumetric imaging data

We developed a 3D convolutional neural network (CNN) architecture for modelling with volume brain imaging data (Goodfellow et al., 2017; Krizhevsky et al., 2012). This CNN was contained three channels, for T1, FLAIR and DWI. The architecture followed the sequence of 3D convolution, GELU hidden activation (Hendrycks and Gimpel, 2016), skip convolution layers (He et al., 2016), batch normalization (Ioffe et al., 2015), GELU hidden activation (Hendrycks and Gimpel, 2016), max 3D pooling (Yamaguchi et al., 1990), dropout (rate 0.1) (Srivastava et al., 2014), flattening to a linear dense layer, batch normalisation (Ioffe et al., 2015), and GELU output activation (Hendrycks and Gimpel, 2016). For training with 64×64×64 images, the CNN channel sizes were 32, 64, 128, 256, 256 with a further final output channel of size 128. For training with 128×128×128 images, the CNN channel sizes were 32, 64, 128, 256, 256, 256 with a further final output channel of size 256. Our architectural design and channel sizes were guided by review of existing literature, and benchmark comparisons to open-source datasets inclusive of MNIST and MNIST-fashion (He et al., 2016; Heinz, 2018; Benchmarks, 2021). Skip convolutions have been shown advantageous to well-known models such as that of ResNet, with a building block composed of two convolution layers and activation operators, then concatenated (He et al., 2016; Ahn and Yim, 2020). The CNN comprised 7 427 680 parameters for 64×64×64 resolution models, and 11 295 968 for 128×128×128 models.

5.3. Feed-forward neural network for functional connectivity

For modelling with functional connectivity, we constructed a FNN which took an input dimension of 64 620, the number of unique pairwise functional connections in the flattened connectivity matrix, with sequential dense layers of 128 and 128, each with batch normalisation (Ioffe et al., 2015), GELU activation (Hendrycks and Gimpel, 2016), and dropout (rate 0.5) (Srivastava et al., 2014). This rsfMRI connectivity FNN comprised 8 288 512 parameters.

5.4. Model concatenation and feed-forward neural network for final prediction

Outputs from the non-imaging data FNN, multi-channel 3D-CNN connectivity, and rsfMRI connectivity FNN, were concatenated and used for a final FNN for target prediction. This took the sum of the output channels from above (288 when training with 64×64×64 MRI, 416 when training with 128×128×128), with further dense layers of size 256, 256, and a final layer of size 2 where the target was categorical and one-hot encoded (e.g., sex), or 1 where the target was continuous (e.g., age). Similar to the non-imaging data and connectivity FNN models, these used sequential batch normalisation (Ioffe et al., 2015), GELU activation (Hendrycks and Gimpel, 2016), and dropout (rate 0.1) (Srivastava et al., 2014). When training with an MRI resolution of 64×64×64, this final FNN comprised 141 057 parameters, and 173 825 when training in 128×128×128.

The total parameter count for this multi-dimensional modelling architecture was 15 871 233 when training with an MRI resolution of 64×64×64, and 19 772 289 when training with an MRI resolution of

128×182×128. The full modelling architecture can be visualised in Supplementary Figure 2.

6. Hyperparameters

All models were compiled with a learning rate of 0.0001, the Adam optimiser (Kingma and Ba, 2017), L_2 regularisation (Cortes et al., 2012), a batch size of 64 for 64×64×64 resolution MRI models, and 32 for 128×128×128 (limited only by GPU size). Models were permitted to train for anywhere up to 100 000 epochs, but with early stopping if there was failure to improve the validation loss function after 50 epochs. For categorical targets, the loss function was binary cross entropy, and for continuous, mean squared error.

6.1. Model evaluation

Models were always trained on the training data only and evaluated at the end of each epoch with the validation set. The best performing model, criterion on the loss function, were saved. After completion of model training, the best performing epoch for each model (on validation data) were used for evaluation of performance on the completely unseen test data. Numerous performance metrics were derived, including accuracy, precision, recall, F1, a confusion matrix (Trimarchi, 2019), the receiver operator characteristic (ROC) curve, the R^2 and r for continuous targets, the model loss, and the amount of time each model required to train (Varoquaux and Colliot, 2023; Poldrack et al., 2020). Metrics of categorical performance were always balanced by macro average to accommodate for any degree of class imbalance. It should be noted that we place our focus here on evaluating categorical models with metrics appropriate for any degree of class imbalance, such as with balanced accuracy or macro-averaged precision, recall, and F1. However, we also provide AUROC as there are a high proportion of research articles that only report this despite its clear limitation to sample imbalance, such that our work is still comparable to the broader literature. To enable large-scale comparison across all 700 models, irrespective of if a categorical or continuous target, we also converted the task to categorical with continuous targets divided by median split with respect to the training dataset. We determined the 'best' performing model by the highest balanced accuracy in the testing set for categorical target, and by the highest R^2 for continuous targets (Varoquaux and Colliot, 2023; Poldrack et al., 2020).

6.2. Model validation with open-source datasets

To ensure that the performance of our CNN model architectures was comparable to existing state-of-the-art performances, we prototyped CNNs initial with both MNIST and the MNIST-fashion (data not shown).

6.3. Model comparison

6.3.1. Linear mixed-effects models

After fitting all possible model combinations, we undertook post-hoc comparisons of models, reviewing the performance metrics to identify both human factors inherently predictable by imaging data, but also the value of data components in fitting each factor. This was undertaken by visual inspection of all performance metrics, including ROC curves and confusion matrices.

We conducted formal statistical comparison of model performance with linear mixed-effect models (Pinheiro et al., 2022; Wickham et al., 2019). These were in the following formulation:

$$\begin{aligned}\textit{Model performance} \sim\ & T1 + FLAIR + DWI + rsfMRI\ Connectivity \\ & + Psychology + Disease + Serology + (1 \mid Target),\end{aligned}$$

where for constitutional, psychological and disease targets, model performance were by the balanced accuracy given the majority were categorical in formulation, but included conversion of continuous targets by median split, as described above. Whereas for serology targets, model performance was R^2.

6.3.2. Graph representations

We derived graph representations of model performances across all targets, across all possible feature combinations. This was undertaken by fitting an undirected graph of all targets as nodes, wherein the edge weights were calculated as the inverse of the Euclidean distance between all performance metrics for a given model (T.P. Peixoto, 2014). We also used this data to derive the weighted eigenvector centrality of each node, weighted by the similarity across different performance metrics, that might further provide insight in explaining the similarities and dissimilarities across model performances. Lastly, we converted these results into a fully interactive HTML object (Haas, 2021) to enable reader visualisation.

7. Compute

7.1. Hardware

Local development and prototyping were predominantly performed on a 32 core (64 thread) CPU Linux workstation housing 135Gb of RAM and an NVIDIA 2080Ti GPU (11Gb size), OS Ubuntu 20.04. All model training was undertaken on a DGX workstation housing 8 x P100 16Gb GPUs, 80 CPU threads and 503Gb of RAM.

7.2. Software

Most of the programming was undertaken within a Python environment (version 3.6.9). Further small operations were completed with Bash for faster IO enabled by GNU parallel (Tange, 2011). The following Python packages were utilised: graph-tool (T.P. Peixoto, 2014), gravis (Haas, 2021), matplotlib (Hunter, 2007), minepy (Albanese et al., 2018), MONAI (Consortium, 2020), nibabel (Brett et al., 2020), NumPy (Harris et al., 2020), pandas (Reback and McKinney, 2020), PyTorch (Paszke et al., 2019), scikit-learn (Pedregosa et al., 2011), SciPy (Virtanen et al., 2020), seaborn (Waskom, 2020), statsmodels (Seabold and Perktold, 2010). GPU-modelling was achieved with the CUDA toolkit version 11.0 (Developers, 2021). Linear mixed effect models were constructed within R version 4.1.2, using packages Tidyverse (Wickham et al., 2019), and nlme (Pinheiro et al., 2022).

7.3. Ethical approval

The study was approved by local institutional review board and conducted in accordance with the "Declaration of Helsinki". Use of UK Biobank data were approved under study application identifier #16,273.

7.4. Code, model, and data availability

All code and models are openly available online at https://github.com/high-dimensional/biobank-megamodeller.git. Supplementary code evaluating age and sex prediction using other models (Cole, 2020; Peng et al., 2021; Gong et al., 2021) is also available. All data is available from the UK Biobank curators.

8. Results

8.1. Cohort

We studied an unselected sample of 23 810 unique UK Biobank participants (11 141 male, 12 669 female, mean age ± 95% confidence interval (CI) 54.775 years (54.66 – 54.85)) who underwent multisequence MRI, including T1-weighted, fluid-attenuated inversion recovery (FLAIR), diffusion weighted imaging (DWI), and resting-state functional MRI (rsfMRI) acquisitions, and for whom a set of 25 constitutional, psychological, disease, and serological domain variables of plausible clinical or scientific interest was available. A compact set of characteristics was chosen to enable comprehensive modelling of the comparative predictability of subsets of variables from the remainder, a task that scales exponentially with the number of subsets. Participants were randomly partitioned into training, validation, and testing sets of 19 048, 2381, and 2381 unique participants, respectively, with no significant differences between them (Fig. 2, Table 1).

8.2. *Graph community representations of non-imaging characteristics*

A generative stochastic block model (Peixoto, 2018), with separate layers (Peixoto, 2015) for positively and negatively related characteristics across the four non-imaging domains, was used to derive a succinct hierarchical representation of the non-imaging data in terms of its patterns of distinct covariance, captured by the graph model as hypergraph ‘community’ structure. The largest community consisted of psychological measures. The second community consisted of sex, age, haemoglobin (Hb), weight, percentage body fat, and high-density lipoprotein (HDL). Male gender was associated with haemoglobin (mutual information (MI) 0.30) and weight (MI 0.18), and inversely related to with body fat (MI 0.31) and HDL (0.14). The final community, consisting of low-density lipoprotein (LDL), HbA1c, reaction time, asthma, atopy, hypertension, smoking, and handedness, was characterised by weak mutual information across all features (Fig. 2). A stochastic block model employing a non-linear index of dependence based on the maximal information coefficien (Reshef et al., 2011), revealed essentially the same structure (Supplementary Figure 1).

8.3. *Imaging models of constitutional characteristics achieve state-of-the-art performance*

Across constitutional characteristics, models of FLAIR, T1, and DWI achieved state-of-the-art sex classification (balanced accuracy (BA) 99.7%, area under the receiver operator characteristic curve (AUROC) > 0.999); age regression (R^2 of 0.859, mean absolute error (MAE) of 2.048 years); and weight regression (R^2 of 0.625, MAE of 7.042 kg) performance (Fig. 3). Models of rsfMRI connectivity alone performed the worst on these characteristics but yielded the best prediction of handedness (BA 57.7%, AUROC of 0.915). Our brain age model outperforms previous top prediction models including that of Peng et al., (Peng et al., 2021) (MAE reported in manuscript=2.14 years, MAE from evaluation on our test set=5.282 years), and Cole et al., (Cole, 2020) (MAE reported in manuscript=3.55 years, MAE from evaluation on our test set=5.115 years). Similarly, our sex classifier outperformed previous top prediction model from Peng et al., (Peng et al., 2021) (accuracy reported in manuscript=99.5%, accuracy from evaluation on our test set=96.6%).

To quantify the relative contribution of each imaging feature we employed a linear mixed-effect model predicting balanced accuracy from the choice of imaging inputs. This showed the inclusion of T1/FLAIR structural sequences to be significantly advantageous to model performance (coefficient 0.041, 95% CI 0.012 to 0.0694, $p = 0.008$) (Fig. 4). There were non-significant trends for the inclusion of both DWI (coefficient 0.0211, 95% CI −0.007 to 0.050, $p = 0.138$) and rsfMRI connectivity (coefficient 0.012, 95% CI −0.017 to 0.040, $p = 0.408$).

8.4. *Psychological characteristics are poorly predicted by imaging*

Models predicting psychological characteristics exhibited poor test-set performance (Fig. 3), in the face of extensive optimisation. Models of non-imaging data *alone,* offered the best prediction fidelity in 10 of the 12 psychological characteristics. Specifically, constitutional non-imaging data *alone* best predicted sensitivity (BA 60.1%, AUROC 0.614), and guilt (BA 58.1%, AUROC 0.659). Models of constitutional and disease non-imaging data best predicted loneliness (BA 57.5%, AUROC 0.685), irritability (BA 57.5%, AUROC 0.628), and the propensity to feel tense (BA 55.8%, AUROC 0.667). Constitutional and serological non-imaging data best predicted anxiety (BA 59.0%, AUROC 0.607). Lastly, models featuring constitutional, serological, and disease non-imaging data best predicted nervousness (BA 57.3%, AUROC 0.650), feeling fed up (BA 57.7%, AUROC 0.649), miserableness (BA 60.5%, AUROC 0.650), and mood swings (balanced accuracy 56.6%, AUROC 0.611). Indeed, only the propensity to worry and reaction time were best predicted with the inclusion of any neuroimaging data. Worry was best predicted with DWI, FLAIR, T1, and constitutional non-imaging data (BA 58.8%, AUROC 0.601), whereas reaction time was best predicted with DWI, constitutional and disease non-imaging data (R^2 of 0.081, MAE of 70.316 ms).

A linear mixed-effect model predicting balanced accuracy from the choice of T1/FLAIR, DWI, rsfMRI connectivity, serology, and disease data for all 32 model input combinations for the 12 psychology targets (384 models) found disease significantly advantageous to model performance (coefficient 0.004, 95% CI 0.001 to 0.007, $p = 0.021$) (Fig. 4).

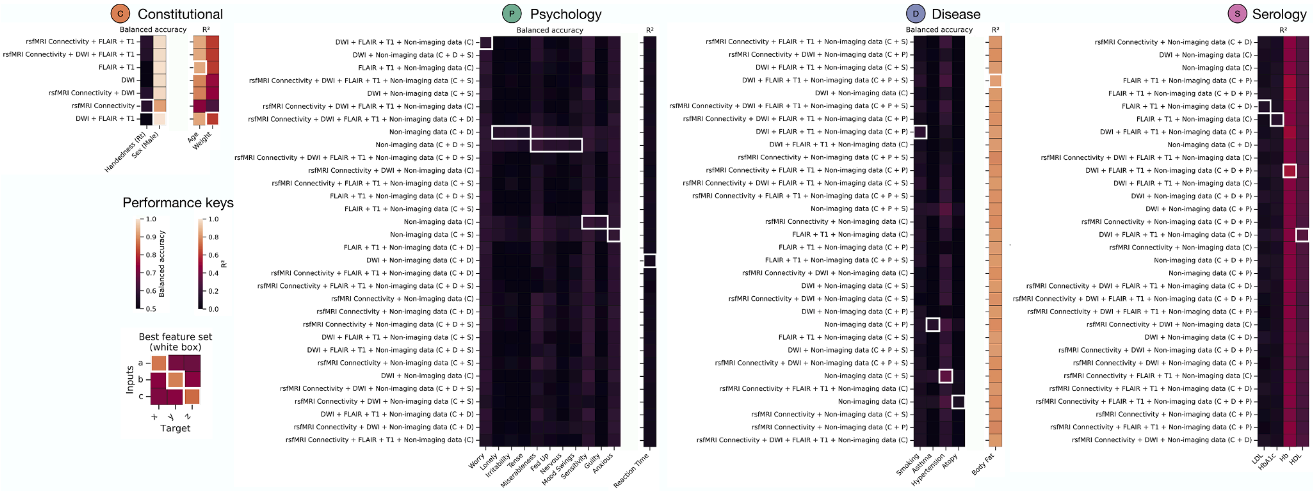


Fig. 3. Model performances. Test set performance for all models across constitutional (C) – orange, disease (D) – blue, psychology (P) – green, and serology (S) – pink feature domains. Index of performance is given as balanced accuracy for classification targets and R^2 for regression fits. The x-axis of all heatmaps depicts the model target, and y-axis depicts the range of feature inputs. White boxes demarcate the best set of inputs to achieve the greatest out-of-sample model performance.

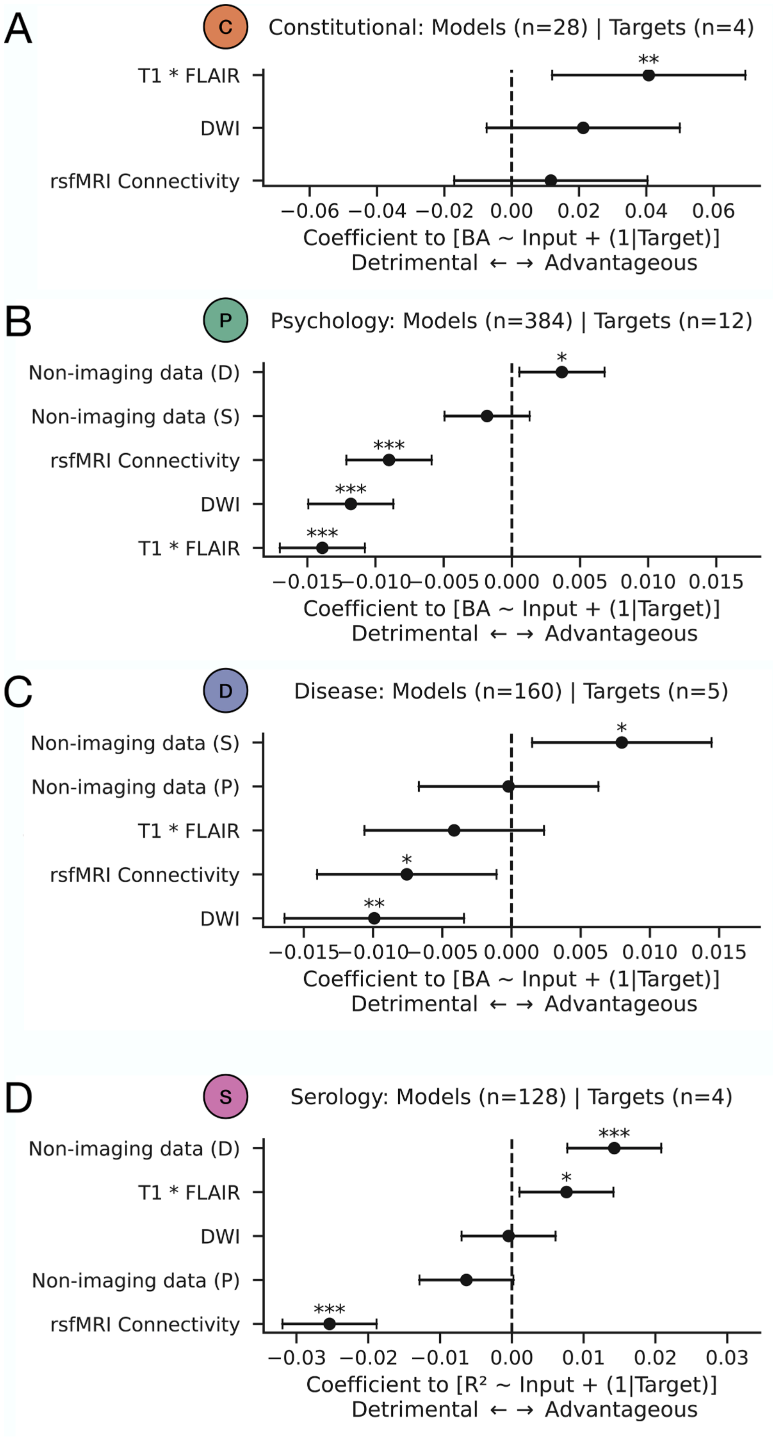


Fig. 4. Domain-specific effects. Linear mixed-effects models for predicting out of sample performance (balanced accuracy or R^2, where applicable) from structural imaging, functional imaging, and non-imaging domain feature sets. Shown are coefficient plots for models whose targets are A) constitutional, B) psychology, C) disease, and D) serology. Inputs with coefficients whose values are positive are associated with increase model performance (advantageous), whilst features with negative coefficients are associated with weaker performance (detrimental). Asterisks stipulate statistical significance as per standard convention: * denotes $p < 0.05$; ** denotes $p < 0.01$; *** denotes $p < 0.001$.

The inclusion of T1/FLAIR (coefficient −0.014, 95% CI −0.017 to −0.011, $p < 0.001$), DWI (coefficient −0.012, 95% CI −0.015 to −0.009, $p < 0.001$) and rsfMRI connectivity (coefficient −0.009, 95% CI −0.012 to −0.006, $p < 0.001$) were all significantly associated with detrimental model performance. There was no significant relationship between serological non-imaging data and model performance ($p = 0.252$).

8.5. *Disease characteristics are best predicted by serological data, not neuroimaging*

Test performance for models predicting individual disease characteristics was more variable, revealing 3 of 5 disease targets to be best predicted from non-imaging data alone. These were: 1) a diagnosis of atopy, using constitutional non-imaging data alone (BA 54.4%, AUROC 0.656), 2) asthma, using constitutional and psychological non-imaging data (BA 57.8%, AUROC 0.608), and 3) hypertension, using constitutional and serological non-imaging data (BA 66.7%, AUROC 0.774). Smoking was best predicted by DWI, FLAIR, T1, constitutional and psychological non-imaging data (balanced accuracy 58.1%, AUROC 0.682), and percentage body fat using DWI, FLAIR, T1, constitutional, psychological, and serological non-imaging data (R^2 of 0.834, r of 0.914, and MAE of 2.653%) (Fig. 3).

A linear mixed-effect model predicting balanced accuracy from the inclusion of T1/FLAIR, DWI, rsfMRI connectivity, serology, and psychology data for all 32 model input combinations for the 5 disease targets (160 models) found serology significantly advantageous to model performance (coefficient 0.008, 95% CI 0.002 to 0.015, $p = 0.016$) (Fig. 4). The inclusion of DWI (coefficient −0.010, 95% CI −0.016 to −0.003, $p = 0.003$) and rsfMRI connectivity (coefficient −0.008, 95% CI −0.014 to −0.001, $p = 0.023$) were significantly associated with detrimental model performance. There was a non-significant trend for the inclusion of T1/FLAIR imaging to also be detrimental (coefficient −0.004, 95% CI −0.011 to 0.002, $p = 0.209$). There was no significant relationship between psychology non-imaging data and model performance ($p = 0.950$).

8.6. *Models of serology perform best with multi-modal imaging and non-imaging data*

Serological targets were best predicted from variable combinations of imaging and non-imaging data. The best performing haemoglobin model included FLAIR, T1, and DWI sequences, with constitutional, psychological, and disease non-imaging data, achieving an R^2 of 0.524, r of 0.725, and MAE of 0.629 g/dl. HDL was best predicted by FLAIR, T1, and DWI sequences, augmented with both constitutional and disease non-imaging data, achieving an R^2 of 0.309, r of 0.556, and MAE of 0.209 mmol/L. Prediction of HbA1c was weaker, the best feature combination of which was FLAIR, T1, and constitutional non-imaging data, achieving an R^2 of 0.146, r of 0.394, and MAE of 2.790 mmol/mol. LDL concentration was similarly weak, though the best performing model utilised FLAIR, T1, constitutional and disease non-imaging data, achieving an R^2 of 0.126, r of 0.355, and MAE of 0.584 mmol/L (Fig. 3).

A linear mixed-effect model predicting R^2 from the inclusion of T1/

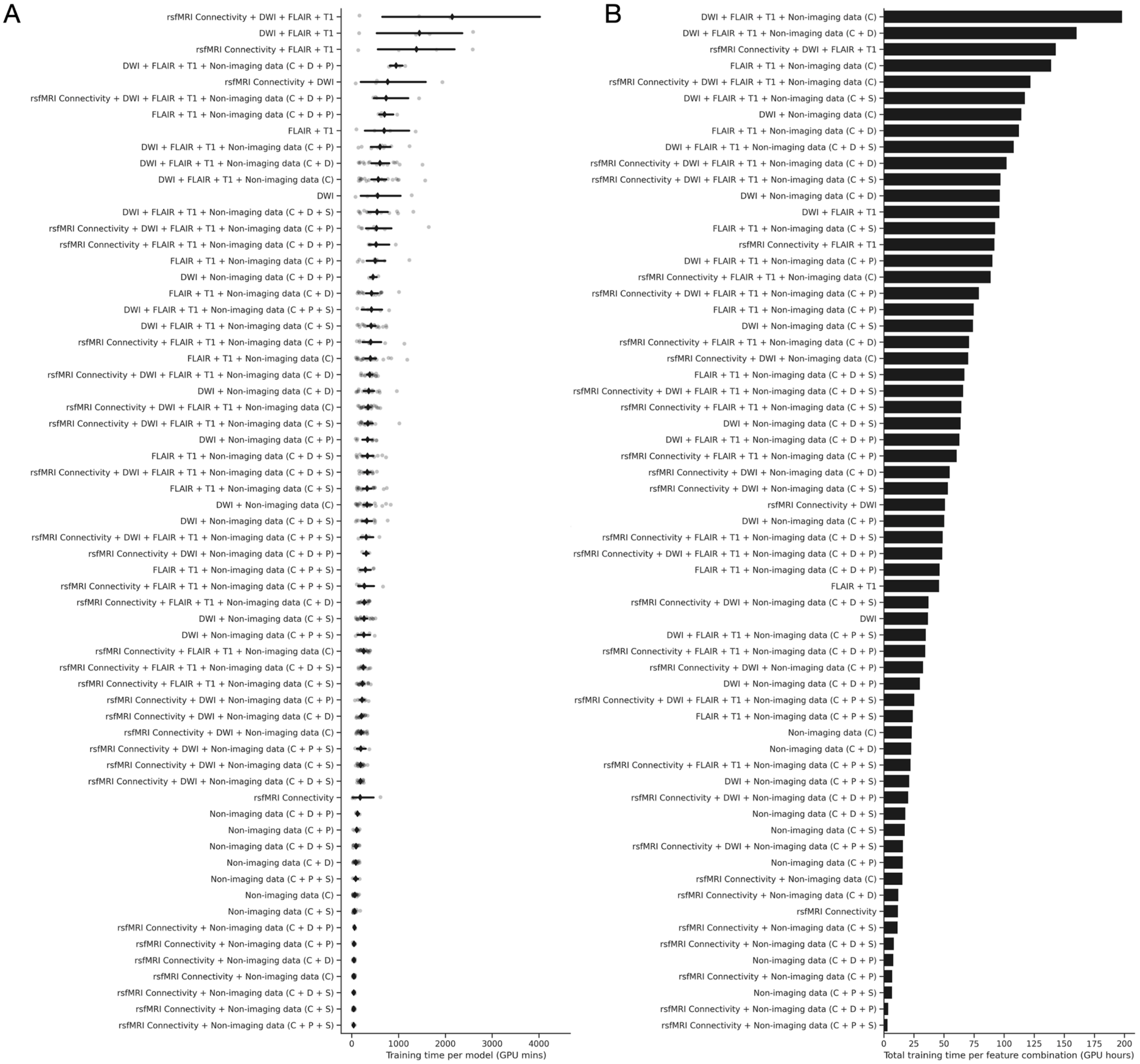


Fig. 5. Time costs in training large medical imaging models. A) Strip plot illustrates training time taken per model in GPU minutes (x-axis) for all possible feature input combinations (y-axis). Grey points indicate individual models, with mean shown as a black diamond, and 95% confidence interval shown as a black line. B) Bar plot of total training time in GPU hours for all feature input combinations. Only 64×64×64 resolution models are shown here for visual simplicity.

FLAIR, DWI, rsfMRI connectivity, disease, and psychology data for all 32 model input combinations for the 4 serology targets (128 models) found the inclusion of both disease (coefficient 0.014, 95% CI 0.008 to 0.021, $p < 0.001$) and T1/FLAIR (coefficient 0.008, 95% CI 0.001 to 0.014, $p = 0.023$) significantly advantageous to model performance (Fig. 4). Conversely, rsfMRI connectivity was significantly associated with detrimental model performance (coefficient −0.025, 95% CI −0.032 to −0.019, $p < 0.001$). There was no significant relationship between DWI and model performance ($p = 0.891$).

8.7. *The trade-off between computational time and performance*

The total time required to train and optimize all models (both low and high resolution) was 188.589 P100 (16Gb) GPU days (4526.135 GPU hours). The total time required to train all 700 individual 64×64×64 resolution models was 146.281 GPU days (3510.734 GPU hours). The total time required to train the 25 individual 128×128×128 resolution models was 42.308 GPU days (1015.401 GPU hours).

For training 3D volumetric 64×64×64 imaging models, mean training time was 300.920 min (95% CI 279.316 to 322.524 min) (Fig. 5). As expected, models that only included rsfMRI connectivity and/or non-imaging data took much less time to train (anywhere from just a few minutes for rsfMRI connectivity alone, to up to 123 min for constitutional, disease and psychology non-imaging data models. Models incorporating 3D volumetric imaging required far longer training times, up to a mean of 1412.750 min for predicting constitutional targets with rsfMRI connectivity, T1, FLAIR, and DWI. We cite training times for two reasons. First, they indicate the computational requirements for training uni- and/or multimodal deep models, including with multi-channel 3D imaging. Second, they show that the model performance reported here is unlikely to be trivially constrained by available compute, and more likely reflects the nature of the data and architectural limitations.

8.8. *Graph relationships of model performance*

Finally, we created an undirected graph to visualise the similarities and differences of the model targets in terms of their predictability from different inputs (Fig. 6). This showed that, whilst pairwise interrelation of participant features generally linked constitutional serological features, whereas psychological were highly interrelated (Fig. 6A-B), pairwise interrelation linked by predictive fidelity revealed a segregation of constitutional features from those serological, with interrelation between features of disease and psychology, This graph is also available as a fully interactive, customizable, and downloadable HTML object (Supplementary Material). We also provide tabular data of performance metrics for all 700 models in the supplementary material.

9. Discussion

In the most comprehensive published analysis of its kind, we have quantified the individual-level predictability of 25 different constitutional, psychological, chronic disease, and serological characteristics, drawing on four different neuroimaging modalities spanning both structural and functional domains, and involving all possible combinations of features. The comparative performance of 700 models, trained over 189 GPU days (4526.135 GPU hours), with large-scale data from 23 810 individuals, casts light on the limits to prediction under a practicably ideal modelling regime: large-scale data, state-of-the-art model architectures, and high-performance compute.

9.1. *Comparative legibility of biological features*

Our study seeks to identify the limit on achievable individual-level predictive fidelity within the prevailing data regime. To do this, we are obliged to use state-of-the-art methods, for any shortfall in performance could otherwise be attributed to a remediable deficiency in the model architecture. Our models of constitutional features set new state-of-the-art benchmarks for age and sex prediction from neuroimaging data (Peng et al., 2021; Gong et al., 2021). Though models combining T1, FLAIR and DWI performed best, achieving balanced accuracy of 99.7% for sex and MAE of 2.048 years for age, even rsfMRI connectivity alone achieved a balanced accuracy of 91.5% for sex, a new state-of-the-art for non-structural imaging (Leming and Suckling, 2021).

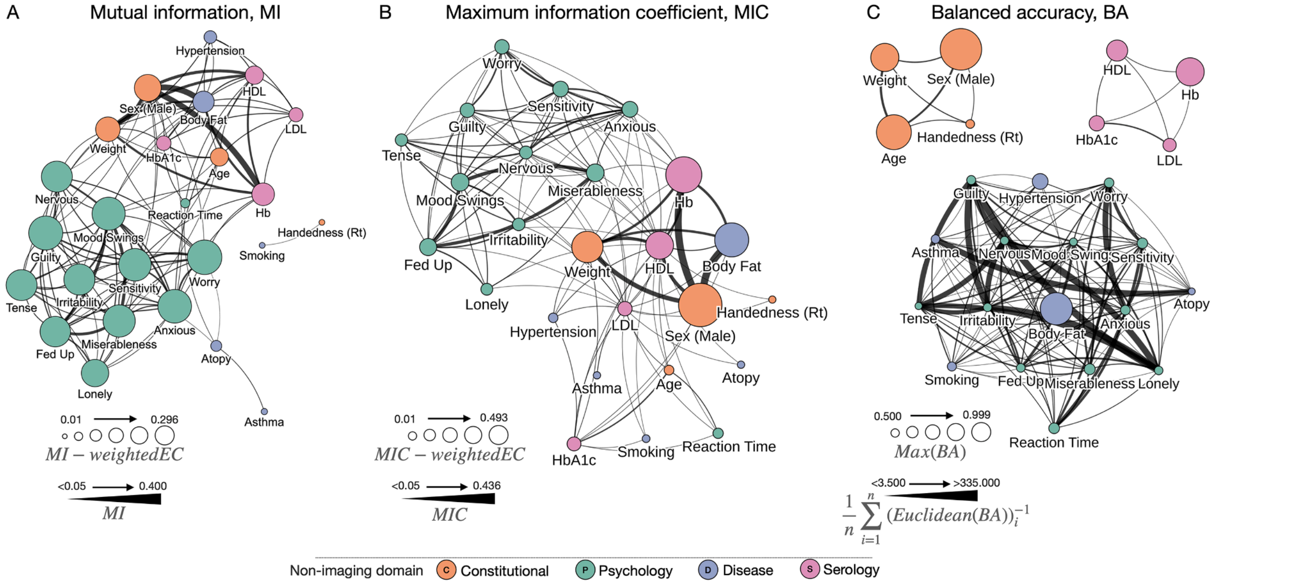


Fig. 6. Static visual network analysis plots of feature relationships and model performances. A) Graph of target features, with nodes sized by mutual information (MI)-weighted eigenvector centrality (EC), and edges sized according to pairwise MI. Eigenvector centrality is a measure of influence of a node across a network. B) Graph of target features, with nodes sized by the maximum information coefficient (MIC)-weighted eigenvector centrality (EC), and edges sized according to the MIC. C) Graph of target features, with nodes sized by the maximum balanced accuracy across all models (BA), with edges sized according to the mean inverse Euclidean distance of all input combinations between each pair of targets. For all panels we depict the top 60% of edges for visualisation purposes. *Note that all graphs are made available as fully interactive and customizable HTML objects within the supplementary material.*

Equally, a MAE of 7.04Kg indicates remarkably high fidelity for weight, a hitherto under-explored task here rendered maximally challenging by the exclusion of all non-brain tissue prior to modelling. These results show that our model architectures and overall analytic pipeline faithfully reflect the highest standards in the field, supporting a rigorous test of the current limits on the individual predictability of other features. That the remaining constitutional feature—handedness, best predicted using rsfMRI connectivity data alone—achieved a balanced accuracy of only 57.7% suggests the difficulty here does not arise from inadequate implementation of current technology. It is striking, even accounting for class imbalance, that handedness is individually so poorly predictable given the magnitude of population-level differences in the organisation of the brain (Sha et al., 2021; Chormai et al., 2022; Good et al., 2001).

Psychological targets showed equally limited predictability, maximal from disease data. The addition of any neuroimaging, whether structural or functional, generally offered no material benefit. Of the 12 psychological targets, only the propensity to worry and reaction time showed some effect from neuroimaging, but only in the context of low overall performance. This is again strikingly at odds with population-level observations, where marked group-level differences are often reported, but individual level predictability is relatively low.

Disease targets, focused on common conditions without gross, diagnostic imaging changes, were also poorly predictable at the individual level, with serology the strongest predictor. Only percentage body fat benefited from imaging (DWI, FLAIR and T1), achieving MAE of 2.65%, but not far from age, sex, and weight alone (3.01%).

Serology was best predicted by disease, followed by T1/FLAIR structural imaging. Haemoglobin offered the best—and most multimodal—performance, drawing on DWI, FLAIR, T1, constitutional, disease, and psychological data (MAE 0.629 g/dl), but its prediction is likely to lean on covarying sex and age. Although we did not explicitly test the question, the difference from constitutional data alone (MAE 0.676 g/dl) is likely explained by global differences in the MRI signal.

These performance figures were essentially invariant to modelled image resolution. Across all 25 model targets, smoking was the only one to demonstrate an increase in model fidelity when training at a higher imaging resolution (128×128×128). The gain, however, was marginal: 58.1% rising to 59.2%. Under current data and architectural regimes, meaningful improvements in performance are unlikely to be achieved merely by increasing image resolution.

9.2. The limits of individual prediction

Our analysis shows that whereas constitutional characteristics—age, sex, and weight—are highly predictable from neuroimaging, psychology, chronic illness, and serological characteristics are not. Crucially, *comparative* differences in predictability are extraordinarily high, suggesting that with currently practicable models, the limits are primarily set by the fundamental informativity of imaging signals. Substantially higher individual resolving power will require either a radical ‘regime change’—in terms of volumes of data, model expressivity, and compute—or new investigational methods.

Three implications are foremost. First, imaging-based clinical decision-support systems with cognitive or behavioural targets—*operating in the absence of overt changes on imaging*—will likely continue to be plagued by underperformance, especially when deployed in real-world scenarios. If fundamental psychological characteristics are illegible under a modelling regime far more conducive to success than clinical realities ordinarily permit, the prospects of such endeavours seem dim. Performance in specific clinical populations inadequately sampled by population-based studies such as UK Biobank (Littlejohns et al., 2020), especially those with overtly abnormal imaging, may well be higher, but the bar is clearly set high.

Second, population-level mechanistic models of cognition and behaviour that seek to ground theories of brain function in terms of normal structural and resting-state functional features will likely leave most individual variability unexplained (Bzdok et al., 2020). Though many in the field consider explanatory and predictive power to be decoupled, the claim to fidelity of a theory contradicted by most instances it describes is bound to strike a disinterested observer as insecure. Were residual variability truly random when inspected at finer scales of observation, with more powerful methods such as intracranial recording, then a case could be made that the observed stochasticity is more aleatoric than epistemic. But such studies near-universally reveal complex structure a suitably expressive model could conceivably capture.

Third, the manifest difficulty of prediction mandates obtaining not just the largest but also the widest possible data support, operating multi-modally, with resilience to the missingness and noise corruption ubiquitous in the real world. This implies a decisive shift not only away from simple models but also from unimodal discriminative models of any complexity. Only multimodal generative models drawing on all available data, complete and incomplete, could adequately corroborate the belief any observed underperformance is irremediable without new or higher quality known biological signals (Pinaya et al., 2023).

9.3. Limitations

Our inferences are supported by the largest and most comprehensive evaluation of its kind. Nonetheless, the data and modelling context give rise to an array of limitations.

First, we chose to model 25 non-imaging characteristics from a far wider range of data available in UK Biobank (Littlejohns et al., 2020; Alfaro-Almagro et al., 2018; Sudlow et al., 2015; Bycroft et al., 2018). A modest number is inevitable where, as here, the objective is to probe the effect of multiple *combinations* of characteristics, with large-scale data and comprehensive model optimization. Indeed, the task of training 700 independent deep learning models of this kind is already onerous enough not to have been previously attempted in this domain. The choice of as wide a range of characteristics as was feasible for the compute at our disposal is deliberate, for it allows us to evaluate the predictive contribution of signals distributed across *multiple* characteristics. Not doing this could have raised the possibility that a given unimodal signal may be present, but camouflaged by multi-modal contextual modulation. Note that each individual characteristic is comprehensively modelled in any event, and the multimodal perspective is an addition, not a substitution.

Second, it is conceivable that architectural—as opposed to hyperparameter—tuning may have obtained better performance for any one individual target. But our objective here was to standardize the model architecture across all possible targets and feature combinations, creating a general-purpose prediction pipeline that enables a fair comparison of the distinct contribution of each input. Moreover, reasoning that non-constitutional targets may require large numbers of parameters, we employed more flexible architectures than current age and sex classifiers (Cole, 2020; Peng et al., 2021). Note the flexibility was not such as to induce overfitting on constitutional targets, so potential overfitting on other targets is not explained by excessive overparameterisation.

Third, it is possible that aspects of the processing upstream of the predictive modelling may have an impact on performance. Such variabilities are common across neuroimaging research (and indeed form a focus of other research groups (Fusar-Poli et al., 2010; Haddad et al., 2023; Zhou et al., 2022)). The same is true for our use of widely available resting-state (as opposed to task-related) fMRI data, or use of other complimentary imaging modalities such as EEG (Chowdhury et al., 2020), where further improvement in model performance is foreseeable. But here we adopt common pre-processing practices (Alfaro-Almagro et al., 2018) and parcellations (M.F. Glasser et al., 2016; M.F. Glasser et al., 2016), and it seems unlikely that the striking differences in performance observed here, especially with state-of-the-art constitutional performance, are thereby explained. Task-based fMRI may reveal greater predictive fidelity in some tasks, especially when compared to resting state. We however deliberately did not include such data for our priority was to maximise generalizability, for the nature of a task-based functional scan could vary greatly across any research study or imaging site.

Fourth, UK Biobank's cohort, though peerlessly large for phenotyping of this richness, is explicitly limited to an older age group and implicitly limited by the time and dispositional demands of participation. The observed variations of the features of interest are, however, both generous and comparable to those likely to obtain in real-world contexts. Equally, the quality and instrumental homogeneity of the source data may theoretically be exceeded elsewhere, even if there is no superior study currently in progress. The critical question we address in this article is the limit on individual-level fidelity imposed by current data regimes, a task that necessitates the largest and richest available unselected dataset, of which UK Biobank is internationally the leading example. While generalisability is definitionally impossible to assure universally, this is as close as anyone could plausibly get at present.

Fifth, the predictive targets are deliberately chosen to exclude those diagnosable from imaging (e.g., acute stroke), for the task then becomes one of recognition rather than prediction (Farazi and Nogga, 2021). Although chronic diseases of high prevalence are included, our focus is on characteristics common enough to span the normal/abnormal divide, at least in statistical terms. This focus reflects the scale of potential population-level benefit in illuminating individual-level patterns of the underlying substrates and processes as reflected in the imaged brain.

Finally, psychological characteristics can only be imperfectly captured by test instruments whose reliability is bound to vary, both across characteristics and datasets. Better tests may, of course, provide targets with higher achievable fidelity. But the variations in observed reliability (Fawns-Ritchie and Deary, 2020) are not so large as to trivially explain the striking differences in comparative predictability here, and the chosen tests are known to correlate reasonably well with other measures (Wu et al., 2022; He et al., 2020).

10. Conclusion

In the largest study of its kind, involving 700 models trained on a comprehensive set of combinations of 25 target biological features, across multiple domains and 23 810 unique participants, we have quantified the individual-level legibility of the human brain. Determining the *comparative* predictability of different targets from each other and from multimodal brain imaging, under the current practical maximum of data quality, algorithmic felicity, and computational resource, we set out to answer a key strategic question: is actionable individual-level predictive fidelity plausibly achievable under current data regimes, or is a radical change necessary? The striking difference in observed comparative predictability suggests the latter, interpretative limitations notwithstanding. If predictive systems are to achieve the individual-level fidelity clinical utility demands, and if mechanistic models are to capture enough variability in the population to be persuasively generalizable, regime change is now unavoidable.

Data sharing

All code and models are openly available at https://github.com/high-dimensional/biobank-megamodeller.git. All data is provided courtesy of the UK Biobank – https://www.ukbiobank.ac.uk.

CRediT authorship contribution statement

James K. Ruffle: Writing – review & editing, Visualization, Validation, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Writing – original draft. **Robert J Gray:** Formal analysis, Software, Writing – review & editing. **Samia Mohinta:** Formal analysis, Software, Writing – review & editing. **Guilherme Pombo:** Software, Writing – review & editing. **Chaitanya Kaul:** Software, Writing – review & editing. **Harpreet Hyare:** Supervision, Writing – review & editing. **Geraint Rees:** Data curation, Funding acquisition, Writing – review & editing. **Parashkev Nachev:** Writing – review & editing, Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft.

Declaration of competing interest

None to declare.

Data availability

See data and code availability statement

Funding

JKR was supported by the Medical Research Council (MR/X00046X/1) and the Guarantors of Brain. PN is supported by the Wellcome Trust (213038/Z/18/Z) and the UCLH NIHR Biomedical Research Centre.

Acknowledgement

We are grateful to the curators of the UK Biobank for providing this data resource.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2024.120600.

References

Abraham, A., et al., 2014. Machine learning for neuroimaging with scikit-learn. Front. Neuroinform. 8, 14. https://doi.org/10.3389/fninf.2014.00014.

Ahn, H., Yim, C., 2020. Convolutional Neural Networks Using Skip Connections with Layer Groups for Super-Resolution Image Reconstruction Based on Deep Learning. Applied Sciences 10, 1959.

Albanese, D., Riccadonna, S., Donati, C., Franceschi, P., 2018. A practical tool for maximal information coefficient analysis. Gigascience 7. https://doi.org/10.1093/gigascience/giy032.

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. https://doi.org/10.1016/j.neuroimage.2017.10.034.

Bakas, S., et al., 2018. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. ArXiv. abs/1811.02629.

Bazinet, V., et al., 2023. Assortative mixing in micro-architecturally annotated brain connectomes. Nat. Commun. 14, 2850. https://doi.org/10.1038/s41467-023-38585-4.

Benchmarks, A.I. MNIST, https://benchmarks.ai/mnist(2021).

Benjamini, Y., Yekutieli, D., 2005. False Discovery Rate–Adjusted Multiple Confidence Intervals for Selected Parameters. J. Am. Stat. Assoc. 100, 71–81. https://doi.org/10.1198/016214504000001907.

Bethlehem, R.A.I., et al., 2022. Brain charts for the human lifespan. Nature 604, 525–533. https://doi.org/10.1038/s41586-022-04554-y.

Brett, Matthew, Markiewicz, Hanke, C, 2020. nipy/nibabel: 3.2.1 (Version 3.2.1). Zenodo. https://doi.org/10.5281/zenodo.4295521.

Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci. 10, 186–198. https://doi.org/10.1038/nrn2575.

Bycroft, C., et al., 2018. The UK Biobank resource with deep phenotyping and genomic data. Nature 562, 203–209. https://doi.org/10.1038/s41586-018-0579-z.

Bzdok, D., Engemann, D., Thirion, B., 2020. Inference and Prediction Diverge in Biomedicine. Patterns 1, 100119. https://doi.org/10.1016/j.patter.2020.100119.

Chormai, P., et al., 2022. Machine learning of large-scale multimodal brain imaging data reveals neural correlates of hand preference. Neuroimage 262, 119534. https://doi.org/10.1016/j.neuroimage.2022.119534.

Chowdhury, M.S.N., et al., 2020. Deep Neural Network for Visual Stimulus-Based Reaction Time Estimation Using the Periodogram of Single-Trial EEG. Sensors. (Basel) 20. https://doi.org/10.3390/s20216090.

Cipolotti, L., et al., 2022. Graph lesion-deficit mapping of fluid intelligence. Brain 146, 167–181. https://doi.org/10.1101/2022.07.28.501722.

Cole, J.H., 2020. Multimodality neuroimaging brain-age in UK biobank: relationship to biomedical, lifestyle, and cognitive factors. Neurobiol. Aging 92, 34–42. https://doi.org/10.1016/j.neurobiolaging.2020.03.014.

Consortium, T.M., 2020. Project MONAI. Zenodo. https://doi.org/10.5281/zenodo.4323059.

Cortes, C., Mohri, M. & Rostamizadeh, A. L2 Regularization for Learning Kernels. (2012). https://ui.adsabs.harvard.edu/abs/2012arXiv1205.2653C.

Developers, N. CUDA Toolkit 11.0, https://developer.nvidia.com/cuda-11.0-download-archive(2021).

Elliott, L.T., et al., 2018. Genome-wide association studies of brain imaging phenotypes in UK Biobank. Nature 562, 210–216. https://doi.org/10.1038/s41586-018-0571-7.

Farazi, H. & Nogga, J. Semantic Prediction: Which One Should Come First, Recognition or Prediction?, (2021), https://ui.adsabs.harvard.edu/abs/2021arXiv211002829F.

Fawns-Ritchie, C., Deary, I.J., 2020. Reliability and validity of the UK Biobank cognitive tests. PLoS. One 15, e0231627. https://doi.org/10.1371/journal.pone.0231627.

Finn, E.S., et al., 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat. Neurosci. 18, 1664–1671. https://doi.org/10.1038/nn.4135.

Fischl, B., et al., 2008. Cortical Folding Patterns and Predicting Cytoarchitecture. Cerebral Cortex 18, 1973–1980. https://doi.org/10.1093/cercor/bhm225.

Fusar-Poli, P., et al., 2010. Effect of image analysis software on neurofunctional activation during processing of emotional human faces. J. Clin. Neurosci. 17, 311–314. https://doi.org/10.1016/j.jocn.2009.06.027.

Glasser, M.F., et al., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178. https://doi.org/10.1038/nature18933.

Glasser, M.F., et al., 2016. The Human Connectome Project's neuroimaging approach. Nat. Neurosci. 19, 1175–1187. https://doi.org/10.1038/nn.4361.

Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., Peng, H., 2021. Optimising a Simple Fully Convolutional Network for Accurate Brain Age Prediction in the PAC 2019 Challenge. Front. Psychiatry 12, 627996. https://doi.org/10.3389/fpsyt.2021.627996.

Good, C.D., et al., 2001. Cerebral asymmetry and the effects of sex and handedness on brain structure: a voxel-based morphometric analysis of 465 normal adult human brains. Neuroimage 14, 685–700. https://doi.org/10.1006/nimg.2001.0857.

Goodfellow, I., Bengio, Y., Courville, A., 2017. Deep Learning. MIT Press.

Grabner, G. et al. in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006. (eds Rasmus Larsen, Mads Nielsen, & Jon Sporring) 58–66 (Springer Berlin Heidelberg).

Haas, R, 2021. gravis. https://github.com/robert-haas/gravis.

Haddad, E., et al., 2023. Multisite test-retest reliability and compatibility of brain metrics derived from FreeSurfer versions 7.1, 6.0, and 5.3. Hum. Brain Mapp. 44, 1515–1532. https://doi.org/10.1002/hbm.26147.

Hansen, J.Y., et al., 2021. Mapping gene transcription and neurocognition across human neocortex. Nat. Hum. Behav. 5, 1240–1250. https://doi.org/10.1038/s41562-021-01082-z.

Hansen, J.Y., et al., 2022. Mapping neurotransmitter systems to the structural and functional organization of the human neocortex. Nat. Neurosci. 25, 1569–1581. https://doi.org/10.1038/s41593-022-01186-3.

Harris, C.R., et al., 2020. Array programming with NumPy. Nature 585, 357–362. https://doi.org/10.1038/s41586-020-2649-2.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

He, T., et al., 2020. Deep neural networks and kernel regression achieve comparable accuracies for functional connectivity prediction of behavior and demographics. Neuroimage 206, 116276. https://doi.org/10.1016/j.neuroimage.2019.116276.

Heinz, S. A performance benchmark of Google AutoML Vision using Fashion-MNIST, https://towardsdatascience.com/a-performance-benchmark-of-google-automl-vision-using-fashion-mnist-a9bf8fc1c74f(2018).

Hendrycks, D., Gimpel, K., 2016. Gaussian Error Linear Units (GELUs). ArXiv.

Honey, C.J., et al., 2009. Predicting human resting-state functional connectivity from structural connectivity. Proceedings of the National Academy of Sciences 106, 2035–2040. https://doi.org/10.1073/pnas.0811168106.

Hunter, J.D., 2007. Matplotlib: A 2D Graphics Environment. Comput. Sci. Eng. 9, 90–95. https://doi.org/10.1109/MCSE.2007.55.

Ioffe, S. & Szegedy, C. in Proceedings of the 32nd International Conference on Machine Learning Vol. 37 (eds Bach Francis & Blei David) 448–456 (PMLR, Proceedings of Machine Learning Research, 2015).

Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods 18, 203–211. https://doi.org/10.1038/s41592-020-01008-z.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17, 825–841.

Jonsson, B.A., et al., 2019. Brain age prediction using deep learning uncovers associated sequence variants. Nat. Commun. 10, 5409. https://doi.org/10.1038/s41467-019-13163-9.

Kingma, D.P., Ba, J.Adam, 2017. A Method for Stochastic Optimization. ArXiv. 1412, 6980.

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. ImageNet Classification with Deep Convolutional Neural Networks. Adv. Neural Inf. Process. Syst. 25.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539.

Leming, M., Suckling, J., 2021. Deep learning for sex classification in resting-state and task functional brain networks from the UK Biobank. Neuroimage 241, 118409. https://doi.org/10.1016/j.neuroimage.2021.118409.

Littlejohns, T.J., et al., 2020. The UK Biobank imaging enhancement of 100,000 participants: rationale, data collection, management and future directions. Nat. Commun. 11, 2624. https://doi.org/10.1038/s41467-020-15948-9.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A., 2001. Neurophysiological investigation of the basis of the fMRI signal. Nature 412, 150–157. https://doi.org/10.1038/35084005.

Marek, S., et al., 2022. Reproducible brain-wide association studies require thousands of individuals. Nature 603, 654–660. https://doi.org/10.1038/s41586-022-04492-9.

Paszke, A., et al., 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. NeurIPS.

Pedregosa, F., Varoquaux, G., Gramfort, A., 2011. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825–2830.

Peixoto, T.P., 2012. Entropy of stochastic blockmodel ensembles. Physical Review E 85, 056122. https://doi.org/10.1103/PhysRevE.85.056122.

Peixoto, T.P., 2014. Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. Physical Review E 89, 012804. https://doi.org/10.1103/PhysRevE.89.012804.

Peixoto, T.P., 2014. The graph-tool python library. figshare. https://doi.org/10.6084/m9.figshare.1164194.

Peixoto, T.P., 2015. Inferring the mesoscale structure of layered, edge-valued, and time-varying networks. Physical Review E 92, 042807. https://doi.org/10.1103/PhysRevE.92.042807.

Peixoto, T.P., 2018. Nonparametric weighted stochastic block models. Physical Review E 97, 012306. https://doi.org/10.1103/PhysRevE.97.012306.

Peng, H., Gong, W., Beckmann, C.F., Vedaldi, A., Smith, S.M., 2021. Accurate brain age prediction with lightweight deep neural networks. Med. Image Anal. 68, 101871 https://doi.org/10.1016/j.media.2020.101871.

Pinaya, W.H.L., et al., 2023. Generative AI for Medical Imaging: extending the MONAI Framework. arXiv e-prints. https://doi.org/10.48550/arXiv.2307.15208.

Pinheiro, J., et al., 2022. Package 'nlme'. cran 1–328.

Poldrack, R.A., Huckins, G., Varoquaux, G., 2020. Establishment of Best Practices for Evidence for Prediction: A Review. JAMa Psychiatry 77, 534–540. https://doi.org/10.1001/jamapsychiatry.2019.3671.

Raamana, P.R., Weiner, M.W., Wang, L., Beg, M.F., 2015. Thickness network features for prognostic applications in dementia. Neurobiol. Aging 36, S91–S102. https://doi.org/10.1016/j.neurobiolaging.2014.05.040.

Reback, J., McKinney, W., 2020. jbrockmendel. pandas-dev/pandas: Pandas 1.0.3 (Version v1.0.3). Zenodo. https://doi.org/10.5281/zenodo.3715232.

Reshef, D.N., et al., 2011. Detecting novel associations in large data sets. Science (1979) 334, 1518–1524. https://doi.org/10.1126/science.1205438.

Richards, B.A., et al., 2019. A deep learning framework for neuroscience. Nat. Neurosci. 22, 1761–1770. https://doi.org/10.1038/s41593-019-0520-2.

Ruffle, J.K., et al., 2021. The autonomic brain: Multi-dimensional generative hierarchical modelling of the autonomic connectome. Cortex 143, 164–179. https://doi.org/10.1016/j.cortex.2021.06.012.

Ruffle, J.K., et al., 2023. Brain tumour genetic network signatures of survival. Brain.

Ruffle, J.K., Mohinta, S., Gray, R., Hyare, H., Nachev, P, 2023. Brain tumour segmentation with incomplete imaging data. Brain Commun. https://doi.org/10.1093/braincomms/fcad118.

Schulz, M.A., et al., 2020. Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. Nat. Commun. 11, 4238. https://doi.org/10.1038/s41467-020-18037-z.

Seabold, S., Perktold, J.Statsmodels, 2010. Econometric and Statistical Modeling with Python. In: Proc of the 9th Python in science conference.

Sha, Z., et al., 2021. Handedness and its genetic influences are associated with structural asymmetries of the cerebral cortex in 31,864 individuals. Proc. Natl. Acad. Sci. u S. a 118. https://doi.org/10.1073/pnas.2113095118.

Smith, S.M., et al., 2004. Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 (Suppl 1), S208–S219. https://doi.org/10.1016/j.neuroimage.2004.07.051.

Smith, S.M., et al., 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. Neuroimage 31, 1487–1505. https://doi.org/10.1016/j.neuroimage.2006.02.024.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155. https://doi.org/10.1002/hbm.10062.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958.

Suárez, L.E., Markello, R.D., Betzel, R.F., Misic, B., 2020. Linking Structure and Function in Macroscale Brain Networks. Trends. Cogn. Sci. 24, 302–315. https://doi.org/10.1016/j.tics.2020.01.008.

Sudlow, C., et al., 2015. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS. Med. 12, e1001779 https://doi.org/10.1371/journal.pmed.1001779.

Szucs, D., Ioannidis, J.P., 2020. Sample size evolution in neuroimaging research: An evaluation of highly-cited studies (1990-2012) and of latest practices (2017-2018) in high-impact journals. Neuroimage 221, 117164. https://doi.org/10.1016/j.neuroimage.2020.117164.

Tange, O., 2011. GNU Parallel - The Command-Line Power Tool. The USENIX Magazine 42–47.

Thiebaut de Schotten, M., Foulon, C., Nachev, P., 2020. Brain disconnections link structural connectivity with function and behaviour. Nat. Commun. 11, 5094. https://doi.org/10.1038/s41467-020-18920-9.

Thomas Yeo, B.T., et al., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106, 1125–1165. https://doi.org/10.1152/jn.00338.2011.

Trimarchi, D. Confusion Matrix, https://github.com/DTrimarchi10/confusion_matrix (2019).

Varoquaux, G. & Colliot, O. in Machine Learning for Brain Disorders (ed Olivier Colliot) 601–630 (Springer US, 2023).

Virtanen, P., et al., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

Wang, C., et al., 2022. Phenotypic and genetic associations of quantitative magnetic susceptibility in UK Biobank brain imaging. Nat. Neurosci. 25, 818–831. https://doi.org/10.1038/s41593-022-01074-w.

Waskom, M., 2020. Seaborn_Development_Team. seaborn. Zenodo. https://doi.org/10.5281/zenodo.592845.

Wickham, H., et al., 2019. Welcome to the Tidyverse. J. Open. Source Softw. 4.

Wu, J., et al., 2022. Cross-cohort replicability and generalizability of connectivity-based psychometric prediction patterns. Neuroimage 262, 119569. https://doi.org/10.1016/j.neuroimage.2022.119569.

Wu, J., Li, J., Eickhoff, S.B., Scheinost, D., Genon, S., 2023. The challenges and prospects of brain-based prediction of behaviour. Nat. Hum. Behav. 7, 1255–1264. https://doi.org/10.1038/s41562-023-01670-1.

Yamaguchi, K., Sakamoto, K., Akabane, T., Fujimoto, Y., 1990. A Neural Network for Speaker-Independent Isolated Word Recognition. ICSLP 90, 1077–1080.

Zalesky, A., Fornito, A., Bullmore, E.T., 2010. Network-based statistic: identifying differences in brain networks. Neuroimage 53, 1197–1207. https://doi.org/10.1016/j.neuroimage.2010.06.041.

Zhou, X., et al., 2022. Choice of Voxel-based Morphometry processing pipeline drives variability in the location of neuroanatomical brain markers. Commun. Biol. 5, 913. https://doi.org/10.1038/s42003-022-03880-1.