Predicting Future Depressive Episodes from Resting-State fMRI with Generative Embedding

Herman Galioulline[1*], Stefan Frässle[1], Sam Harrison[1], Inês Pereira[1], Jakob Heinzle[1‡], Klaas Enno Stephan[1,2‡]

[1] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich & ETH Zurich, 8032 Zurich, Switzerland.

[2] Max Planck Institute for Metabolism Research, Cologne, Germany.

Keywords: depression, prediction, early detection, generative embedding, UK Biobank, translational neuromodeling, computational psychiatry

Running title: Predicting Depressive Episodes with Generative Embedding

[*] Corresponding author: galioulline@biomed.ee.ethz.ch

‡ shared last authorship

Abstract

After a first episode of major depressive disorder (MDD), there is substantial risk for a long-term remitting-relapsing course. Prevention and early interventions are thus critically important. Various studies have examined the feasibility of detecting at-risk individuals based on out-of-sample predictions about the future occurrence of depression. However, functional magnetic resonance imaging (MRI) has received very little attention for this purpose so far.

Here, we explored the utility of generative models (i.e. different dynamic causal models, DCMs) as well as functional connectivity (FC) for predicting future episodes of depression in never-depressed adults, using a large dataset (N=906) of task-free ("resting state") fMRI data from the UK Biobank. Connectivity analyses were conducted using timeseries from pre-computed spatially independent components of different dimensionalities. Over a three year period, 50% of participants showed indications of at least one depressive episode, while the other 50% did not. Using nested cross-validation for training and a held-out test set (80/20 split), we systematically examined the combination of 8 connectivity feature sets and 17 classifiers. We found that a generative embedding procedure based on combining regression DCM (rDCM) with a support vector machine (SVM) enabled the best predictions, both on the training set (0.63 accuracy, 0.66 area under the curve, AUC) and the test set (0.62 accuracy, 0.64 AUC; $p<0.001$). However, on the test set, rDCM was only slightly superior to predictions based on FC (0.59 accuracy, 0.61 AUC). Interpreting model predictions based on SHAP (SHapley Additive exPlanations) values suggested that the most predictive connections were widely distributed and not confined to specific networks. Overall, our analyses suggest (i) ways of improving future fMRI-based generative embedding approaches for the early detection of individuals at-risk for depression and that (ii) achieving accuracies of clinical utility may require combination of fMRI with other data modalities.

Introduction

Major depressive disorder (MDD) causes tremendous personal suffering and, amongst all medical conditions, has one of the highest burden of disease globally (GBD 2019 Mental Disorders Collaborators, 2022; Vos et al., 2020). It has a profoundly negative impact on social and occupational functions (Adler et al., 2006; Kupferberg et al., 2016) and is associated with increased risk for other mental and somatic (e.g. cardiovascular) disorders . After the onset of a first episode of MDD, there is a substantial risk for a long-term remitting-relapsing course (Eaton et al., 2008), accompanied by prolonged trial-and-error treatment attempts (Correll et al., 2017; Steffen et al., 2020). Prevention and early interventions are thus crucial for reducing the burden of MDD, both at an individual and societal level (Cuijpers et al., 2012, 2021). The challenge is to detect at-risk individuals early so that preventive measures and interventions can be administered in a timely and targeted fashion.

Detecting at-risk individuals requires prediction models that enable out-of-sample predictions about the future occurrence of (symptoms of) depression with clinically adequate accuracy. In the recent past, there have been numerous attempts to establish such models both in adolescents and adults, based on combinations of various data types, e.g. demographic, socioeconomic, cognitive, and clinical variables as well as motor activity (Caldirola et al., 2022; Chikersal et al., 2021; Gu et al., 2020; King et al., 2008; Librenza-Garcia et al., 2021; Lin et al., 2022; Na et al., 2020; Rocha et al., 2021; Rosellini et al., 2020; Sampson et al., 2021; van Eeden et al., 2021; Voorhees et al., 2008; Xu et al., 2019).

Neuroimaging has played a minor role in this endeavour so far. This may be partly due to difficulties of obtaining datasets that are longitudinal in nature and sufficiently large to allow for robust out-of-sample predictions. Several longitudinal magnetic resonance imaging (MRI) studies of depressive symptoms do exist (e.g. Barch et al., 2019; Pagliaccio et al., 2014; Papmeyer et al., 2016; Shapero et al., 2019), but almost all have small to moderate sample sizes and employ within-sample association analyses. However, association is not prediction: prediction requires out-of-sample analyses, i.e. " ... testing of the model on data separate from those used to estimate the model's parameters" (Poldrack et al., 2020). A recent exception is the study by Toenders et al. (2021) which predicted depression onset out-of-sample, based on structural MRI (and other) data from a large sample of 544 adolescents. Concerning functional MRI (fMRI), however, we are aware of only one previous fMRI study (Hirshfeld-Becker et al., 2019) that has attempted out-of-sample predictions of future depressive episodes in hitherto depression-free individuals, albeit with a small sample (total N=33). The predictive value of fMRI for identifying individuals at risk for future depression is thus not well known.

One might wonder why fMRI should be considered at all for establishing predictor models of depressive episodes, given that fMRI data are more difficult to obtain and more costly than many other types of measurements? There are several reasons why fMRI – and particularly generative models for estimating connectivity – may have particular utility for clinical predictions. First, fMRI may afford high sensitivity since it assesses the functional status quo of neural circuits (Stephan et al., 2015), the biological level that is closest to psychiatric symptoms (Gordon, 2016). Second, clinical predictions are most valuable if they afford a mechanistic interpretation (Stephan et al., 2017); for example, this may guide the development of novel treatments. Analyses of functional interactions based on fMRI can potentially give insights into circuit mechanisms that increase risk for depression. Ideally, this requires generative models which offer an explanation how activity distributed throughout a circuit could have been generated (Stephan et al., 2015) and provide estimates of effective (directed) connectivity.

An approach that blends generative modeling with prediction is "generative embedding" (GE) (Brodersen et al., 2011, 2014; Frässle et al., 2020; Stephan et al., 2017). GE uses parameter estimates of a system (circuit) of interest, obtained by inverting a generative model, as features for subsequent machine learning (ML). This often improves prediction accuracy since the parameter estimates of a generative model offer a low-dimensional, de-noised representation of neural dynamics. Furthermore, provided the generative model is biologically plausible, GE may reveal which biological processes or properties (e.g. specific connections in a neural circuit) are most relevant for successful clinical predictions.

In this study, we used a large dataset (N=906) of task-free ("resting state") fMRI data from the UK Biobank (Miller et al., 2016) to explore the utility of fMRI-based connectivity measures for predicting future episodes of depression in never-depressed adults. Over a three year follow-up period, half of the selected participants (N=453) exhibited at least one indicator of depression, according to clinical records and/or self-report, while the other half remained free from depression. Both groups were carefully matched with regard to 7 potentially confounding variables (age, sex, handedness, tobacco, alcohol, illicit drugs, cannabis).

We emphasise that the goal of this work was not to test whether predictions based on fMRI data are better or worse than predictions based on other data types, e.g. socioeconomic or clinical variables. Instead, because there are numerous options of utilising fMRI for predictive analyses, this initial study focused on fMRI only and assessed the relative performance of different connectivity approaches – including generative embedding based on different variants of dynamic causal modeling (DCM)[13] as well as functional connectivity (FC) – for predicting future depressive episodes. Concretely, in our training set (N=724), we systematically combined different connectivity

121 approaches with different ML classifiers, using nested cross-validation, and tested how well they
122 predicted the occurrence of at least one indicator of a depressive episode over a follow-up period of
123 three years. We then used the best-performing combination to make the same prediction in a held-
124 out test set (N=182) that was completely independent from the training data. Notably, predicting
125 the occurrence of indicators of depressive episodes represents a more challenging scenario than
126 predicting a full clinical diagnosis of MDD. Our study can thus be seen as a "stress test" whether
127 fMRI-based assessments of connectivity, and generative models in particular, are likely to be useful
128 at all for early detection of at-risk individuals.

Materials and Methods

The following sections describe the dataset and methodology used in this study. Briefly, the data consist of task-free fMRI measurements (i.e. unconstrained cognition or "resting state") and questionnaire data from the UK Biobank (www.ukbiobank.ac.uk). Based on entries in UK Biobank, we selected participants that had good quality fMRI recordings and consistent questionnaire information that allowed us to assign them to one of two groups: a group that initially had no signs of depressive symptoms but exhibited indicators of depressive episodes (e.g. questionnaire data, prescription of antidepressants) within three years after the fMRI session (D+ group), or a control group that did not show any such indicators during the same period (D- group).

We used different connectivity metrics (different variants of DCM as well as functional connectivity, FC) in combination with different ML classifiers for prediction of future indicators of depressive episodes. DCM and FC analyses were applied to time series of rs-fMRI networks (with 6, 21, or 55 nodes) defined by independent components analysis (ICA) of the preprocessed "resting-state" fMRI (rs-fMRI) data and provided by UK Biobank. Posterior parameter estimates (DCM) and Pearson correlation coefficients (FC), respectively, served as input features to various discriminative classifiers. The classifiers were trained using nested cross-validation to avoid overfitting and to provide the best possible estimate of generalizability. Finally, the best models were chosen, and a prediction was made on held-out (and completely independent) test data.

It is worth noting that our analysis was pre-specified in an ex ante analysis plan, prior to performing any of the analyses. The analysis plan was time-stamped by uploading it to the Git repository of the Translational Neuromodeling Unit (TNU); it is available at https://gitlab.ethz.ch/tnu/analysis-plans/galioullineetal_ukbb_pred_depr. Furthermore, code reviews were performed by three of the co-authors (SF, SH and JH) who were not involved in the data analysis, both before the beginning of the analysis of the training data, and once again before running models on the test data. The code can be found at https://gitlab.ethz.ch/tnu/code/galioullineetal_ukbb_pred_depr.

Dataset: groups with/without depressive episodes

The process of data extraction from the UK Biobank is summarised by Figure 1. To avoid confusion, it is worth explaining that participants of the neuroimaging branch of UK Biobank (which started in 2014) underwent two fMRI scans, approx. three years apart, each of which involved both task fMRI and rs-fMRI data. In this study, we only used the rs-fMRI data acquired during the first scan.

Overall, selected individuals were required to have rs-fMRI data of good quality (as indicated by UK Biobank quality control) and no indication of any previous or current depressive episodes at the time

of their first fMRI scan. From the subset of participants that fulfilled these criteria, we aimed to select two groups, one of which continued to indicate no signs of depressive episodes (D- group) three years after their first scan, and one that showed at least one indicator for at least one depressive episode over this three-year period (D+ group).

Concretely, we first identified participants who had both task (UKB field 20249-2.0) and "resting-state" (UKB field 20227-2.0) fMRI scans in NIFTI format, ensuring quality controlled images already preprocessed by UK Biobank (Alfaro-Almagro et al., 2018), resulting in 35,485 participants. In order to define the D- group, we chose the subset of participants who responded “no” to the questionnaire item “Looking back over your life, have you ever had a time when you were feeling depressed or down for at least a whole week?” (UKB field 4598-2.0) when they were first scanned (2014+). This resulted in 15,739 participants. We further shrunk this set by selecting those individuals who continued to show no evidence of depression in following years (2014 to 2019) and again replied, at the second fMRI session in 2019+, “no" to the previous question (UKB field 4598-3.0). This resulted in 1,085 potential D- participants who could be searched for matching criteria once the D+ group had been determined.

Concerning the D+ group, we also selected individuals from the set of 15 739 participants who had preprocessed imaging data and who – during their first imaging questionnaire (2014+) – indicated never having been depressed. Since the UK Biobank does not include information about the absence or presence of a clinical diagnosis of depression for all participants, we used multiple sources of information to identify indicators of depression. Specifically, we searched selected UK Biobank data fields which plausibly indicated the occurrence of at least one depressive episode in the years after the first fMRI scan. The following list summarizes the data fields in UKB and number of hits.

- Medical records in UKB:
    - First Clinically Recorded Depressive Episode [UKB 130894] (5 hits)
    - Clinical Depression-Related Encounter [UKB 41270] (31 hits)
    - Prescription of Antidepressants [UKB 20003] (6 hits)
    - Depression Diagnosis Report in UKB Assessment [UKB 20002] (12 hits)
- Self-report data in UKB:
    - Depressed for at Least a Week Report [UKB 4598-2.0] (203 hits)
    - Depression Diagnosis Report in Mental Health Questionnaire [UKB 20544] (90 hits)
    - High Score (Coleman et al., 2020) on CIDI in Mental Health Questionnaire [UKB 20446] (165 hits)
    - High Score (sum > 4) on Patient Health Questionnaire 3-subset [UKB 2050, 2060, 2080] (6 hits)

Overall, this resulted in 518 potential D+ participants. Since for any given participant a previous depressive episode could be reflected by multiple hits, we took the union of the above 8 sets of hits. This resulted in a total of 464 participants in the D+ group.

Having completed the initial definition of D+ and D- groups, we searched for data entries showing inconsistent or logically incompatible responses from participants (e.g. participants stating "never depressed for at least a week" but with a clinical report of depression). This process led to the removal of 9 participants in total, resulting in 455 participants in the D+ group and 1,076 participants in the D- group.

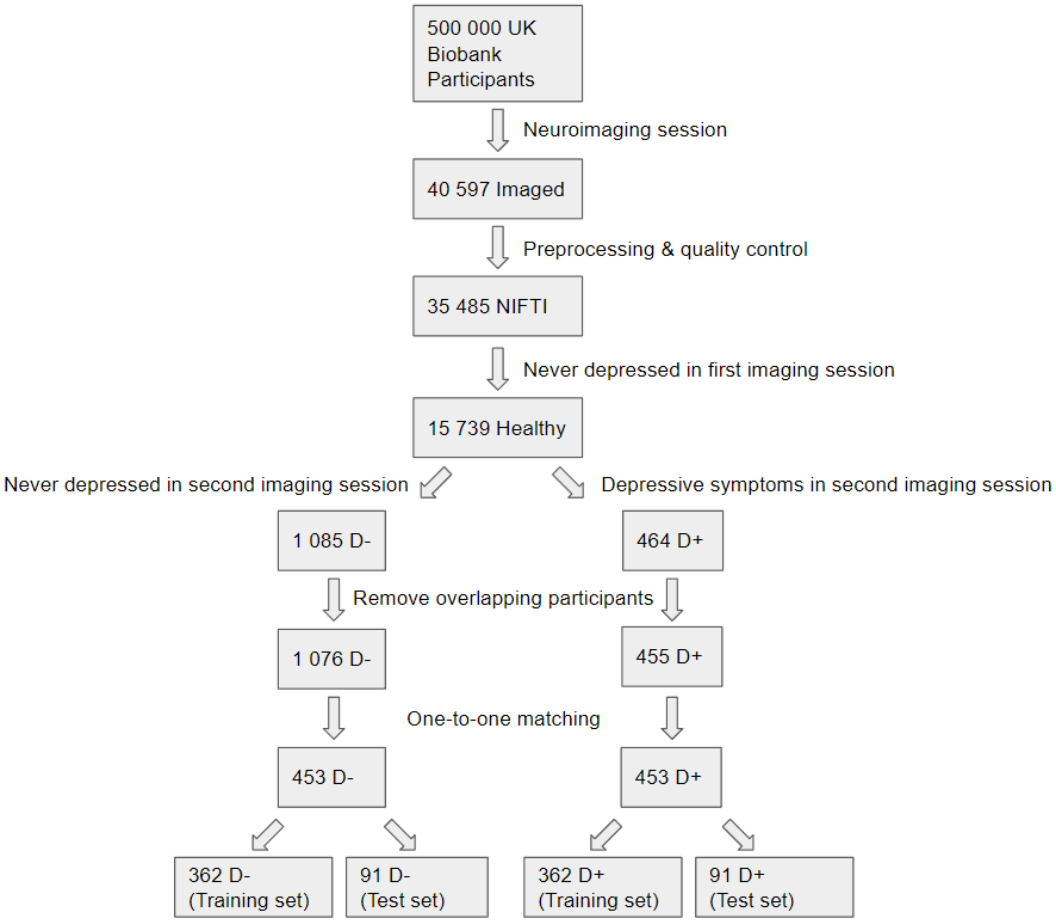

Figure 1: Flow diagram representing the dataset selection process. D- represents subjects who indicated no signs of depression, whereas D+ represents subjects who showed at least one indicator of depression.

Matching of participants and definition of training/test sets

To minimize any effects of potentially confounding variables, we matched participants with respect to multiple criteria. Specifically, for each D+ participant we tried to find a matching D- participant according to the following seven criteria (where a tolerance range was only allowed for age, as indicated):

- Sex (UKB field 31)
- Age ± 5 years (UKB field 34)
- Handedness (UKB field 1707)
- Tobacco smoking frequency (UKB field 1249)
- Alcohol consumption frequency (UKB field 1558)
- Ongoing addiction or dependence on illicit or recreational drugs (UKB field 20457)
- Historical cannabis consumption (UKB field 20453)

All but 57 D+ participants could be matched exactly. Out of these, 55 could be matched almost exactly, with at most one criterion deviating. Two D+ participants could not be matched and were excluded from further analyses. This provided us with a dataset of 906 participants in total: 453 D+ participants and 453 matched D- participants.

Finally, we performed an 80/20 split to partition the data into training and test sets. Both datasets were strictly separated from each other during data analysis to prevent any leakage of information that could affect the prediction results. We also addressed an unlikely, but theoretically possible, information leakage stemming from UK Biobank itself: the templates of major functional networks in the brain (Miller et al., 2016) which are offered by UK Biobank and which our study used for data extraction had been created using rs-fMRI data from the first 4,181 individuals in UKB. We resolved this potential problem by ensuring that all participants from this set that were also part of our extracted data were assigned to the training set. This resulted in patient/control training sets with 362 individuals each and test sets with 91 individuals (Figure 1).

FMRI Data Analysis

Wherever possible we used data that is directly available on UK Biobank and did not require additional processing. The rs-fMRI data are of 6 minute duration (490 images, TR=0.735s), with a spatial resolution of 2.4mm isotropic, and were acquired with 8x multislice acceleration (Alfaro-Almagro et al., 2018). We used the data after the standard preprocessing pipeline executed by UK Biobank. The processing steps performed at UK Biobank included realignment, EPI distortion correction, and high-pass temporal filtering (with a 50s cut-off). The rs-fMRI data were further processed using single-subject spatial ICA decomposition using MELODIC in FSL (Jenkinson et al., 2012). The resulting independent components (ICs) were classified as signal vs. noise, and a cleaned version of the data was provided. UK Biobank then fed these data into a dual regression (Nickerson et al., 2017) based on a set of group-level templates of "resting-state" networks (based on data from 4'181 subjects) at dimensionalities of either 25 or 100. For subsequent analyses, 21/25 and 55/100

ICs were kept, as the others were found in previous work to be "...clearly identifiable as artefactual (i.e., not neuronally driven)" (Alfaro-Almagro et al., 2018).

ICs of resting-state data can be thought of as distinct functional networks (Smith et al., 2009), and interactions between these networks can be investigated by applying functional and effective connectivity methods to IC timeseries (for previous examples, see Goulden et al., 2014; Hyett et al., 2015; Motlaghian et al., 2022). In this work, we selected three sets of networks, which differed in the number of ICs included. Since we were interested in major functional networks implicated in depression (Brakowski et al., 2017; Kaiser et al., 2015), our first IC selection targeted the default mode network (DMN), central executive network (CEN), salience network (SN), and the dorsal attention network (DAN). The DMN and DAN are mapped (Miller et al., 2016) to IC indices 1 and 3, respectively, while the left/right SN and left/right CEN are mapped (Gratton et al., 2018; Shen et al., 2018) to IC indices 6, 5 and 13, 21, respectively. Furthermore, we considered IC sets of size 21 and 55 components (as provided by UK Biobank) in order to explore the impact of increasing the number of networks/ICs on prediction performance. The 55 components can be interrogated interactively via a web-based visualisation tool provided by UK Biobank:
https://www.fmrib.ox.ac.uk/ukbiobank/group_means/rfMRI_ICA_d100_good_nodes.html

Generative Embedding

Having completed the selection of timeseries, our analysis proceeded to generative embedding (GE). GE requires two choices: (i) a generative model, and (ii) a ML method that uses posterior estimates from the generative model as features.

Concerning the choice of generative models, our analysis considered three different variants of DCM that are suitable for task-free fMRI data: stochastic DCM (Li et al., 2011), spectral DCM (Friston et al., 2014), and regression DCM (Frässle, Harrison, et al., 2021). For all models and all IC sets, we assumed a fully connected network. As a reference, we also obtained functional connectivity estimates, based on Pearson correlation coefficients.

To invert stochastic DCMs, we used the *spm_dcm_estimate* function in SPM12, with a DCM struct as input which had its *Y.y* set to the 6 timeseries, $\boldsymbol{a}$ set to a 6x6 matrix of ones (fully connected network of endogenous connections), and *Y.dt* set to 0.735 (interscan interval). This resulted in a 6x6 matrix of effective connectivity estimates, giving us 36 features for subsequent ML. Due to its high computational complexity, it was not possible to run stochastic DCM with 21 and 55 IC timeseries.

For spectral DCM, we used the SPM12 function *spm_dcm_fmri_csd* with the same exact DCM struct as for stochastic DCM as input. The performance was notably faster than for stochastic DCM, but given that it still took a few hours to run on the Euler high-performance computing cluster of ETH Zurich and that the scaling of the computational complexity is supra-linear in the number of ICs (i.e., number of nodes in the DCM), we estimated that it would still take weeks or even months to run the entire analysis (i.e., inversion of the DCMs for all subjects) for 21 or 55 IC timeseries. Hence, just like in the stochastic DCM case, we restricted the spectral DCM analysis to 6 IC timeseries.

Concerning rDCM, its high computational efficiency enabled us to analyse networks consisting of more components (6, 21, and 55 ICs), resulting in 36, 441, and 3025 features, respectively. We used the rDCM code in TAPAS 4.0 (Frässle, Aponte, et al., 2021), with *Y.y* set to the respective time series, and *Y.dt* set to 0.735.

Finally, FC matrices were computed using the *corrcoef* function in MATLAB. Since these matrices are symmetric along the diagonal, and the diagonal is always 1, we took the upper triangle of these matrices to be our features, resulting in 15, 210, and 1485 features for the respective IC sets. It is important to note that FC does not capture any information about the directionality of connections, as opposed to the effective connectivity measures from the DCM variants described above.

Classification

From the previous generative modeling, we had eight feature sets in place – functional connectivity for each IC set (6, 21, 55), stochastic and spectral DCM for 6 ICs each, and three rDCM feature sets for 6, 21 and 55 ICs. These feature sets were subsequently used as input to discriminative classifiers. Initially, we restricted all analyses to the training set data, and only touched the test data once we had selected a feature set / classifier combination that performed best. Regardless of the specific classifier chosen, the steps taken to arrive at reported metrics are the same.

Classifier training was performed using nested cross-validation (CV). Nested CV provides robustness against overfitting by optimizing hyperparameters in an inner CV loop while averaging the performance against other partitions of the data in an outer CV loop (Cawley & Talbot, 2010; Stone, 1974). In our case, we used 10 folds in the outer loop, and 5 folds in the inner loop. At the beginning of each iteration of the outer loop (before training with hyperparameter optimization), the confounds (sex, age, handedness, smoking, alcohol, illicit drugs, cannabis) were linearly regressed out using scikit-learn's *LinearRegression* module. Then the data were normalized using the *StandardScaler* module and then the classifier was finally fit with the *GridSearchCV* module. This

procedure yielded a set of performance measures for each feature/classifier combination (see Results).

After evaluation of the feature/classifier pairs on the training data, there are several possibilities how models fitted on the training data could be applied to the test data. First, we evaluated whether the feature set/classifier combination that had performed best on the training set generalised to the test set. Second, we performed a post hoc analysis in which we examined each feature set together with the classifier that had been optimal for this specific feature set on the training data.

In addition to evaluating classifiers based on their performance metrics, we also ran permutation tests to check for statistical significance of the classification results. These tests were run both on our training and test set. To generate an empirical null distribution for a given feature/classifier pair, we randomly permuted the labels while considering subject pairs between the D- and D+ groups, originating from the matching of confounds. This is done by identifying a pair and flipping their labels with 0.5 probability. For the resulting permuted labels, the classifier is trained again by re-running the entire nested CV procedure, yielding performance metrics under random conditions. This process is repeated many times (n = 1,000 in our case) to construct the empirical null distribution of performance metrics. We then compute the rank of the true performance metrics (obtained from the prediction without shuffling the labels) by calculating how many instances of the null distribution performed better. Dividing the rank by the number of permutations yields the p-value which we report.

A separate question concerned the choice of hyperparameters for the test set. While there are multiple options how hyperparameters for prediction on the test set could be chosen, we decided to use all data from the training set for optimising hyperparameters: we ran a non-nested 5-fold CV on the entire training set, picked the best-performing hyperparameters, and used those to predict on the test set. Other aspects relevant for classification on the test set, such as permutation testing, and regression of confounds were identical to the training set. Please see Figure 2 for a summary of the Materials and Methods described above.

Finally, we ran an interpretability analysis on our best-performing feature set/classifier combination (rDCM estimates based on 55 ICs and an SVM with a sigmoid kernel). This analysis based on SHAP (SHapley Additive exPlanations) (Lundberg & Lee, 2017), a generalisation of Shapley values from game theory (Shapley, 1953). For each feature, SHAP assigns an importance or attribution value that describes how much that feature contributes to the overall prediction. We used the *shap* software (https://github.com/slundberg/shap) to create a *KernelExplainer* that took as arguments:

- a sigmoid SVM classifier trained on rDCM with 55 ICs,

- a low-dimensional representation of the training data using *shap.kmeans* with five clusters (for computational tractability; see *shap* documentation)

Then, we computed the SHAP values using the explainer's *shap_values* function which takes the test data as an argument. This gives us a SHAP value for each feature for each subject, which we process (mean of SHAP value magnitude across subjects) to obtain the average impact of each feature on model output magnitude.

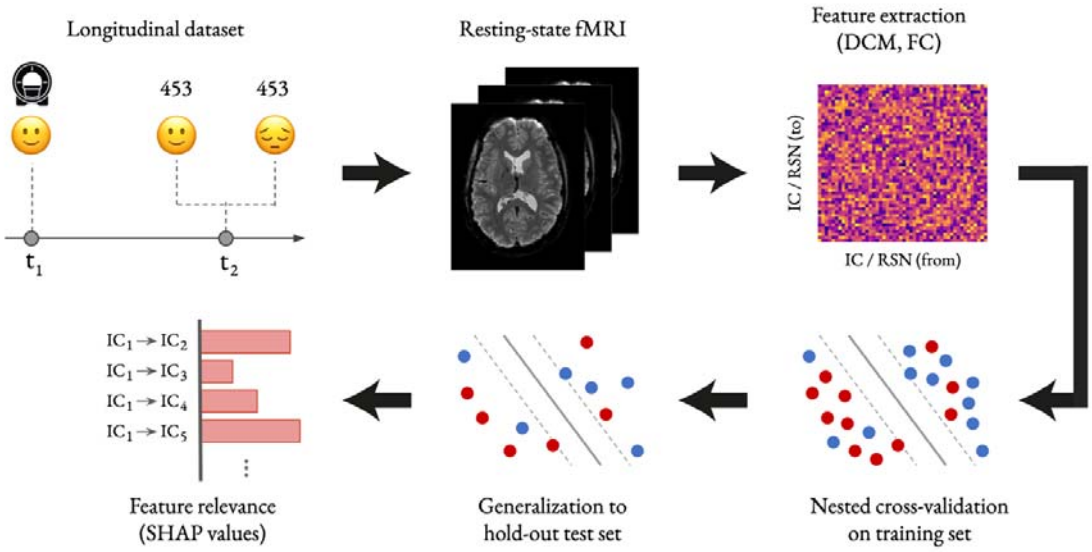

*Figure 2: Illustration of the generative embedding pipeline utilized in the present study for predicting indicators of future depressive episodes. The pipeline comprises: Definition of the longitudinal dataset (*top, left*), identification of data features from the resting-state fMRI (rs-fMRI) data (*top, middle*), feature extraction, representing effective connectivity (dynamic causal modeling, DCM) or functional connectivity (FC) estimates amongst independent components (IC) or resting state networks (RSN) derived from the rs-fMRI data (*top, right*), nested cross-validation on the training set (*bottom, right*), generalization to the test set (*bottom, middle*), and feature relevance analysis based on SHAP values (*bottom, left*). Parts of the figure contain material from shutterstock.com (with permission).*

Choice and Implementation of Classifiers

A total of 17 classifiers were evaluated (please see Table 1 in the Results section), including six support vector machine (SVM) variants and three neural network (NN) variants. As described in the following, for most classifiers, we chose hyperparameters to optimize within the inner nested CV loop. For two classifiers (Gaussian naive Bayes and quadratic discriminant analysis) where hyperparameter tuning is less common, we kept scikit-learn's default parameters.

A first classifier was *logistic regression*. Following the default parameters of the scikit-learn version, we also used *L2* regularization, used *lbfgs* as our solver, and iterated at maximum 100 times. In the inner CV loop, we optimized for the regularization parameter C (0.01, 0.1, 1, 10, 100), which is the inverse of regularization strength (smaller values enforce stronger regularization).

For the *SVMs*, we made use of Platt scaling (Platt, 1999) to get probabilistic outputs for use in an AUROC (area under the receiver-operating characteristic curve) metric. We attempted classification with all types of kernels that scikit-learn has to offer, which include linear, radial basis function (RBF), sigmoid, and polynomial kernels with order 3, 4, and 5. We again treated the regularization parameter *C* (0.01, 0.1, 1) as a hyperparameter and additionally tuned gamma (1, 0.1, 0.01, 0.001) — the kernel coefficient for the RBF, sigmoid, and polynomial kernels.

We included three *neural network* variants (1, 2, and 3 hidden layers) which treat their layer sizes as hyperparameters. All other parameters are scikit-learn defaults (version 0.23.2), which means that – unlike logistic regression – the activation function used is actually ReLU (Rectified Linear Unit; Nair & Hinton, 2010).

Neural Network Hyperparameters

- 1 hidden layer sizes: 100, 150, 300, 500
- 2 hidden layers sizes: (100, 50), (150, 20), (300, 100), (500, 250)
- 3 hidden layers sizes: (100, 50, 5), (150, 20, 10), (500, 250, 50)

Ensemble methods combine multiple base models to (hopefully) produce better results than each individual model would have on its own. One such algorithm we used is *AdaBoost* (Freund & Schapire, 1997). We employed the scikit-learn default base classifier (decision tree) treating the number of estimators (30, 50, 70) as a hyperparameter. For a baseline comparison, we also attempted classification with a *single decision tree* with default scikit-learn parameters. Another ensemble method used in our classification is *gradient boosting* (Friedman, 2001) with 50, 100, 150 estimators as hyperparameters. Finally, we also tried *random forest* (Breiman, 2001), with options to tune 100, 500, 1000 estimators. For random forest, we additionally treated the maximum tree depth (10, 30, 60) as a hyperparameter.

We also explored prediction with three supervised learning algorithms that do not fall under the previous categories. Namely, *Gaussian naive Bayes* (Zhang, 2004), *quadratic discriminant analysis* (Cover, 1965) – both of which use scikit-learn default parameters – and *k-nearest neighbors* (Cover & Hart, 1967) where we treated the number of neighbors (3, 5, 7, 9) and leaf size (20, 30, 40) as hyperparameters.

We used a variety of metrics to evaluate classifier performance in order to ensure a holistic view and to avoid potential pitfalls (such as overemphasizing the importance of one metric). We report recall (sensitivity), precision (positive predictive value), F_1 score, accuracy, and AUROC (area under the receiver-operating characteristic curve).

Deviations from the original analysis plan

Our analyses were pre-specified and are described in a time-stamped analysis plan (https://gitlab.ethz.ch/tnu/analysis-plans/galioullineetal_ukbb_pred_depr). We subsequently extended this analysis plan in three ways:

1. We extended the coverage of networks and, in addition to the 6 networks (represented by IC timeseries), also considered sets of networks consisting of 21 and 55 ICs, as provided by UK Biobank.
2. We extended the connectivity methods by considering functional connectivity (Pearson correlation coefficients) in addition to variants of DCM as generative models.
3. In addition to SVMs, we decided to test a larger set of classifiers in order to avoid that our results may depend on the particular choice of classifier.
4. We included an analysis of feature importance on the test set using SHAP values.

The decision to extend the analyses in this manner took place before any prediction analyses of the training or test data were conducted.

Results

We first present the performance of the cross-validated classifiers for the training dataset and then proceed with the most promising feature/classifier combinations to the test dataset. Note: since our dataset is balanced, accuracy as a metric implies balanced accuracy.

Training set

Table 1 provides an overview of prediction performance on the training set in the nested cross-validation setting. Altogether, 17 different classifiers were evaluated (including four SVM variants and three neural network variants). We report AUROC of all the features run with each classifier (Table 1) and we also report all metrics for the five best feature/classifier combinations (Table 2).

Classifiers \ Features	FC (6)	FC (21)	FC (55)	St. DCM	Sp. DCM	rDCM (6)	rDCM (21)	rDCM (55)
Ada	0.52	0.49	0.56	0.49	0.47	0.55	0.54	0.56
DTC	0.49	0.49	0.52	0.50	0.51	0.51	0.53	0.52
GBC	0.51	0.50	**0.57**	0.49	0.45	0.60*	0.60*	0.64*
GNB	0.53	**0.54**	0.54	0.50	0.49	**0.61***	0.61*	0.63*
kNN	0.52	**0.54**	0.51	0.49	0.47	0.53	0.58	0.59
LR	**0.54**	0.52	0.55	0.51	0.51	0.59	0.58	0.58
NN (1)	0.47	0.52	0.55	0.47	**0.55**	0.50	0.60*	0.61*
NN (2)	0.47	0.52	0.54	0.48	0.53	0.50	0.58	0.63*
NN (3)	0.48	0.52	0.55	0.50	0.54	0.53	0.59	0.62*
QDA	0.47	0.52	0.50	**0.52**	0.53	0.52	0.48	0.51
RF	0.51	0.52	0.56	0.49	0.47	0.57	0.63*	**0.66***
SVM (lin)	**0.54**	0.50	0.55	0.51	0.48	0.58	0.58	0.56
SVM (3)	0.47	0.49	0.52	**0.52**	**0.55**	0.59	0.61*	0.65*
SVM (4)	0.49	0.50	0.47	0.50	0.47	0.49	0.46	0.47
SVM (5)	0.49	0.47	0.47	0.49	**0.55**	0.56	0.47	0.46
SVM (rbf)	0.51	0.51	0.55	0.50	0.50	0.60*	0.63*	0.65*
SVM (sig)	0.52	0.52	0.47	0.48	0.48	**0.61***	**0.64***	**0.66***

Table 1: Summary table of AUROC for each feature/classifier combination as determined by nested cross-validation on the training set. Bold represents the best result across classifiers for a given feature set, orange shading represents the best result across feature sets for a given classifier, and a star denotes a statistically significant result ($p \leq 0.05$). Ada: AdaBoost, DTC: Decision Tree Classifier, GBC: Gradient Boosting Classifier, GNB: Gaussian Naïve Bayes, kNN: k-Nearest Neighbors, SVM (lin): Support Vector Machine with linear kernel, LR: Logistic Regression, NN (n): Neural Network with n layers, SVM (n): Support Vector Machine with polynomial kernel order n, QDA: Quadratic Discriminant Analysis, SVM (rbf): Support Vector Machine with radial basis function kernel, RF: Random Forest, SVM (sig): Support Vector Machine with sigmoid kernel.

Features + Model	Precision	Recall	F_1 Score	Accuracy	AUC
rDCM(55) + SVM (sig)	**0.64**	**0.60**	**0.62**	**0.63**	**0.66**
rDCM(55) + RF	0.63	0.58	0.60	0.62	**0.66**
rDCM(55) + SVM (3)	0.63	0.59	0.61	0.62	0.65
rDCM(55) + SVM (rbf)	**0.64**	0.58	0.61	0.62	0.65
rDCM(22) + SVM (sig)	0.62	0.57	0.59	0.61	0.64

Table 2: Summary of all metrics on the top five feature set/classifier combination as determined by nested cross-validation on the training set. Bold indicates the best model for the given metric.

Having run all feature set/classifier pairs on the training data using nested cross-validation, we found that a sigmoid SVM paired with an rDCM taking 55 ICs – referred to subsequently as rDCM(55) – as input performed best (Tables 1, 2). In terms of performance, applying a sigmoid SVM to rDCM connectivity estimates based on 55 ICs resulted in an AUROC of 0.66 (Figure 3A). The other performance metrics for this combination were: precision=0.64, recall=0.60, F1 score=0.62, accuracy=0.63 (Table 2).

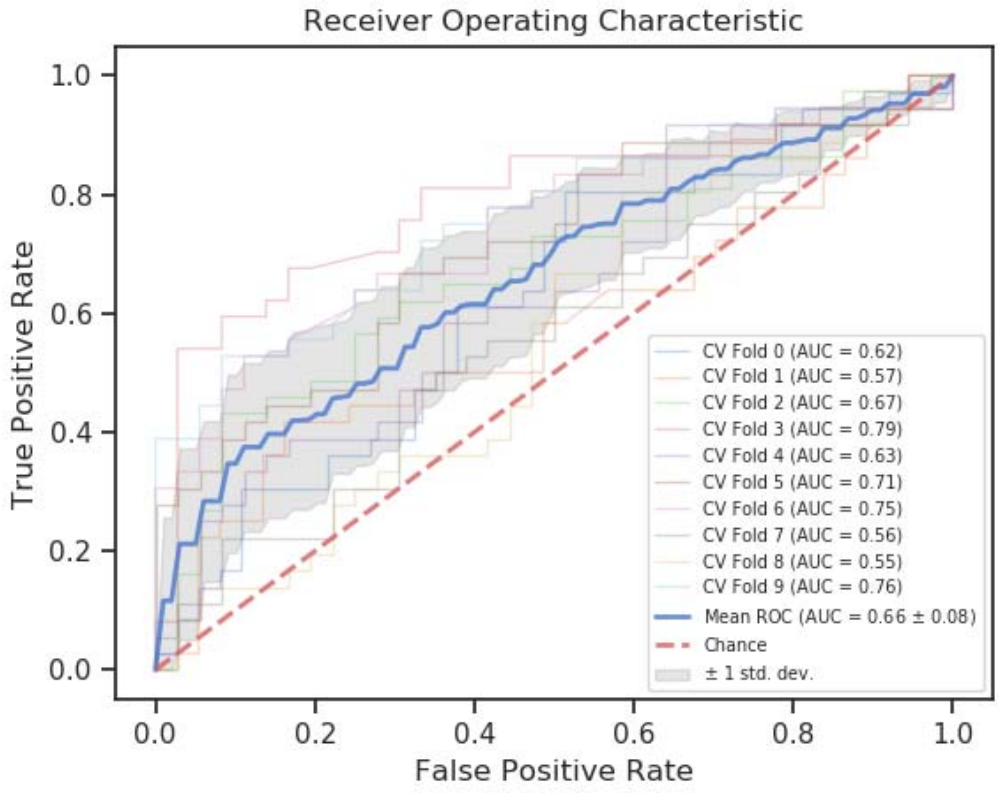

Figure 3: ROC curve for sigmoid SVM paired with rDCM(55) on the training set run with nested cross validation.

In general, connectivity estimates by rDCM enabled better predictions, regardless of classifier (see Table 1, orange shading): for 15 out of the 17 classifiers tested, one of the rDCM feature sets

resulted in the best AUROC (in 11/15 cases, the best feature set was rDCM(55)). Furthermore, SVMs tended to perform better than other classifiers, with SVM variants having highest accuracy for 6 out of 8 connectivity feature sets. In particular, the sigmoid SVM was the best-performing classifier for 3 feature sets, more than any other classifier.

Based on these results, we chose rDCM(55) with sigmoid SVM to move forward to the test set. The test data had not been touched up until this point to prevent any leakage of information and ensure a thorough verification of the generalizability of our prediction model.

Test set

The prediction of the best-performing approach on the training set generalized to the test data: the application of a sigmoid SVM to connectivity estimates by rDCM (55 ICs) from the test set showed an AUROC of 0.64 (Figure 4A). This prediction performance was significantly above chance: Figure 4B shows that the achieved accuracy of 62% is well outside the null distribution generated by predictions on randomly permuted labels ($p<0.001$).

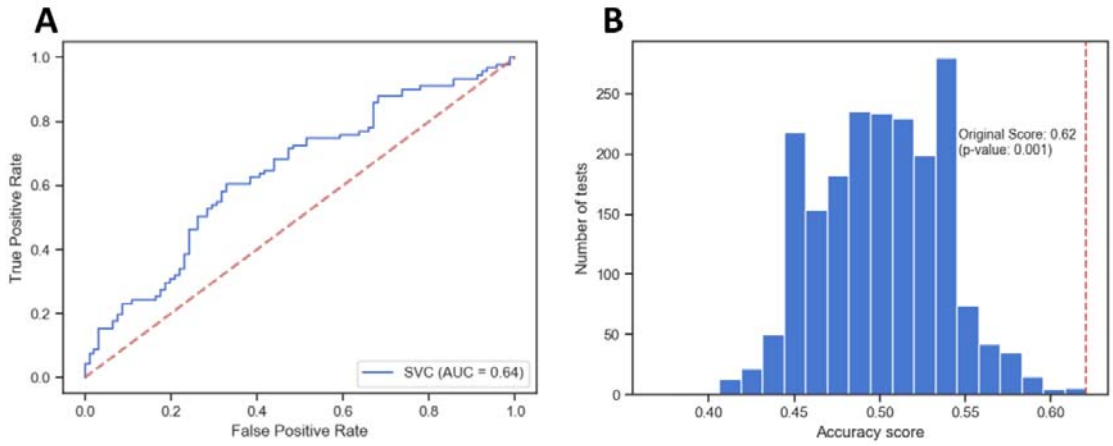

Figure 4: (A) ROC curve of rDCM (55 ICs) with sigmoid SVM run on test data. (B) Permutation test (n=1,000) run on test data with accuracy as the metric.

To get a better understanding of the generalization performance we conducted a post-hoc analysis on other well-performing feature/classifier pairs (Table 3) from the nested cross-validation and assessed their performance on the test data. We defined "well-performing" as the best classifier in general, but for feature sets where another classifier performed better, we selected the latter instead. From the 13 classifiers tested post-hoc, only three other pairs had above-chance performance on the test set, namely rDCM (21 ICs) with sigmoid SVM (58% accuracy, p=0.019), functional connectivity (6 ICs) with sigmoid SVM (59% accuracy, p=0.007), and functional

connectivity (21 ICs) with Gaussian Naïve Bayes (59% accuracy, p-value=0.012). All of these performed worse with at least a 3% drop in accuracy, leaving the rDCM (55 ICs) with sigmoid SVM as the best-performing model overall on the test data.

Features + Model	Precision	Recall	F_1 Score	Accuracy	AUC
rDCM(55) + SVM (sig)	**0.63**	0.56	**0.60**	**0.62**	**0.64**
FC(21) + GNB	0.59	**0.60**	0.59	0.59	0.58
FC(6) + SVM (sig)	0.59	0.59	0.59	0.59	0.61
rDCM(21) + SVM (sig)	0.60	0.51	0.55	0.58	0.59
FC(6) + Log Res	0.57	0.52	0.54	0.57	0.61

Table 3: Summary of all metrics on the top five feature/classifier combinations on the test set.

Finally, we computed the SHAP values (Figure 5) for our best-performing model, the sigmoid SVM classifier paired with rDCM(55). This assesses the contribution of each connection to the prediction performance. Since rDCM provides estimates of effective (directed) connectivity, we have two SHAP value estimates for each IC, one for the outgoing connection, and one for the incoming connection. We visualized the top 100 SHAP values as a circular plot, where each IC is shown twice, and the bottom half represents the associated values for the outgoing connections. The width of each displayed connection reflects the magnitude of the SHAP value, and the width of the coloured IC label on the circle represents the cumulative SHAP value for outgoing or incoming connections of that node. Figure 5 shows that connections with the top 100 SHAP values were not confined to a few networks but included almost all ICs, with very few exceptions. Put simply, during the "resting" state of unconstrained cognition the participants were in, the most predictive connections were found all over the brain.

Furthermore, we examined the entire distribution of SHAP values, which is shown as a histogram in Figure 6. This demonstrates that all connections contribute to the model's prediction, albeit most of them to a small degree. The distribution shows considerable spread and a long tail, where the contribution of the most important connection (from IC 34 to IC 24; compare Figure 5) is two orders of magnitude larger than connections at the mode of the histogram.

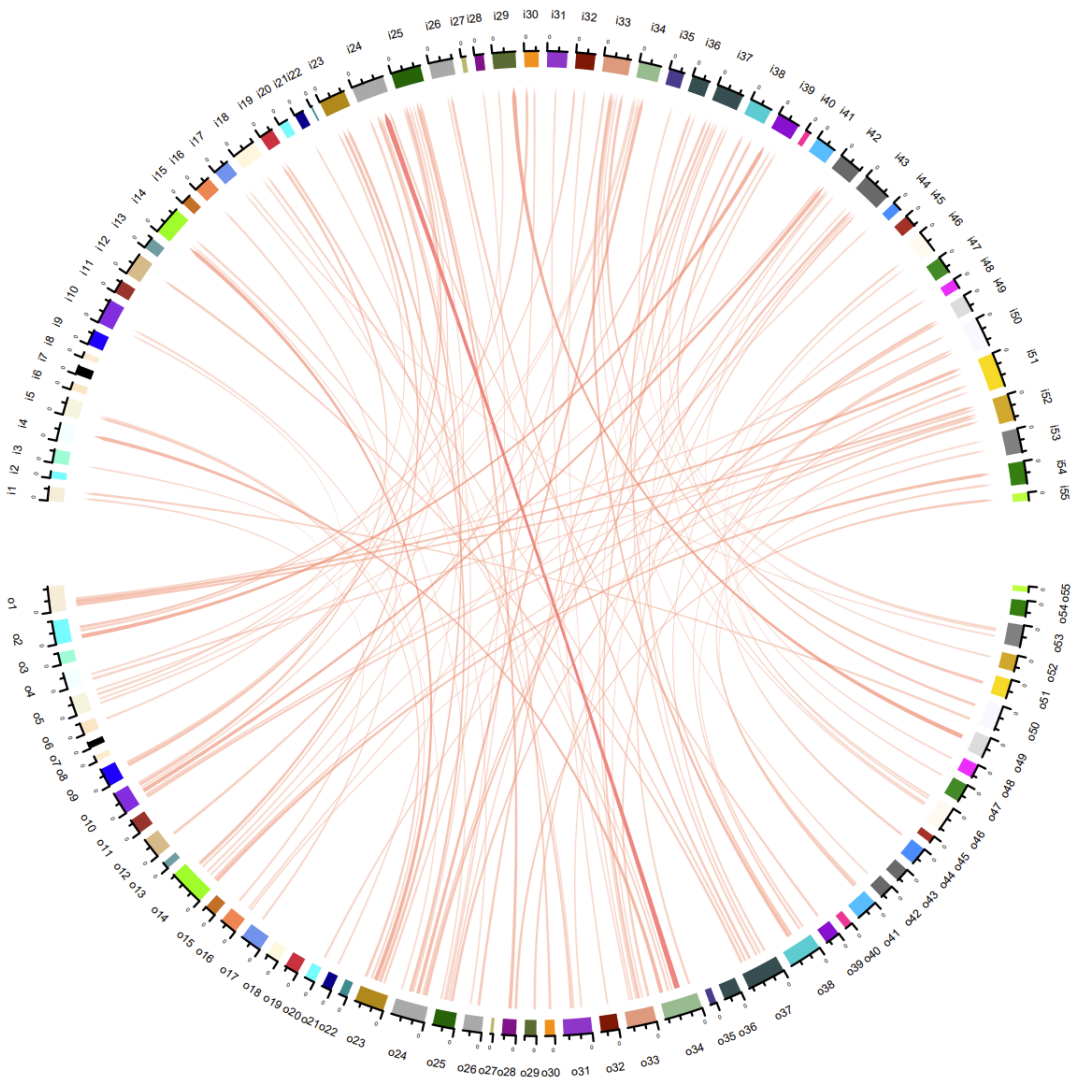

Figure 5: Top 100 Shapley values for sigmoid SVM paired with rDCM(55) on the test data. Bottom half are outgoing connections from each of the 55 ICs, and top half are incoming connections.

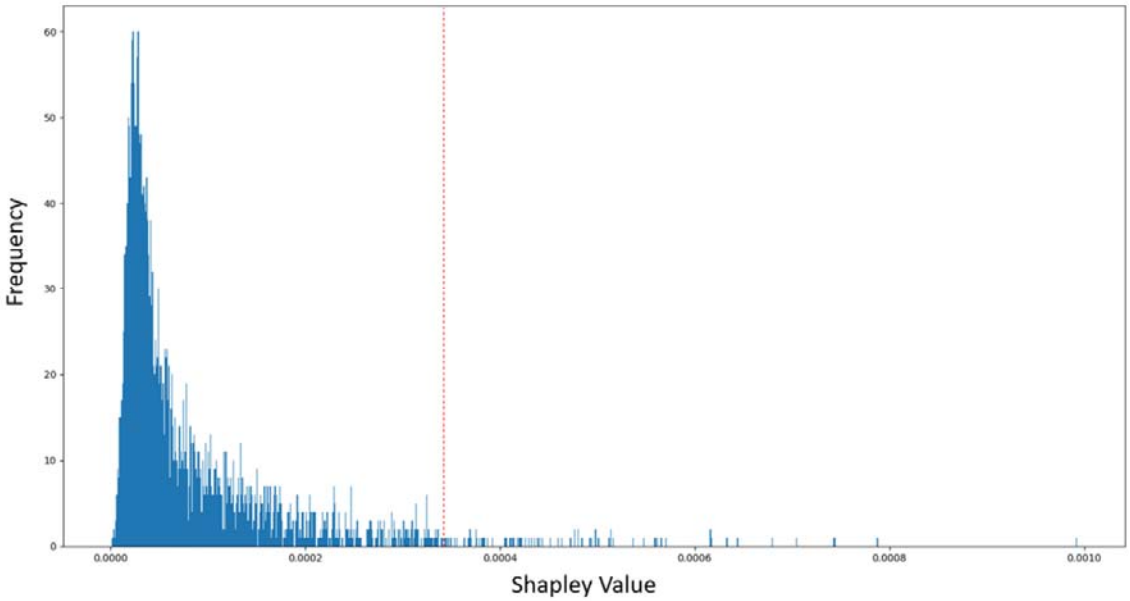

Figure 6: Shapley value distribution for sigmoid SVM paired with rDCM(55) on the test data. Values to the right of the dashed red line are in the top 100.

Discussion

MDD is a syndrome with heterogenous disease trajectories (Merikangas et al., 1994) and variable treatment responses (Rush et al., 2006). Given the importance for clinical management, predicting future clinical outcomes of individual MDD patients has become an important topic in computational psychiatry. In particular, various fMRI studies have examined the feasibility of predicting treatment response (e.g. Harris et al., 2022; Hopman et al., 2021; Ju et al., 2020; Osuch et al., 2018; Queirazza et al., 2019), relapse (e.g. Berwian et al., 2020; Lawrence et al., 2022), or disease trajectories (e.g. Frässle et al., 2020; Schmaal et al., 2015) in individuals with MDD.

By contrast, there have been hardly any attempts to use fMRI to address another challenge of similar importance: the early detection of individuals who are at risk of experiencing a future episode of depression. Given the high frequency of a prolonged remitting-relapsing disease course after a first episode of MDD (Eaton et al., 2008), identifying at-risk individuals is crucial for enabling the targeted deployment of preventive measures and early interventions. So far, to our knowledge, there has only been a single study that used fMRI for detecting individuals at-risk for future depression (Hirshfeld-Becker et al., 2019). This previous study used rs-fMRI and functional connectivity measures in a small sample of individuals with familial risk for MDD (N=33 for prediction).

The study presented in this paper is novel in several ways. It is the first study using generative models of fMRI data as a basis for predicting future depressive episodes, using three different

variants of DCM, in comparison to simpler functional connectivity measures. It uses a large balanced sample size (N=906), carefully matches groups with presence and absence of depressive symptoms, examines the combination of 8 connectivity feature sets with 17 classifiers in a training set, and evaluates the generalisability of the best predictions using a held-out test set.

The results from the training set (Table 1) indicated that the combination of the rDCM(55) feature set (i.e. rDCM-based connectivity estimates between 55 networks or ICs) and a SVM (with a sigmoid kernel) performed best, showing an AUROC of 0.66 and an accuracy of 63%. This result was significantly above chance, as indicated by permutation testing (p=0.001, Figure 3B). Moreover, across classifiers, rDCM demonstrated higher predictive value than other connectivity methods (see Table 1): for 15 out of the 17 classifiers tested, one of the rDCM feature sets resulted in the best AUROC; in 11/15 cases, the best feature set was rDCM(55). Examining the results along the other dimension of our investigation, i.e. across all connectivity feature sets, SVMs performed better than other classifiers: for 6 out of 8 connectivity feature sets, one of the SVM variants had the highest accuracy. In particular, a SVM with a sigmoid kernel performed best for 3 feature sets, surpassing any other classifier.

Evaluating the best combination (i.e. rDCM(55) + SVM with sigmoid kernel) on the test set confirmed the generalisability of the predictions, resulting in an AUROC of 0.64 and an accuracy of 62%. This was significantly above chance (p=0.001), as confirmed by permutation testing (Figure 4B). In a post-hoc analysis, we also evaluated the predictive value of all other connectivity feature sets on the test set; notably, for each feature set, we used the classifier that had performed best on the training set. These analyses showed that three other combinations of connectivity features/classifiers (rDCM(21) + sigmoid SVM, FC(6) + sigmoid SVM, and FC(21) + Gaussian Naïve Bayes) also achieved significant results, although with slightly lower accuracy (58-59%).

In short, our results thus demonstrate that a GE procedure – based on applying rDCM to rs-fMRI timeseries from a large number of ICs (55) – enabled the best predictions about the occurrence of future depressive episodes within a 3-year period. Having said this, the superiority of GE over a simpler prediction procedure based on FC estimates was not large, amounting to 3% higher accuracy and 0.03 higher AUROC compared to the combination of FC(6) + sigmoid SVM. A binomial test indicated that this difference in accuracy was not significant (p=0.315).

The lack of a decisive advantage of generative embedding in this rs-fMRI study contrasts with previous task-based fMRI studies in which GE based on DCM was clearly superior to predictions based on FC estimates (e.g. Brodersen et al., 2011, 2014; Frässle et al., 2018, 2020). For example, DCM estimates of effective connectivity during a face perception task allowed for substantially more

accurate predictions of MDD disease trajectories than FC estimates: balanced accuracies for predicting a chronic course vs. remission were 79% for DCM and 50% for FC, a difference that was highly significant (Frässle et al. 2020).

In order to understand the limited advantage of GE over FC-based prediction in this study, it is useful to first consider the general reasons why one would, in general, expect GE to show superior performance. In brief:

(i) GE exploits the fact that a generative model partitions data into signal and noise. Using model parameter estimates (as a low-dimensional representation of signal) as features for subsequent ML ensures that only meaningful information underpins training of classifiers. This makes it less likely that predictions are informed by noise and do not generalise. By contrast, measures of functional connectivity, such as correlation coefficients, reflect both signal and noise. As highlighted by Friston (2011), functional connectivity estimates based on correlations are highly susceptible to changes in the signal-to-noise ratio of data.

(ii) A generative model like DCM distinguishes different mechanisms how measured signal in a system of interest is caused, e.g. connections between system nodes or external inputs. This allows predictions to be differentially informed by distinct system mechanisms. By contrast, FC cannot distinguish whether co-varying signal in two brain regions is caused by shared input or by connections between the regions.

(iii) DCM provides directed connectivity estimates, allowing one to obtain separate weights for reciprocal connections between regions. By contrast, FC can only provide undirected estimates of connection strengths.

(iv) From a classical test theory perspective, test-retest reliability of connection strength estimates would be considered an important prerequisite for predictive validity. Concerning rs-FC, test-retest reliability has been examined in numerous studies; a recent meta-analysis reported that, on average, individual connection estimates have limited test-retest reliability (Noble et al., 2019). A direct comparison between FC and rDCM-based estimates of connectivity on identical data (rs-fMRI and multiple tasks) demonstrated that rDCM performed more favourably in this regard (see Figure 3 in Frässle & Stephan, 2022).

Considering these general factors, one possibility why we only found a limited advantage of GE over FC-based predictions in this study relates to (i) above: in the present study, connectivity was estimated from timeseries that resulted from ICA decomposition and subsequent (manual) removal of components that were identified as noise (Alfaro-Almagro et al., 2018). This approach may have

584 diminished the difference between GE and FC-based prediction with regard to denoising. For
585 comparison, in previous comparisons of GE and FC-based predictions (e.g. Brodersen et al., 2011,
586 2014; Frässle et al., 2018, 2020), timeseries were obtained by computing the first principal
587 component from regional BOLD measurements, which does not involve a specific distinction
588 between signal and noise. Another possible explanation derives from (ii): application of rDCM to rs-
589 fMRI data essentially means that the model "switches off" external inputs (Frässle, Harrison, et al.,
590 2021). This reduces the superiority in representational richness of GE.

591 In summary, this suggests that, in the current setting of IC-based rs-fMRI timeseries, only factors (iii)
592 and (iv) – but not factors (i) and (ii) – could potentially contribute to higher performance of GE. In
593 order to obtain an impression of the potential impact of factor (iii) – the ability of DCM to obtain
594 separate weights for reciprocal connections between network nodes – we visually explored the
595 asymmetries of node-level SHAP values for incoming versus outgoing connections. For each of the
596 55 network nodes (ICs), Fig. 7 plots SHAP values summed across all incoming (afferent) and outgoing
597 (efferent) connections, respectively. Visually, it is apparent that for many of the network nodes, the
598 explanatory contributions of incoming versus outgoing connections differ considerably (up to 59%).
599 A more fine-grained plot of connection-specific SHAP values is provided by Fig. 8.

600 These plots also illustrate a disadvantage of the analysis approach we have chosen in the current
601 study. Specifically, using ICs as network nodes diminish the advantage GE usually enjoys in terms of
602 rendering predictions neurophysiologically interpretable. For example, as shown by Figures 5 and 8,
603 the connection with the largest SHAP value is the connection from IC 34 to IC 24. Both of these
604 components include a set of fronto-parietal areas: IC 34 includes bilateral frontal regions that appear
605 to match the location of the frontal eye fields as well as more anterior parts of the superior parietal
606 cortex. By contrast, IC 24 contains more posterior bilateral parietal areas, including large parts of
607 bilateral intraparietal sulcus, as well as parts of right middle frontal gyrus and right middle/inferior
608 temporal gyrus. Given this complex anatomical configuration, the biological interpretation of a
609 (directed) functional coupling between IC 34 and IC 24 is not as straightforward as a functional
610 coupling between specific frontal and/or parietal areas. While the FC between these components
611 does not enable any easier interpretations, this example illustrates that the usual interpretive
612 advantage of GE tends to be lost when using IC components as nodes of networks.

613

614

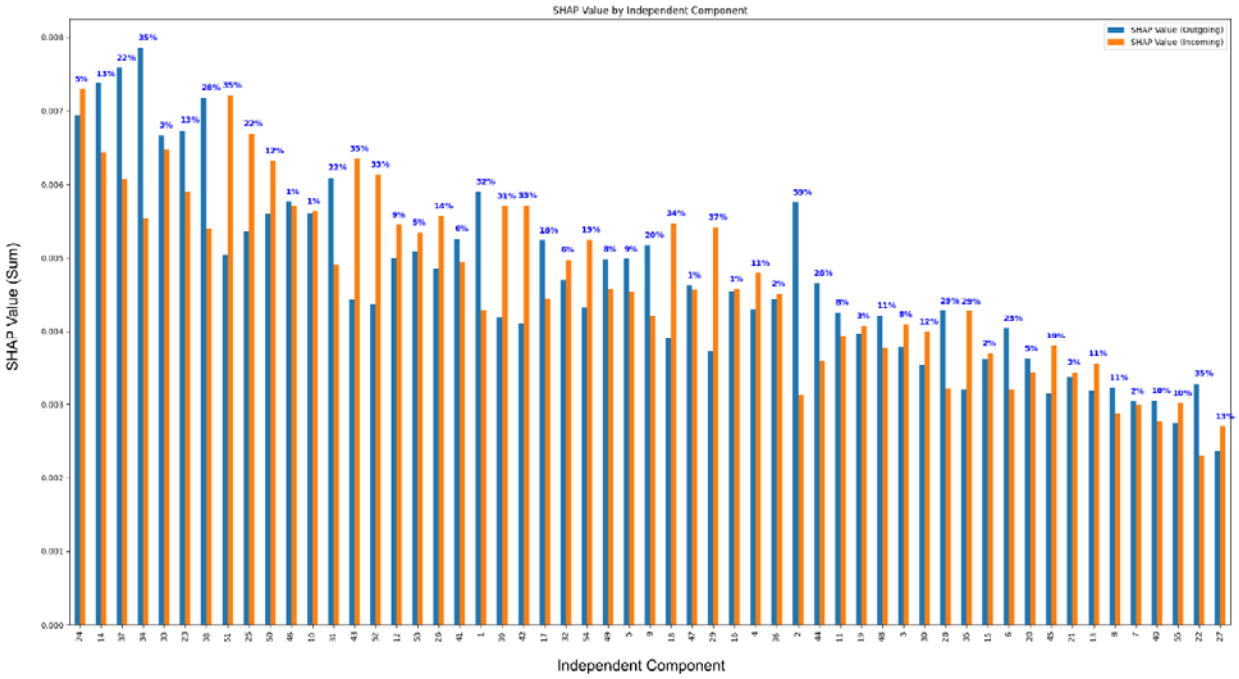

Figure 7: SHAP values summed across the incoming (afferent) and outgoing (efferent) connections of each IC. Percentages indicate the % difference in SHAP values for afferent and efferent connections. The plot concerns predictions based on rDCM(55) estimates and SVM with a sigmoid kernel.

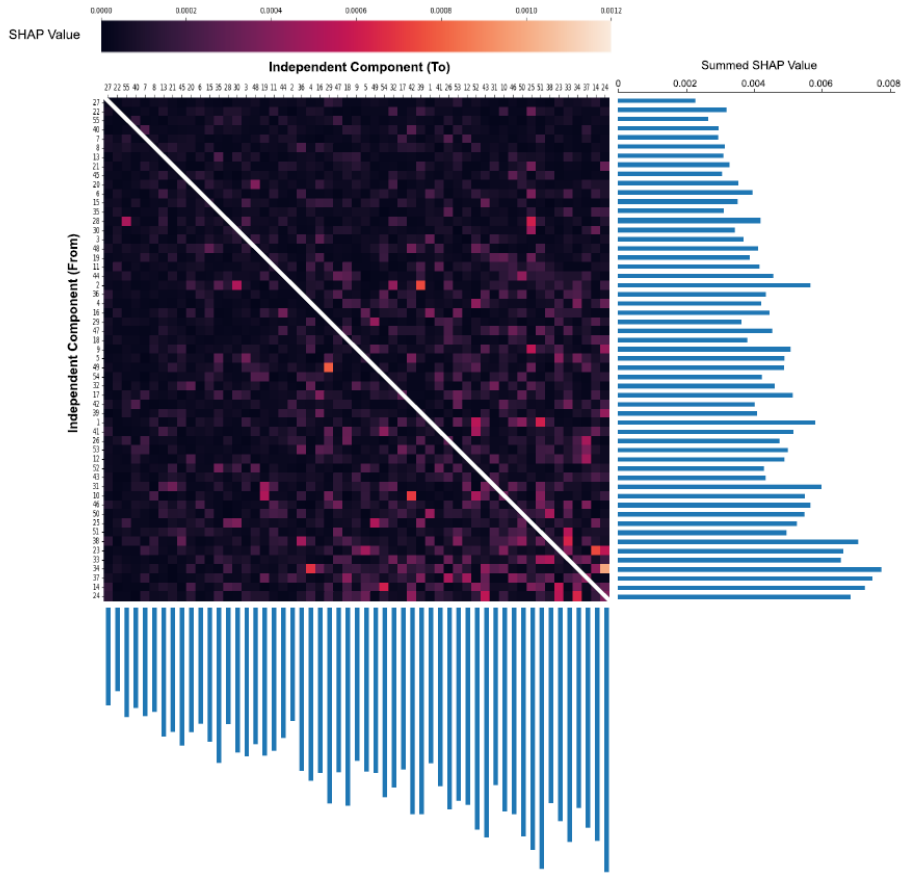

Figure 8: Matrix of connections between all ICs, showing the connections' colour-coded SHAP values (test data) for predictions based on rDCM(55) estimates and sigmoid SVM. ICs are ordered according to summed SHAP values.

Three further aspects of the results deserve discussion. First, it may initially seem surprising that SVM turned out to be the most successful classifier in our comparison, surpassing potentially more powerful methods like neural networks. However, this result is compatible with several recent reports that, for neuroimaging data, kernel-based methods like SVMs (and, in some cases, even simpler linear models) perform equivalently to neural networks for sample sizes up to 10,000 (Cole et al., 2017; He et al., 2020; Schulz et al., 2020).

Second, in our post-hoc analysis of connectivity features/classifier combinations on the test set, three of four significant predictions used the same classifier, an SVM with a sigmoid kernel. Strikingly, FC achieved a significant 59% predictive accuracy using only 6 ICs, whereas the more accurate prediction by rDCM (62%) used 55 ICs, respectively. The resulting difference in the number of features is substantial (15 for FC versus 3025 for rDCM), and it is not immediately clear why predictions based on functional vs. effective connectivity differed greatly in the preferred dimensionality of the feature set. One speculative explanation – which would be consistent with the findings in Figures 7 and 8 – is that differences in the strengths of reciprocal between-network connections provide subtle but meaningful information that is distributed over many connections (compare factor (iii) above). This type of information would only be reflected by rDCM, but not by FC-based, connectivity estimates. More generally, it is not clear why FC(6) performed so well on the test data at all. The nested CV on the training data did not indicate that this feature set might be particularly predictive (maximum accuracy of FC-based predictions with any classifier was 54%, none of them significant). The finding of a higher accuracy (59%) on the test set our post-hoc analysis was surprising. It might be a chance result due to the variance inherent in CV procedures (Varoquaux, 2018) but otherwise lacks a compelling explanation.

Third, contrary to our expectations, predictions based on stochastic and spectral DCM did not generalise to the test set. One possible reason for the lack of successful generalisation is that the higher complexity of the model formulation (e.g. the flexible hemodynamic component and the more sophisticated noise model) could make parameter estimation less reliable, e.g. due to greater abundance of local extrema in the objective function, which would be expected to harm generalisability. This possibility is supported by a recent investigation of parameter recovery of spectral DCM and rDCM which found more accurate parameter recovery for the latter (Frässle et al. 2021). Perhaps even more importantly, however, we could only run spectral and stochastic DCM for 6 ICs on our cluster; for larger feature sets, their compute time (within the context of our entire

analysis pipeline) became prohibitively long. However, considering the success of rDCM based on 55 ICs, it is plausible that spectral and stochastic DCM may have performed better if we had been able to run them with larger IC sets (21 and 55).

How does the prediction performance achieved in this study compare to previous results in the literature? The only previous fMRI study on predicting future depression (Hirshfeld-Becker et al., 2019) used FC estimates based on rs-fMRI data from six regions, achieving 92% accuracy. However, this study recruited never-depressed children with familial risk for MDD, as opposed to never-depressed participants from the general population as in our study. Additionally, given the more specific focus of the previous study, only 33 participants were available for classification (25 at-risk children, eight controls); this small sample size did not allow for verification in a held-out dataset. Another useful (although not fMRI-based) comparison study utilized structural MRI together with clinical data, questionnaires, and environmental variables (Toenders et al. 2021). The study used a large training set (N=407 adolescents) and an independent test set (N=137), achieving an AUROC between 0.68-0.72.

It is also instructive to consider the results from non-imaging studies that used demographic, socioeconomic, and clinical variables for predicting the future onset of depression. When considering those studies that had large sample sizes (i.e. N>500) and tested for generalisability in an independent test set, the reported AUROC values in the literature range between 0.71-0.87 (Caldirola et al., 2022; King et al., 2008; Librenza-Garcia et al., 2021; Na et al., 2020; Xu et al., 2019). It is noteworthy, however, that these studies mostly used imbalanced datasets where the number of negative cases (no future depressive episode) far outnumber the positive cases. For example, in the two studies with the highest prediction performance – i.e., AUROC of 0.87 (Na et al., 2020) and 0.85 (Caldirola et al., 2022) – individuals with future depressive episodes amounted to approx. only 8% and 7% of the respective samples. Even when techniques such as oversampling are used (as in Na et al., 2020; but not always the case in other studies), such imbalance can lead to overly optimistic estimates of prediction performance.

Our study has strengths and limitations. Its strengths include an ex ante analysis plan (https://gitlab.ethz.ch/tnu/analysis-plans/galioullineetal_ukbb_pred_depr) and a large (N>900) and balanced sample in which groups were carefully matched for 7 potentially confounding variables (age, sex, handedness, tobacco smoking frequency, alcohol consumption frequency, ongoing addictions to illicit drugs, and historical cannabis consumption). This degree of matching is unusually comprehensive (for comparison, in clinical trials and observational studies, it is rarely possible to match for more than two variables) and only made possible by the large resource of the UK Biobank.

Furthermore, we conducted a comprehensive comparison of 8 different connectivity measures and 17 classifiers, ensuring that training and test data were strictly separated throughout all analyses.

Concerning weaknesses, our study has a retrospective design which allows for less robust conclusions than from a prospective study. Furthermore, one potential weakness of the variants of DCM used in this study is that they all rely on variational Bayesian techniques, rendering model inversion susceptible to local extrema in the objective function (Daunizeau et al., 2011). In theory, this could have been addressed by a multi-start procedure, as in previous work with DCM (Schöbi et al., 2021; van Wijk et al., 2018). In practice, however, we were unable to implement this approach given that it would have led to an explosion of the already very substantial compute time. Finally, the greatest limitation of our study is the definition of depressive episodes. Given the heterogeneity of clinical data in the UK Biobank and the lack of systematic information about absence/presence of a clinical diagnosis of depression, we combined multiple sources of information within UK Biobank – i.e., clinical records, questionnaires (PHQ, MHQ) and self-report specifically on issues of depression – to identify indicators of at least one depressive episode within three years after the fMRI scan. Clearly, this partial reliance on self-report is not ideal; additionally, the resulting group of participants with a putative depressive episode (D+ group) is likely heterogeneous and might include people with very different severities of depression. Furthermore, there is rarely information on when exactly within the 3-year period a depressive episode occurred; the likely interindividual variability in the latency of symptom onset after the fMRI scan would further add to the heterogeneity of the D+ group. Having said this, our approach is similar to previous analyses of depression in the UK Biobank that also relied on self-report and questionnaires like the MHQ (Howard et al., 2020). More generally, a pragmatic approach to identifying individuals with likely clinical characteristics is often unavoidable when working with large heterogeneous databases (for an example using self-reported depression in genetics, see Wray et al., 2018). The challenge how to optimally extract data from the UK Biobank for studies of MDD is being addressed by ongoing methodological developments (Dutt et al., 2022) which will help to improve and standardise future studies.

Overall, our results have four implications. First, given the challenging nature of the prediction problem tackled in the study (i.e. occurrence of indicators of depressive episodes, as opposed to full clinical diagnoses, over a three year period), it is encouraging that significant predictions on held-out data can be obtained at all. Second, despite this success and the potential for further optimisation, our study suggests that fMRI on its own may not be sufficient for clinically useful predictions. Future studies of predicting depression should utilise fMRI-based connectivity estimates in conjunction with additional data (e.g. demographic, socioeconomic, clinical). Third, while GE results based on rDCM

722 were consistently successful across all classifiers and enjoyed a numerical advantage over FC for
723 clinical predictions, performance differences were modest and nonsignificant. The magnitude of
724 performance differences between GE and FC in this study and previous work suggests that adding
725 task-based fMRI may enhance the difference in predictive accuracy. Finally, using IC components as
726 network nodes diminishes the usual advantage of GE with regard to biological interpretability of
727 predictions. In order to maintain the interpretability of GE based predictions, it would seem
728 advantageous to compute effective connectivity between disjoint areas from parcellations based on
729 combined anatomical-functional criteria (e.g. Fan et al., 2016; Glasser et al., 2016). We hope that
730 these conclusions will be useful for future work on predicting the occurrence of depressive episodes.
731

Author contributions according to CRediT (Contributor Roles Taxonomy)

Herman Galioulline: Investigation, Software, Formal analysis, Writing – Original Draft, Visualization

Stefan Frässle: Investigation, Supervision, Validation, Software, Methodology, Writing – Review & Editing

Samuel Harrison: Investigation, Supervision, Validation, Software, Methodology, Writing – Review & Editing

Inês Pereira: Investigation, Validation, Writing – Review & Editing

Jakob Heinzle: Investigation, Supervision, Validation, Software, Methodology, Writing – Review & Editing

Klaas Enno Stephan: Investigation, Conceptualization, Supervision, Methodology, Project administration, Funding acquisition, Writing – Review & Editing

Acknowledgements

This work was supported by the René and Susanne Braginsky Foundation (KES), the ETH Foundation (KES), and project grant 320030_179377 by the Swiss National Science Foundation (KES).

References

Adler, D. A., McLaughlin, T. J., Rogers, W. H., Chang, H., Lapitsky, L., & Lerner, D. (2006). Job Performance Deficits Due to Depression. *American Journal of Psychiatry*, *163*(9), 1569–1576. https://doi.org/10.1176/ajp.2006.163.9.1569

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L. R., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P., Rorden, C., Daducci, A., Alexander, D. C., Zhang, H., Dragonu, I., Matthews, P. M., … Smith, S. M. (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

Barch, D. M., Harms, M. P., Tillman, R., Hawkey, E., & Luby, J. L. (2019). Early Childhood Depression, Emotion Regulation, Episodic Memory and Hippocampal Development. *Journal of Abnormal Psychology*, *128*(1), 81–95. https://doi.org/10.1037/abn0000392

Berwian, I. M., Wenzel, J. G., Kuehn, L., Schnuerer, I., Kasper, L., Veer, I. M., Seifritz, E., Stephan, K. E., Walter, H., & Huys, Q. J. M. (2020). The relationship between resting-state functional connectivity, antidepressant discontinuation and depression relapse. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-79170-9

Brakowski, J., Spinelli, S., Dörig, N., Bosch, O. G., Manoliu, A., Holtforth, M. G., & Seifritz, E. (2017). Resting state brain network function in major depression—Depression symptomatology, antidepressant treatment effects, future research. *Journal of Psychiatric Research*, *92*, 147–159. https://doi.org/10.1016/j.jpsychires.2017.04.007

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brodersen, K. H., Deserno, L., Schlagenhauf, F., Lin, Z., Penny, W. D., Buhmann, J. M., & Stephan, K. E. (2014). Dissecting psychiatric spectrum disorders by generative embedding. *NeuroImage: Clinical*, *4*, 98–111. https://doi.org/10.1016/j.nicl.2013.11.002

Brodersen, K. H., Schofield, T. M., Leff, A. P., Ong, C. S., Lomakina, E. I., Buhmann, J. M., & Stephan, K. E. (2011). Generative Embedding for Model-Based Classification of fMRI Data. *PLOS Computational Biology*, *7*(6), e1002079. https://doi.org/10.1371/journal.pcbi.1002079

Caldirola, D., Daccò, S., Cuniberti, F., Grassi, M., Alciati, A., Torti, T., & Perna, G. (2022). First-onset major depression during the COVID-19 pandemic: A predictive machine learning model. *Journal of Affective Disorders*, *310*, 75–86. https://doi.org/10.1016/j.jad.2022.04.145

Cawley, G. C., & Talbot, N. L. C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, *11*(70), 2079–2107.

Chikersal, P., Doryab, A., Tumminia, M., Villalba, D. K., Dutcher, J. M., Liu, X., Cohen, S., Creswell, K. G., Mankoff, J., Creswell, J. D., Goel, M., & Dey, A. K. (2021). Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach With Robust Feature Selection. *ACM Transactions on Computer-Human Interaction*, *28*(1), 3:1-3:41. https://doi.org/10.1145/3422821

Cole, J. H., Poudel, R. P. K., Tsagkrasoulis, D., Caan, M. W. A., Steves, C., Spector, T. D., & Montana, G. (2017). Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, *163*, 115–124. https://doi.org/10.1016/j.neuroimage.2017.07.059

Coleman, J. R. I., Peyrot, W. J., Purves, K. L., Davis, K. A. S., Rayner, C., Choi, S. W., Hübel, C., Gaspar, H. A., Kan, C., Van der Auwera, S., Adams, M. J., Lyall, D. M., Choi, K. W., Dunn, E. C., Vassos, E., Danese, A., Maughan, B., Grabe, H. J., Lewis, C. M., … Breen, G. (2020). Genome-wide gene-environment analyses of major depressive disorder and reported lifetime traumatic experiences in UK Biobank. *Molecular Psychiatry*, *25*(7), Article 7. https://doi.org/10.1038/s41380-019-0546-6

Correll, C. U., Solmi, M., Veronese, N., Bortolato, B., Rosson, S., Santonastaso, P., Thapa-Chhetri, N., Fornaro, M., Gallicchio, D., Collantoni, E., Pigato, G., Favaro, A., Monaco, F., Kohler, C., Vancampfort, D., Ward, P. B., Gaughran, F., Carvalho, A. F., & Stubbs, B. (2017). Prevalence, incidence and mortality from cardiovascular disease in patients with pooled and specific severe mental illness: A large-scale meta-analysis of 3,211,768 patients and 113,383,368 controls. *World Psychiatry*, *16*(2), 163–180. https://doi.org/10.1002/wps.20420

Cover. (1965). Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Transactions on Electronic Computers*, *EC-14*(3), 326–334. https://doi.org/10.1109/PGEC.1965.264137

Cover, & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

Cuijpers, P., Beekman, A. T. F., & Reynolds, C. F. (2012). Preventing Depression: A Global Priority. *JAMA*, *307*(10), 1033–1034. https://doi.org/10.1001/jama.2012.271

Cuijpers, P., Pineda, B. S., Quero, S., Karyotaki, E., Struijs, S. Y., Figueroa, C. A., Llamas, J. A., Furukawa, T. A., & Muñoz, R. F. (2021). Psychological interventions to prevent the onset of depressive disorders: A meta-analysis of randomized controlled trials. *Clinical Psychology Review*, *83*, 101955. https://doi.org/10.1016/j.cpr.2020.101955

Daunizeau, J., David, O., & Stephan, K. E. (2011). Dynamic causal modelling: A critical review of the biophysical and statistical foundations. *NeuroImage*, *58*(2), 312–322. https://doi.org/10.1016/j.neuroimage.2009.11.062

Dutt, R. K., Hannon, K., Easley, T. O., Griffis, J. C., Zhang, W., & Bijsterbosch, J. D. (2022). Mental health in the UK Biobank: A roadmap to self-report measures and neuroimaging correlates. *Human Brain Mapping*, *43*(2), 816–832. https://doi.org/10.1002/hbm.25690

Eaton, W. W., Shao, H., Nestadt, G., Lee, B. H., Bienvenu, O. J., & Zandi, P. (2008). Population-Based Study of First Onset and Chronicity in Major Depressive Disorder. *Archives of General Psychiatry*, *65*(5), 513–520. https://doi.org/10.1001/archpsyc.65.5.513

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2016). The Human Brainnetome Atlas: A New Brain Atlas Based on Connectional Architecture. *Cerebral Cortex*, *26*(8), 3508–3526. https://doi.org/10.1093/cercor/bhw157

Frässle, S., Aponte, E. A., Bollmann, S., Brodersen, K. H., Do, C. T., Harrison, O. K., Harrison, S. J., Heinzle, J., Iglesias, S., Kasper, L., Lomakina, E. I., Mathys, C., Müller-Schrader, M., Pereira, I., Petzschner, F. H., Raman, S., Schöbi, D., Toussaint, B., Weber, L. A., … Stephan, K. E. (2021). TAPAS: An Open-Source Software Package for Translational Neuromodeling and Computational Psychiatry. *Frontiers in Psychiatry*, *12*, 680811. https://doi.org/10.3389/fpsyt.2021.680811

Frässle, S., Harrison, S. J., Heinzle, J., Clementz, B. A., Tamminga, C. A., Sweeney, J. A., Gershon, E. S., Keshavan, M. S., Pearlson, G. D., Powers, A., & Stephan, K. E. (2021). Regression dynamic causal

833 modeling for resting-state fMRI. *Human Brain Mapping*, *42*(7), 2159–2180.
834 https://doi.org/10.1002/hbm.25357

835 Frässle, S., Marquand, A. F., Schmaal, L., Dinga, R., Veltman, D. J., van der Wee, N. J. A., van Tol, M.-
836 J., Schöbi, D., Penninx, B. W. J. H., & Stephan, K. E. (2020). Predicting individual clinical trajectories of
837 depression with generative embedding. *NeuroImage: Clinical*, *26*, 102213.
838 https://doi.org/10.1016/j.nicl.2020.102213

839 Frässle, S., & Stephan, K. E. (2022). Test-retest reliability of regression dynamic causal modeling.
840 *Network Neuroscience*, *6*(1), 135–160. https://doi.org/10.1162/netn_a_00215

841 Frässle, S., Yao, Y., Schöbi, D., Aponte, E. A., Heinzle, J., & Stephan, K. E. (2018). Generative models
842 for clinical applications in computational psychiatry. *WIREs Cognitive Science*, *9*(3), e1460.
843 https://doi.org/10.1002/wcs.1460

844 Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an
845 Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139.
846 https://doi.org/10.1006/jcss.1997.1504

847 Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of*
848 *Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

849 Friston, K. J. (2011, June 1). *Functional and Effective Connectivity: A Review* (140 Huguenot Street,
850 3rd Floor New Rochelle, NY 10801 USA) [Research-article]. Https://Home.Liebertpub.Com/Brain;
851 Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA.
852 https://doi.org/10.1089/brain.2011.0008

853 Friston, K. J., Kahan, J., Biswal, B., & Razi, A. (2014). A DCM for resting state fMRI. *NeuroImage*, *94*,
854 396–407. https://doi.org/10.1016/j.neuroimage.2013.12.009

855 GBD 2019 Mental Disorders Collaborators. (2022). Global, regional, and national burden of 12
856 mental disorders in 204 countries and territories, 1990–2019: A systematic analysis for the Global
857 Burden of Disease Study 2019. *The Lancet Psychiatry*, *9*(2), 137–150. https://doi.org/10.1016/S2215-
858 0366(21)00395-3

859 Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K.,
860 Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal
861 parcellation of human cerebral cortex. *Nature*, *536*(7615), Article 7615.
862 https://doi.org/10.1038/nature18933

863 Gordon, J. A. (2016). On being a circuit psychiatrist. *Nature Neuroscience*, *19*(11), Article 11.
864 https://doi.org/10.1038/nn.4419

865 Goulden, N., Khusnulina, A., Davis, N. J., Bracewell, R. M., Bokde, A. L., McNulty, J. P., & Mullins, P. G.
866 (2014). The salience network is responsible for switching between the default mode network and
867 the central executive network: Replication from DCM. *NeuroImage*, *99*, 180–190.
868 https://doi.org/10.1016/j.neuroimage.2014.05.052

869 Gratton, C., Sun, H., & Petersen, S. E. (2018). Control networks and hubs. *Psychophysiology*, *55*(3).
870 https://doi.org/10.1111/psyp.13032

871 Gu, S.-C., Zhou, J., Yuan, C.-X., & Ye, Q. (2020). Personalized prediction of depression in patients with
872 newly diagnosed Parkinson's disease: A prospective cohort study. *Journal of Affective Disorders*, *268*,
873 118–126. https://doi.org/10.1016/j.jad.2020.02.046

874 Harris, J. K., Hassel, S., Davis, A. D., Zamyadi, M., Arnott, S. R., Milev, R., Lam, R. W., Frey, B. N., Hall,
875 G. B., Müller, D. J., Rotzinger, S., Kennedy, S. H., Strother, S. C., MacQueen, G. M., & Greiner, R.
876 (2022). Predicting escitalopram treatment response from pre-treatment and early response resting
877 state fMRI in a multi-site sample: A CAN-BIND-1 report. *NeuroImage: Clinical*, *35*, 103120.
878 https://doi.org/10.1016/j.nicl.2022.103120

879 He, T., Kong, R., Holmes, A. J., Nguyen, M., Sabuncu, M. R., Eickhoff, S. B., Bzdok, D., Feng, J., & Yeo,
880 B. T. T. (2020). Deep neural networks and kernel regression achieve comparable accuracies for
881 functional connectivity prediction of behavior and demographics. *NeuroImage*, *206*, 116276.
882 https://doi.org/10.1016/j.neuroimage.2019.116276

883 Hirshfeld-Becker, D. R., Gabrieli, J. D. E., Shapero, B. G., Biederman, J., Whitfield-Gabrieli, S., & Chai,
884 X. J. (2019). Intrinsic Functional Brain Connectivity Predicts Onset of Major Depression Disorder in
885 Adolescence: A Pilot Study. *Brain Connectivity*, *9*(5), 388–398.
886 https://doi.org/10.1089/brain.2018.0646

887 Hopman, H. J., Chan, S. M. S., Chu, W. C. W., Lu, H., Tse, C.-Y., Chau, S. W. H., Lam, L. C. W., Mak, A.
888 D. P., & Neggers, S. F. W. (2021). Personalized prediction of transcranial magnetic stimulation clinical
889 response in patients with treatment-refractory depression using neuroimaging biomarkers and
890 machine learning. *Journal of Affective Disorders*, *290*, 261–271.
891 https://doi.org/10.1016/j.jad.2021.04.081

892 Howard, D. M., Folkersen, L., Coleman, J. R. I., Adams, M. J., Glanville, K., Werge, T., Hagenaars, S. P.,
893 Han, B., Porteous, D., Campbell, A., Clarke, T.-K., Breen, G., Sullivan, P. F., Wray, N. R., Lewis, C. M., &
894 McIntosh, A. M. (2020). Genetic stratification of depression in UK Biobank. *Translational Psychiatry*,
895 *10*(1), Article 1. https://doi.org/10.1038/s41398-020-0848-0

896 Hyett, M. P., Parker, G. B., Guo, C. C., Zalesky, A., Nguyen, V. T., Yuen, T., & Breakspear, M. (2015).
897 Scene unseen: Disrupted neuronal adaptation in melancholia during emotional film viewing.
898 *NeuroImage: Clinical*, *9*, 660–667. https://doi.org/10.1016/j.nicl.2015.10.011

899 Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL.
900 *NeuroImage*, *62*(2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

901 Ju, Y., Horien, C., Chen, W., Guo, W., Lu, X., Sun, J., Dong, Q., Liu, B., Liu, J., Yan, D., Wang, M., Zhang,
902 L., Guo, H., Zhao, F., Zhang, Y., Shen, X., Constable, R. T., & Li, L. (2020). Connectome-based models
903 can predict early symptom improvement in major depressive disorder. *Journal of Affective Disorders*,
904 *273*, 442–452. https://doi.org/10.1016/j.jad.2020.04.028

905 Kaiser, R. H., Andrews-Hanna, J. R., Wager, T. D., & Pizzagalli, D. A. (2015). Large-Scale Network
906 Dysfunction in Major Depressive Disorder: A Meta-analysis of Resting-State Functional Connectivity.
907 *JAMA Psychiatry*, *72*(6), 603–611. https://doi.org/10.1001/jamapsychiatry.2015.0071

908 King, M., Walker, C., Levy, G., Bottomley, C., Royston, P., Weich, S., Bellón-Saameño, J. Á., Moreno,
909 B., Švab, I., Rotar, D., Rifel, J., Maaroos, H.-I., Aluoja, A., Kalda, R., Neeleman, J., Geerlings, M. I.,
910 Xavier, M., Carraça, I., Gonçalves-Pereira, M., … Nazareth, I. (2008). Development and Validation of
911 an International Risk Prediction Algorithm for Episodes of Major Depression in General Practice
912 Attendees: The PredictD Study. *Archives of General Psychiatry*, *65*(12), 1368–1376.
913 https://doi.org/10.1001/archpsyc.65.12.1368

914 Kupferberg, A., Bicks, L., & Hasler, G. (2016). Social functioning in major depressive disorder.
915 *Neuroscience & Biobehavioral Reviews*, *69*, 313–332.
916 https://doi.org/10.1016/j.neubiorev.2016.07.002

917 Lawrence, A. J., Stahl, D., Duan, S., Fennema, D., Jaeckle, T., Young, A. H., Dazzan, P., Moll, J., & Zahn,
918 R. (2022). Neurocognitive Measures of Self-blame and Risk Prediction Models of Recurrence in Major
919 Depressive Disorder. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *7*(3), 256–264.
920 https://doi.org/10.1016/j.bpsc.2021.06.010

921 Li, B., Daunizeau, J., Stephan, K. E., Penny, W., Hu, D., & Friston, K. (2011). Generalised filtering and
922 stochastic DCM for fMRI. *NeuroImage*, *58*(2), 442–457.
923 https://doi.org/10.1016/j.neuroimage.2011.01.085

924 Librenza-Garcia, D., Passos, I. C., Feiten, J. G., Lotufo, P. A., Goulart, A. C., Santos, I. de S., Viana, M.
925 C., Benseñor, I. M., & Brunoni, A. R. (2021). Prediction of depression cases, incidence, and chronicity
926 in a large occupational cohort using machine learning techniques: An analysis of the ELSA-Brasil
927 study. *Psychological Medicine*, *51*(16), 2895–2903. https://doi.org/10.1017/S0033291720001579

928 Lin, S., Wu, Y., He, L., & Fang, Y. (2022). Prediction of depressive symptoms onset and long-term
929 trajectories in home-based older adults using machine learning techniques. *Aging & Mental Health*,
930 *0*(0), 1–10. https://doi.org/10.1080/13607863.2022.2031868

931 Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I.
932 Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.),
933 *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc.
934 https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

935 Merikangas, K. R., Wicki, W., & Angst, J. (1994). Heterogeneity of Depression: Classification of
936 Depressive Subtypes by Longitudinal Course. *The British Journal of Psychiatry*, *164*(3), 342–348.
937 https://doi.org/10.1192/bjp.164.3.342

938 Miller, K. L., Alfaro-Almagro, F., Bangerter, N. K., Thomas, D. L., Yacoub, E., Xu, J., Bartsch, A. J.,
939 Jbabdi, S., Sotiropoulos, S. N., Andersson, J. L. R., Griffanti, L., Douaud, G., Okell, T. W., Weale, P.,
940 Dragonu, I., Garratt, S., Hudson, S., Collins, R., Jenkinson, M., … Smith, S. M. (2016). Multimodal
941 population brain imaging in the UK Biobank prospective epidemiological study. *Nature Neuroscience*,
942 *19*(11), Article 11. https://doi.org/10.1038/nn.4393

943 Motlaghian, S. M., Belger, A., Bustillo, J. R., Ford, J. M., Iraji, A., Lim, K., Mathalon, D. H., Mueller, B.
944 A., O'Leary, D., Pearlson, G., Potkin, S. G., Preda, A., van Erp, T. G. M., & Calhoun, V. D. (2022).
945 Nonlinear functional network connectivity in resting functional magnetic resonance imaging data.
946 *Human Brain Mapping*, *43*(15), 4556–4566. https://doi.org/10.1002/hbm.25972

947 Na, K.-S., Cho, S.-E., Geem, Z. W., & Kim, Y.-K. (2020). Predicting future onset of depression among
948 community dwelling adults in the Republic of Korea using a machine learning algorithm.
949 *Neuroscience Letters*, *721*, 134804. https://doi.org/10.1016/j.neulet.2020.134804

950 Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines.
951 *Proceedings of the 27th International Conference on International Conference on Machine Learning*,
952 807–814.

953 Nickerson, L. D., Smith, S. M., Öngür, D., & Beckmann, C. F. (2017). Using Dual Regression to
954 Investigate Network Shape and Amplitude in Functional Connectivity Analyses. *Frontiers in*
955 *Neuroscience*, *11*. https://www.frontiersin.org/articles/10.3389/fnins.2017.00115

Noble, S., Scheinost, D., & Constable, R. T. (2019). A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *NeuroImage*, *203*, 116157. https://doi.org/10.1016/j.neuroimage.2019.116157

Osuch, E., Gao, S., Wammes, M., Théberge, J., Williamson, P., Neufeld, R. J., Du, Y., Sui, J., & Calhoun, V. (2018). Complexity in mood disorder diagnosis: FMRI connectivity networks predicted medication-class of response in complex patients. *Acta Psychiatrica Scandinavica*, *138*(5), 472–482. https://doi.org/10.1111/acps.12945

Pagliaccio, D., Luby, J. L., Luking, K. R., Belden, A. C., & Barch, D. M. (2014). Brain–behavior relationships in the experience and regulation of negative emotion in healthy children: Implications for risk for childhood depression. *Development and Psychopathology*, *26*(4pt2), 1289–1303. https://doi.org/10.1017/S0954579414001035

Papmeyer, M., Sussmann, J. E., Stewart, T., Giles, S., Centola, J. G., Zannias, V., Lawrie, S. M., Whalley, H. C., & McIntosh, A. M. (2016). Prospective longitudinal study of subcortical brain volumes in individuals at high familial risk of mood disorders with or without subsequent onset of depression. *Psychiatry Research: Neuroimaging*, *248*, 119–125. https://doi.org/10.1016/j.pscychresns.2015.12.009

Platt, J. C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, 61–74.

Poldrack, R. A., Huckins, G., & Varoquaux, G. (2020). Establishment of Best Practices for Evidence for Prediction: A Review. *JAMA Psychiatry*, *77*(5), 534–540. https://doi.org/10.1001/jamapsychiatry.2019.3671

Queirazza, F., Fouragnan, E., Steele, J. D., Cavanagh, J., & Philiastides, M. G. (2019). Neural correlates of weighted reward prediction error during reinforcement learning classify response to cognitive behavioral therapy in depression. *Science Advances*, *5*(7), eaav4962. https://doi.org/10.1126/sciadv.aav4962

Raposo Pereira, F., Zhutovsky, P., Mcmaster, M. T. B., Polderman, N., de Vries, Y. D. A. T., van den Brink, W., & van Wingen, G. A. (2019). Recreational use of GHB is associated with alterations of resting state functional connectivity of the central executive and default mode networks. *Human Brain Mapping*, *40*(8), 2413–2421. https://doi.org/10.1002/hbm.24532

Rocha, T. B.-M., Fisher, H. L., Caye, A., Anselmi, L., Arseneault, L., Barros, F. C., Caspi, A., Danese, A., Gonçalves, H., Harrington, H. L., Houts, R., Menezes, A. M. B., Moffitt, T. E., Mondelli, V., Poulton, R., Rohde, L. A., Wehrmeister, F., & Kieling, C. (2021). Identifying Adolescents at Risk for Depression: A Prediction Score Performance in Cohorts Based in 3 Different Continents. *Journal of the American Academy of Child & Adolescent Psychiatry*, *60*(2), 262–273. https://doi.org/10.1016/j.jaac.2019.12.004

Rosellini, A. J., Liu, S., Anderson, G. N., Sbi, S., Tung, E. S., & Knyazhanskaya, E. (2020). Developing algorithms to predict adult onset internalizing disorders: An ensemble learning approach. *Journal of Psychiatric Research*, *121*, 189–196. https://doi.org/10.1016/j.jpsychires.2019.12.006

Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., Niederehe, G., Thase, M. E., Lavori, P. W., Lebowitz, B. D., McGrath, P. J., Rosenbaum, J. F., Sackeim, H. A., Kupfer, D. J., Luther, J., & Fava, M. (2006). Acute and Longer-Term Outcomes in Depressed

Outpatients Requiring One or Several Treatment Steps: A STAR*D Report. *American Journal of Psychiatry*, *163*(11), 1905–1917. https://doi.org/10.1176/ajp.2006.163.11.1905

Sampson, L., Jiang, T., Gradus, J. L., Cabral, H. J., Rosellini, A. J., Calabrese, J. R., Cohen, G. H., Fink, D. S., King, A. P., Liberzon, I., & Galea, S. (2021). A Machine Learning Approach to Predicting New-onset Depression in a Military Population. *Psychiatric Research and Clinical Practice*, *3*(3), 115–122. https://doi.org/10.1176/appi.prcp.20200031

Schmaal, L., Marquand, A. F., Rhebergen, D., van Tol, M.-J., Ruhé, H. G., van der Wee, N. J. A., Veltman, D. J., & Penninx, B. W. J. H. (2015). Predicting the Naturalistic Course of Major Depressive Disorder Using Clinical and Multimodal Neuroimaging Information: A Multivariate Pattern Recognition Study. *Biological Psychiatry*, *78*(4), 278–286. https://doi.org/10.1016/j.biopsych.2014.11.018

Schöbi, D., Homberg, F., Frässle, S., Endepols, H., Moran, R. J., Friston, K. J., Tittgemeyer, M., Heinzle, J., & Stephan, K. E. (2021). Model-based prediction of muscarinic receptor function from auditory mismatch negativity responses. *NeuroImage*, *237*, 118096. https://doi.org/10.1016/j.neuroimage.2021.118096

Schulz, M.-A., Yeo, B. T. T., Vogelstein, J. T., Mourao-Miranada, J., Kather, J. N., Kording, K., Richards, B., & Bzdok, D. (2020). Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-18037-z

Shapero, B. G., Chai, X. J., Vangel, M., Biederman, J., Hoover, C. S., Whitfield-Gabrieli, S., Gabrieli, J. D. E., & Hirshfeld-Becker, D. R. (2019). Neural markers of depression risk predict the onset of depression. *Psychiatry Research: Neuroimaging*, *285*, 31–39. https://doi.org/10.1016/j.pscychresns.2019.01.006

Shapley, L. S. (1953). 17. A Value for n-Person Games. In *17. A Value for n-Person Games* (pp. 307–318). Princeton University Press. https://doi.org/10.1515/9781400881970-018

Shen, X., Cox, S. R., Adams, M. J., Howard, D. M., Lawrie, S. M., Ritchie, S. J., Bastin, M. E., Deary, I. J., McIntosh, A. M., & Whalley, H. C. (2018). Resting-State Connectivity and Its Association With Cognitive Performance, Educational Attainment, and Household Income in the UK Biobank. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(10), 878–886. https://doi.org/10.1016/j.bpsc.2018.06.007

Smith, S. M., Fox, P. T., Miller, K. L., Glahn, D. C., Fox, P. M., Mackay, C. E., Filippini, N., Watkins, K. E., Toro, R., Laird, A. R., & Beckmann, C. F. (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences*, *106*(31), 13040–13045. https://doi.org/10.1073/pnas.0905267106

Steffen, A., Nübel, J., Jacobi, F., Bätzing, J., & Holstiege, J. (2020). Mental and somatic comorbidity of depression: A comprehensive cross-sectional analysis of 202 diagnosis groups using German nationwide ambulatory claims data. *BMC Psychiatry*, *20*(1), 142. https://doi.org/10.1186/s12888-020-02546-8

Stephan, K. E., Iglesias, S., Heinzle, J., & Diaconescu, A. O. (2015). Translational Perspectives for Computational Neuroimaging. *Neuron*, *87*(4), 716–732. https://doi.org/10.1016/j.neuron.2015.07.008

1038 Stephan, K. E., Schlagenhauf, F., Huys, Q. J. M., Raman, S., Aponte, E. A., Brodersen, K. H., Rigoux, L.,
1039 Moran, R. J., Daunizeau, J., Dolan, R. J., Friston, K. J., & Heinz, A. (2017). Computational
1040 neuroimaging strategies for single patient predictions. *NeuroImage*, *145*, 180–199.
1041 https://doi.org/10.1016/j.neuroimage.2016.06.038

1042 Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the*
1043 *Royal Statistical Society. Series B (Methodological)*, *36*(2), 111–147.

1044 van Eeden, W. A., Luo, C., van Hemert, A. M., Carlier, I. V. E., Penninx, B. W., Wardenaar, K. J., Hoos,
1045 H., & Giltay, E. J. (2021). Predicting the 9-year course of mood and anxiety disorders with automated
1046 machine learning: A comparison between auto-sklearn, naïve Bayes classifier, and traditional logistic
1047 regression. *Psychiatry Research*, *299*, 113823. https://doi.org/10.1016/j.psychres.2021.113823

1048 van Wijk, B. C. M., Cagnan, H., Litvak, V., Kühn, A. A., & Friston, K. J. (2018). Generic dynamic causal
1049 modelling: An illustrative application to Parkinson's disease. *NeuroImage*, *181*, 818–830.
1050 https://doi.org/10.1016/j.neuroimage.2018.08.039

1051 Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars.
1052 *NeuroImage*, *180*, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

1053 Voorhees, B. W. V., Paunesku, D., Gollan, J., Kuwabara, S., Reinecke, M., & Basu, A. (2008). Predicting
1054 Future Risk of Depressive Episode in Adolescents: The Chicago Adolescent Depression Risk
1055 Assessment (CADRA). *The Annals of Family Medicine*, *6*(6), 503–511.
1056 https://doi.org/10.1370/afm.887

1057 Vos, T., Lim, S. S., Abbafati, C., Abbas, K. M., Abbasi, M., Abbasifard, M., Abbasi-Kangevari, M.,
1058 Abbastabar, H., Abd-Allah, F., Abdelalim, A., Abdollahi, M., Abdollahpour, I., Abolhassani, H.,
1059 Aboyans, V., Abrams, E. M., Abreu, L. G., Abrigo, M. R. M., Abu-Raddad, L. J., Abushouk, A. I., …
1060 Murray, C. J. L. (2020). Global burden of 369 diseases and injuries in 204 countries and territories,
1061 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*,
1062 *396*(10258), 1204–1222. https://doi.org/10.1016/S0140-6736(20)30925-9

1063 Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., Adams, M. J.,
1064 Agerbo, E., Air, T. M., Andlauer, T. M. F., Bacanu, S.-A., Bækvad-Hansen, M., Beekman, A. F. T.,
1065 Bigdeli, T. B., Binder, E. B., Blackwood, D. R. H., Bryois, J., Buttenschøn, H. N., Bybjerg-Grauholm, J., …
1066 Sullivan, P. F. (2018). Genome-wide association analyses identify 44 risk variants and refine the
1067 genetic architecture of major depression. *Nature Genetics*, *50*(5), Article 5.
1068 https://doi.org/10.1038/s41588-018-0090-3

1069 Xia, Y., Chen, Q., Shi, L., Li, M., Gong, W., Chen, H., & Qiu, J. (2018). Tracking the dynamic functional
1070 connectivity structure of the human brain across the adult lifespan. *Human Brain Mapping*, *40*(3),
1071 717–728. https://doi.org/10.1002/hbm.24385

1072 Xu, Z., Zhang, Q., Li, W., Li, M., & Yip, P. S. F. (2019). Individualized prediction of depressive disorder
1073 in the elderly: A multitask deep learning approach. *International Journal of Medical Informatics*, *132*,
1074 103973. https://doi.org/10.1016/j.ijmedinf.2019.103973

1075 Zhang, H. (2004). *The Optimality of Naive Bayes*. *2*.

1076