frontiers | Frontiers in Psychiatry

TYPE Original Research
PUBLISHED 18 October 2022
DOI 10.3389/fpsyt.2022.938694



OPEN ACCESS

EDITED BY
Glenn Saxe,
New York University, United States

REVIEWED BY
John Kounios,
Drexel University, United States
Vineet Tiruvadi,
Emory University, United States

*CORRESPONDENCE
D. A. Pinotsis
pinotsis@mit.edu

SPECIALTY SECTION
This article was submitted to Computational Psychiatry, a section of the journal Frontiers in Psychiatry

RECEIVED 07 May 2022
ACCEPTED 22 September 2022
PUBLISHED 18 October 2022

CITATION
Pinotsis DA, Fitzgerald S, See C, Sementsova A and Widge AS (2022) Toward biophysical markers of depression vulnerability. *Front. Psychiatry* 13:938694. doi: 10.3389/fpsyt.2022.938694



# Toward biophysical markers of depression vulnerability

D. A. Pinotsis[1,2]*, S. Fitzgerald[1], C. See[3], A. Sementsova[3] and A. S. Widge[4]

[1]Centre for Mathematical Neuroscience and Psychology, Department of Psychology, City, University of London, London, United Kingdom, [2]The Picower Institute for Learning and Memory, Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, United States, [3]Department of Computer Science, City, University of London, London, United Kingdom, [4]Department of Psychiatry and Behavioral Sciences, University of Minnesota, Minneapolis, MN, United States

A major difficulty with treating psychiatric disorders is their heterogeneity: different neural causes can lead to the same phenotype. To address this, we propose describing the underlying pathophysiology in terms of interpretable, biophysical parameters of a neural model derived from the electroencephalogram. We analyzed data from a small patient cohort of patients with depression and controls. Using DCM, we constructed biophysical models that describe neural dynamics in a cortical network activated during a task that is used to assess depression state. We show that biophysical model parameters are biomarkers, that is, variables that allow subtyping of depression at a biological level. They yield a low dimensional, interpretable feature space that allowed description of differences between individual patients with depressive symptoms. They could capture internal heterogeneity/variance of depression state and achieve significantly better classification than commonly used EEG features. Our work is a proof of concept that a combination of biophysical models and machine learning may outperform earlier approaches based on classical statistics and raw brain data.

KEYWORDS

**depression, dynamic causal modeling (DCM), biomarkers, event-related potentials (ERPs), machine learning**

## Introduction

Depression affects roughly one in six people (1), and its prevalence may be increasing (2). A major difficulty with treating depression, and psychiatric disorders in general, is their heterogeneity: a clinical phenotype or classification can arise from multiple different neural causes (3, 4). To address this heterogeneity, we propose describing

Abbreviations: DCM, Dynamic Causal Modeling; ERP, Event-related Potentials; MSIT, Multi-Source Interference Task; JR, Jansen and Rit mass model; ICA, Independent Component Analysis; MCC, Matthew's Correlation Coefficient; SHAP, Shapley Additive Explanation; PCA, Principal Component Analysis.

depression in terms of interpretable, biophysical parameters of a neural model, derived from the electroencephalogram (EEG). These parameters may serve as biomarkers, variables that allow subtyping of depression at a biological level. They can be thought of as latent variables that may capture individual differences between patients. As a proof of concept, we show that this idea works even in a small patient cohort.

Several studies have used EEG data to identify potential biomarkers for psychiatric disorders (5–8). These studies emphasized EEG features: components of event-related potentials (ERPs) or oscillatory responses, not biophysical parameters. For example, multiple papers report smaller N1 amplitudes in depression (6, 9, 10). However, the results from these studies are difficult to connect back to biology: The data features (e.g., ERPs, oscillatory responses, and resting state activity) do not directly map back to brain structures or to physiologic changes. There are exceptions, e.g., the loudness dependence of the auditory evoked potential (11–13), but in general these analyses are mainly phenomenological. They also fail to consider depression's internal heterogeneity, which limits generalizability of the derived biomarkers. A recent meta-analysis suggested that no EEG marker had reliable clinical utility (14), although newer work has tried to address this (15, 16).

One approach to overcoming depression's heterogeneity might be to shift the level of analysis. For instance, source localization techniques can interpret scalp phenomena in terms of their underlying cortical generators (8, 17). Still these methods emphasize waves/patterns in the electrical activity whose neural basis remains unclear. A deeper level of analysis more grounded in cellular physiology may be possible when using biophysical models. This is the approach we take here. Their parameters describe the neurobiology or neural populations (e.g., synaptic time constants, intrinsic, and extrinsic connectivity) that give rise to the scalp-recorded patterns. They capture important developmental, structural and functional properties of cortical sources. Synaptic time constants are important for determining the EEG signal (18). Intrinsic and extrinsic connectivity go through characteristic changes throughout the development of the brain and can exhibit differences with age or in the presence of a disease (19). For instance, we and others have used biophysical models to analyze data from patients with neurological diseases recorded using M/EEG and fMRI (20–24).

Here, we constructed biophysical models using Dynamic Causal Modelling (DCM). These describe the cortical network activated during a cognitive conflict task that activates depression-relevant brain areas (25–27). Our model transforms high-dimensional EEG data onto a mechanistically interpretable feature space (20); in which, we show below that we can better measure depression's internal heterogeneity. We present a proof of concept for the following idea: that biophysical model parameters yield a low dimensional, interpretable feature space. As a result of that better capture of internal heterogeneity/variance, model-derived features achieved significantly better classification than manifest EEG features. Our work shows that a combination of biophysical models and machine learning may be an alternative to earlier approaches based on classical statistics and raw brain data.

# Materials and methods

## Dataset

The dataset included 15 psychiatric patients who reported current or past depressive symptoms and 34 non-diagnosed controls. Importantly, this dataset was not limited to patients diagnosed with unipolar depression, but included bipolar and unspecified depression. We considered this a better demonstration of our approach to heterogeneity. This is a secondary analysis of a cohort collected in a previous study (28). For details of the EEG recordings, see C Methods.

Electroencephalogram (EEGs) were collected as participants performed the Multi-Source Interference Task (MSIT), see **Figure 1C** (29). MSIT has been validated to produce robust cortical activations at the single-participant level, in both fMRI (29) and EEG (30) studies, and is used for assessing depression state. For more details regarding the task and dataset, see **Supplementary methods**.

## Event-related potentials analysis

We emphasized a Positive Potential signal, defined as evoked responses between 250 and 350°ms after event onset. Positive Potential components were extracted from 70 EEG channels (average ERPs over participants are included in **Figure 1A**; see also **Supplementary Figure 1**). Previous MSIT studies found differences in similar Positive Potential components between task conditions (31, 32). These early Positive Potentials are a common signature of conflict and cognitive control, and arise when incongruent stimuli are processed (33, 34). We used Positive Potential peak amplitude and latency as EEG classification features (see Class Balancing and Model Training section below). To test for the effect of conditions (interference vs. control), Wilcoxon tests were conducted on each channel to compare the peak amplitude and latencies. Latency exhibited significant differences, see **Supplementary Table 1**. See also **Supplementary methods** for more details on the ERP extraction pipeline.

## Dynamic causal modeling

We used Dynamic Causal Modelling (DCM) (20–22, 35–40) to infer processes at the neuronal level from scalp EEG

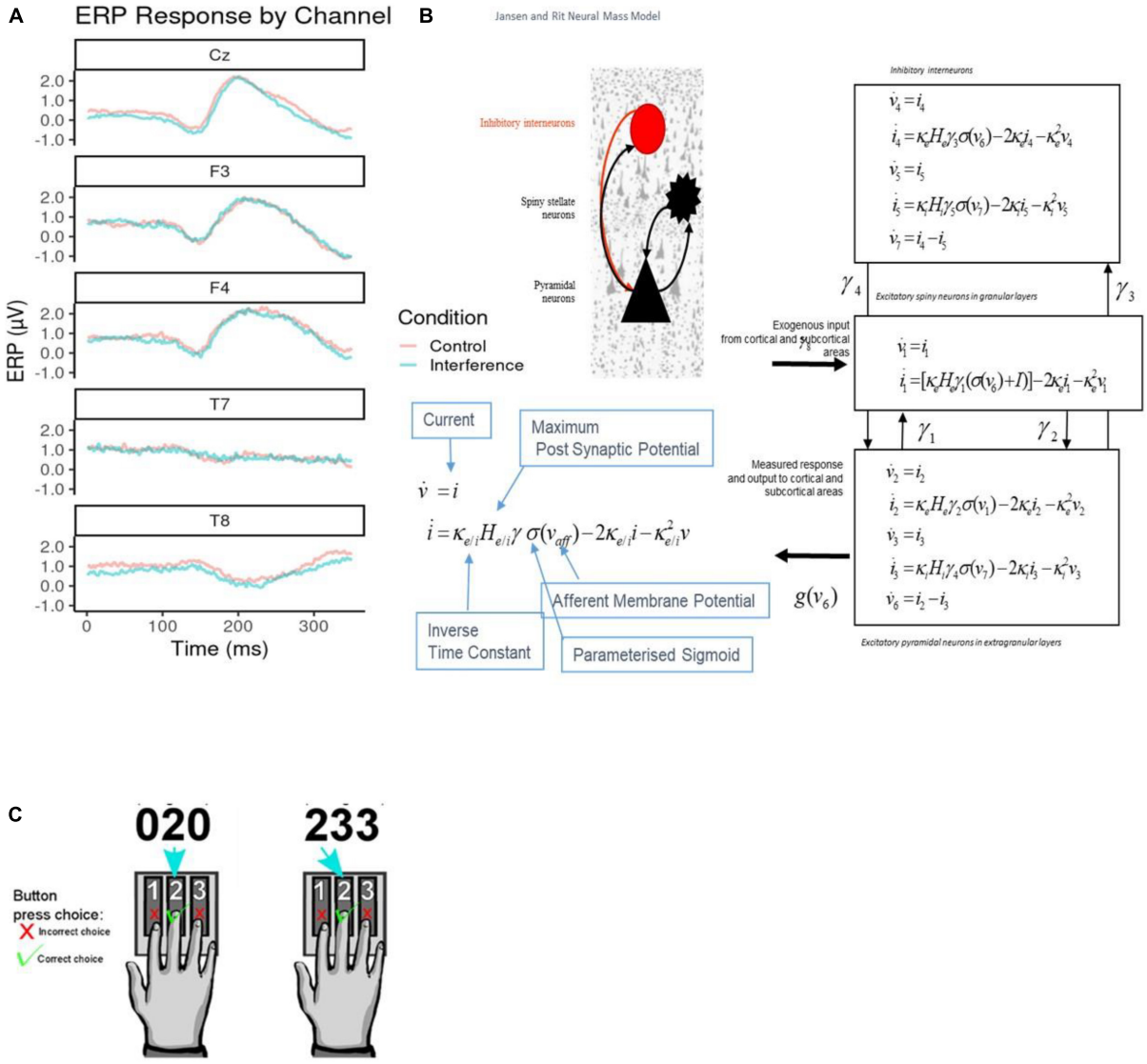


FIGURE 1
**(A)** Sample average electroencephalogram (EEG) responses. These are shown per channel retained for classification analysis. Both control (red) and interference condition (blue) averages are plotted across the 350°ms response time post stimulus onset. Channels Cz, F3, and F4 show typical event-related potentials (ERP) amplitudes, while T7 and T8 show flat EEG responses across the duration. **(B)** Jansen and Rit Model. The schematic diagram summarizes the evolution equations that specify a Jansen and Rit (JR) neural mass model of a single source. This model contains three populations, each loosely associated with a specific cortical sub-population or layer: pyramidal and spiny stellate neurons and inhibitory interneurons. Second-order differential equations mediate a linear convolution of presynaptic activity to produce postsynaptic depolarization. This depolarization gives rise to firing rates within each sub-population that provides inputs to other populations. The operations are captured by the equations shown on the right-hand side, which are explained in the bottom-left comer inset that includes the parameters appearing in these equations and then definitions. For a thorough discussion of these equations, see the study by (26). **(C)** Multiple Source Interference Task (MSIT). The task requires participants to report on a presented stimulus by using their index, middle or ring finger to press three buttons corresponding to numbers 1, 2, and 3 respectively. The stimulus appears on a screen displaying three numbers; one number (the target) is different from the other two (distractors). The participant identifies the target by pressing the corresponding buttons. There are two task conditions, control, and interference. During control trials, the distractor numbers are zeros and the location of the target number is aligned with its corresponding button. In interference trials, the distractors are non-zero numbers, and the target is in a location misaligned with that of the button.

measurements (2). We characterized changes of intrinsic (within area) and extrinsic (between area) connections across task conditions and between individuals. We assessed whether information flow changed in the same way (top-down, bottom-up or both) between the two task conditions across all participants. We here used DCM for Evoked Responses and Jansen and Rit (JR) mass model (Figure 1B). JR models can predict both evoked and induced responses and have been used

in theoretical and experimental studies (27, 41–44). DCM was implemented using SPM12. For more details about DCM, see **Supplementary methods**.

## Functional network

The functional network modeled with DCM can be seen in **Figure 2** (cf. model M1 in top left corner, all other models include the same network and assume changes in different connections, explained below). This network is comprised of areas activated during the MSIT (29, 45). The network included sensory, temporal, parietal, dorsal and ventral frontal areas, and ACC: V1, dACC and the following areas in both hemispheres: ITG, SPL, vlPFC, dlPFC. Changes in functional connectivity within this network were observed, at the group level, in patients with depression (46–49). For details about the coordinates of these areas, and why we chose this network and no other areas, see **Supplementary methods**. We used DCM parameter estimates as data features for classification and clustering.

## Dynamic causal modeling parameters

Dynamic causal modeling (DCM) parameter estimates were obtained by fitting ERPs, i.e., Positive Potentials evoked during the MSIT task. We fitted ERP recordings from different participants, patients, and controls. We thus obtained DCM parameter estimates. Noise or heterogeneity in the scalp-level recordings might arise from a small number of disruptions in the underlying network. After fitting, variability in ERP recordings leads to variability in the biophysical model (DCM) parameter estimates across participants. This, in turn, could describe biotypes or endophenotypes of depression. We hypothesize: (1) If DCM can capture that variability, then DCM-derived model parameters might be more effective than raw ERPs at classifying patients from controls; (2) Clustering of DCM parameters may help identify clusters of endophenotypes. We tested these hypotheses below.

Dynamic causal modeling (DCM) parameters were obtained after fitting data from individual subjects. These included the

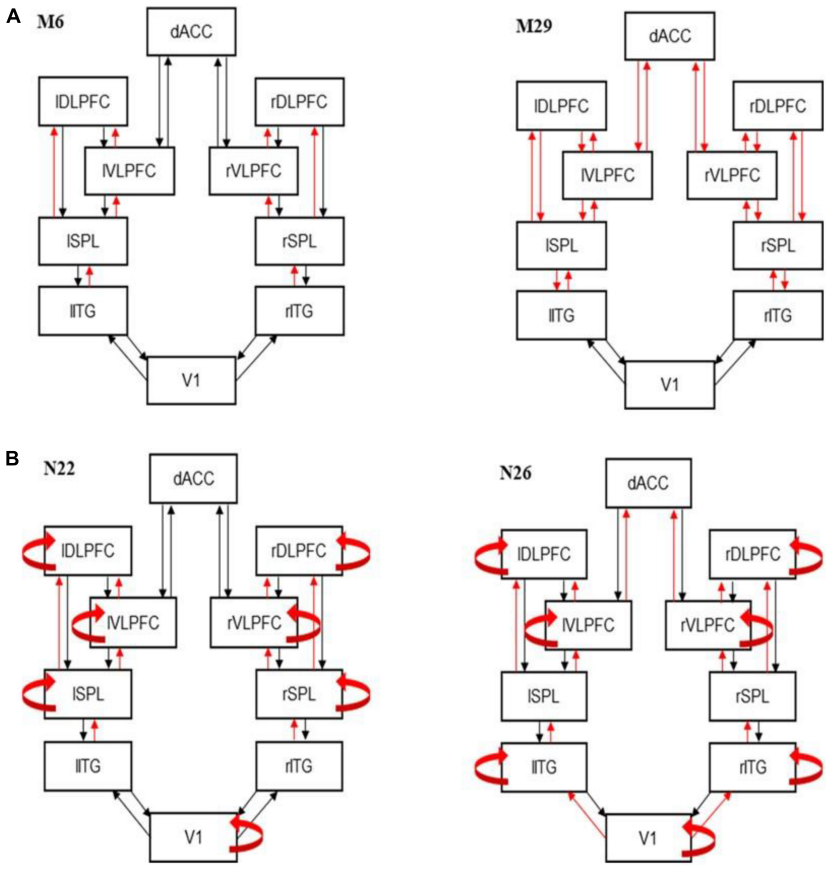


FIGURE 2
Best fitting models. **(A)** Dynamic causal modeling (DCM) best fitting model M6 (left) and runner up model M29 (right). Model M6 includes changes in forward connections at all levels except VI. The runner up (M29) is very similar. It also includes the corresponding feedback connections on top of the forward connections included in M6. **(B)** Best fitting DCM models showing modulations of intrinsic connections for controls (N22; left) and patients (N26; right).

following parameters: extrinsic connectivity, $A$ ($12 \times 2 = 24$ parameters), differences in extrinsic connectivity between MSIT conditions, $B$ (24 parameters, derived from the model fitting as shown in Results), excitatory and inhibitory receptor density, $G$ ($2 \times 10 = 20$ parameters), strength of connections between the three populations of the JR model shown in **Figure 1B**, $H$ (4 parameters; see arrows in **Figure 1B**), and excitatory and inhibitory synaptic time constants, $T$ (20 parameters).

## Model comparison

Because we did not know how connectivity changed between task conditions, we compared several variants of the biophysical model describing the network of **Figure 2**. We considered a network containing all of our modeled brain regions: V1, ITG, SPL, vlPFC, dlPFC, and dACC. We assumed forward and backward connections between specific areas, as well as lateral connections between homologous areas in the right and left hemispheres. We asked which connections might change between MSIT conditions and considered all possible changes. The alternative model variants differed in the connections that could change. Following (39), we first considered changes of extrinsic connections (i.e., between nodes) only. Then in step 2, changes in intrinsic connections. The first twenty candidate models with extrinsic connections changes that we considered, are shown in **Supplementary Figure 2**. Overall, the candidate model space comprised 45 models. We describe in detail these 45 models in **Supplementary methods**. The 45 models included all possible models where forward or backward connections changed between different parts of the brain network. Finding the most likely among these models yielded the extrinsic connections that were modulated during the task. For model comparison, we used an approach known as Bayesian model selection (BMS). This was performed assuming fixed-effects (FFX) (50). BMS fits competing models to EEG data and assesses the most likely model. See **Supplementary methods** for more details.

We considered variations of the network shown in **Figure 2**. We assumed changes in intrinsic connections from each node to itself (in addition to changes in extrinsic connections that the winning model above assumed) (39). We thus assumed that intrinsic connections could change at any (combination of) brain areas: V1, ITG, SPL, {vlPFC, dlPFC}, and dACC. We thus compared 32 candidate models in total.

## Classification features

The biophysical parameters of the best DCM model obtained *via* BMS were used as features for patient classification and subtyping.

Two sets of classification features were used. DCM parameters and EEG features. DCM parameters were directly compared to EEG features. The DCM parameters included intrinsic and extrinsic connections that were found to differ between MSIT conditions in both the patient and control DCM fits. This resulted in 92 DCM parameters. These were used as DCM predictors in machine learning classifiers.

To compare the predictive power of the DCM parameter estimates with raw EEG features, we used an equal number of ERP features (51). The full set of potential EEG features included 240 variables (60 EEG Channels x 2 conditions x 2 variables, i.e., ERP peak amplitude and latency differences between the two MSIT conditions). We reduced the number of channels to 23 so that the total number of EEG features was the same as the number of DCM parameters. This reduces bias in the comparison between ERP and DCM feature sets. To choose these 23 channels, we performed permutation testing that assesses the change in prediction error of classification after permuting a feature (52, 53). The ERP features were chosen based on their contribution to a random forest model (constructed without hyperparameter tuning). This "naive" random forest allowed us to select channels with features that were most beneficial in separating classes while still allowing for multiple interaction effects between features. The tradeoff of this method stems from using a reduced number of features with the benefit that they are potentially more meaningful, and easier to interpret.

Critically, the above selection of ERP features, biases the subsequent machine learning analysis against our *a priori* hypothesis that DCM-based features will provide at least equivalent classification and clustering–the DCM analysis considers an unselected set of model parameters, whereas the ERP analysis begins with features already known to have some classification power. The selected channels are included in **Supplementary Table 2**.

## Class balancing and algorithm training

The dataset was imbalanced between control and patient classes. Only 15 of the 49 participants being patients with depressive symptoms. We implemented Synthetic Minority Over-sampling Technique (SMOTE) to correct for this imbalance (54). For more details, about oversampling, see **Supplementary methods**. This brought parity to the classes with 34 observations each (patients and controls).

Given the sample size limitations, we used 10-fold cross-validation to train and tune machine learning classifiers (55, 56). See **Supplementary Figure 9** for a visual depiction of the sampling strategy and **Supplementary methods**. The cross-validation was used to train classifiers and assess whether DCM features can better measure depression's internal heterogeneity, compared to EEG features (56). We compared the performance

of different machine learning algorithms to distinguish patients with depression vulnerability from controls in both the EEG and DCM feature sets. The algorithms included Support Vector Machines (SVM) (57), Random Forests (52, 58), and Gradient Boosted Trees (59). Comparing DCM and EEG features using multiple algorithms ensures that conclusions are not sensitive to the specific algorithm used. We also used multiple performance metrics, including *F*-measure (F1-score), and the Matthews' Correlation Coefficient (MCC) (60). See **Supplementary methods** for more details.

## Feature importance

The best performing classifier as determined by mean MCC score was used to compute feature importance. Shapley additive explanation (SHAP) values were constructed for predictions on the original data set (49 participants, no SMOTE augmentation) (61). This reveals how efficient the low dimensional space spanned by DCM and EEG classification features is in describing the internal heterogeneity of patients with depressive symptoms. SHAP values were constructed using subsampling of different combinations of input features and attributing a weight representing how much credit features should receive for class prediction. The predictive power of EEG and DCM features was compared directly because the corresponding SHAP values take on the same scale and are predicting the same underlying data.

## Unsupervised clustering

The ten most important features as determined by SHAP values from both the ERP and DCM feature sets were used to construct embedding scores with t-stochastic neighbor embeddings (*t*-SNEs). *t*-SNEs are useful for exploring higher-dimensional data in lower dimensional representations when non-linear relationships exist in the data (62). For more details on t-SNE see **Supplementary methods**. This provided visualizations of the data that were convenient for assessing subtypes or clusters of patients with depressive symptoms.

Clustering was performed using *k*-means in the three dimensional space obtained by *t*-SNE. *K*-means is an unsupervised machine learning method that groups observations to reduce within-cluster sum squares distances and increase the sum squared distance between cluster centroids (63, 64). *K*-means depends on an *a priori* number of clusters. The optimal cluster number can be found by computing Silhouette scores across candidate values of *k*. Observations which been classified appropriately have a lower mean distance between points within their assigned cluster compared to the mean distance to points in the next-nearest cluster neighbors (65). This ratio is given by Silhouette scores.

# Results

## Dynamic causal modeling

We first asked how information flow changes between the congruent and incongruent condition of the MSIT. We used Bayesian Model Selection (BMS; see **Supplementary methods**) to find the connectivity pattern between our six modeled areas. We fitted ERP data using a biophysical DCM model (**Figure 1B**) and scored all possible model variants that corresponded to different subsets of connections that might change between MSIT conditions. Different competing models represented different combinations of modulated forward or backward extrinsic connections (see also Methods). BMS identified the winning model as M6 ($BF > 3$; **Figure 2A**, see also **Supplementary Figure 4**). Model M6 includes changes in forward connections between ITG→SPL, SPL→vlPFC, SPL→dlPFC, and vlPFC→dLFPC. The runner up (M29) is very similar to M6 (**Figure 2B**). It includes the corresponding feedback connections on top of the forward connections included in M6. M29 also includes changes in feedforward and feedback connections between vlPFC, dLFPC, and dACC. Thus, the prominent difference between task conditions was changes in the forward (bottom-up) information flow between sensory processing and associative regions. This corresponds to different processing of sensory input between the two MSIT conditions that gives rise to ERP differences in the first 350°ms after stimulus presentation.

In the second step, BMS was used to test the modulation of the intrinsic connections. In this step, we fixed the extrinsic connections to those shown in the winning model M6 above. We then considered variants of M6, where intrinsic connections (within each brain area) were allowed to change. These variants formed a different model space to the one considered above (Methods and **Supplementary Figure 3**). BMS identified that variant N22 had the highest evidence ($BF > 3$; **Figure 2B**, see also **Supplementary Figure 5**). This included modulated intrinsic connections in V1, SPL, vlPFC, and dlPFC areas. The runner up (N23) is exactly the same as N22 with the additional modulation of dACC intrinsic connections (**Supplementary Figure 5**). We also performed the same BMS for the patient cohort.

We repeated the earlier analysis, where we fitted the neural mass model to patient data only. We wanted to capture the intrinsic heterogeneity. We first found that model M8 best described the changes in extrinsic connections (**Supplementary Figure 6**). This included changes in all feedforward connections. The runner up (M41) was similar to the winning model, M8, in that it included changes in the feedback connections too. The difference from M8, was that it did not include changes in feedforward (and feedback) connections between V1 and SPL and vlPFC and dACC. The winning model for patients, M8, was similar to the winning model for the control cohort (M6).

The difference between the winning models was that two more brain regions showed modulations of feedforward connections. Patients showed changes in vlPFC to dACC and V1 to ITG, in addition to changes in forward (bottom-up) information flow between the ITG to SPL, SPL to vlPFC and dlPFC, and vlPFC to dlPFC that we had found for controls. However, we did not use this difference for our clustering analyses below. Thus, we do not claim that the additional connection changes that we found for the patient cohort are biomarkers of depression or depressive vulnerability. Rather, the similarities (common connections) between M6 (controls) and M8 (patients) were the inputs to machine learning algorithms together with the biophysical parameters described in the last paragraph of this section below.

In the second step, similar to the analysis above, we identified N26 as the model with highest evidence ($BF > 3$; **Figure 2B**, see also **Supplementary Figure 7**). This is very similar to N22 that was the winning model for controls. The only difference between N22 (controls) and N26 (patients) is that N22 includes changes in intrinsic connections in ITG instead of SPL. The runner up (N9) is also very similar and assumes modulations of intrinsic connections in dACC instead of V1. Again, we are not claiming this change to be a biomarker of depression, but as an example of how DCM can identify underlying neurological variability. The clustering/classification analysis used only the intrinsic connections that were common to N26 and N22.

Several connections were independently shown to be modulated between conditions in both cohorts (controls and patients): ITG to SPL, SPL to vlPFC, and dlPFC, vlPFC to dlPFC, and the intrinsic connections in V1, vlPFC, and dlPFC. After finding the set of extrinsic and intrinsic connections that changed between MSIT conditions, we fitted the winning model (N22 for controls; N26 for patients) to each participant. Example model predictions for each of the three populations are shown in **Supplementary Figure 8**. Blue and red lines correspond to the two MSIT conditions (control and interference). There are three pairs of lines, corresponding to the three populations of the JR model (cf. **Figure 1**, right panel). After fitting the model, we obtained connections (*A, B*) and other biophysical parameter estimates from each participant (*G, H, T*; see Methods). These were used as DCM predictors in the next section.

## Classification

Our goal was to assess whether DCM features can better measure depression's internal heterogeneity, compared to EEG features. To do this, we asked whether DCM features achieved significantly better classification than EEG features (56). EEG features included ERP parameters, i.e., ERP peak amplitude and latency differences between the two MSIT conditions. We used ERPs from 24 channels so that the number of EEG predictors was equal to the number of DCM predictors. This allowed us to perform head-to-head comparisons between ERP and DCM predictors. Crucially, we chose ERP predictors in such a way that it biases subsequent analysis against DCM predictors (they were the best classifiers in an initial random forest model; see Methods for more details). Due to the small sample (15 patients), it was not possible to test out of sample predictions for algorithm robustness. To assess the relative ability of DCM and ERP parameters to capture diagnostic heterogeneity (and thus support better classification), we computed classification performance using three algorithms: SVM, Gradient Boosted Tree and Random Forest (**Figure 3**). The SVM algorithm performed best for both the DCM and ERP data sets (highest average MCC scores).

Overall, DCM features led to better classification accuracy than EEG features across all three algorithms tested. SVM was used for assessing feature importance because it had the highest classification performance (see **Supplementary Table 3** for full results of MCC and F-Score). It also had a more parsimonious hyperparameter set (two–kernel gamma and cost) compared to either decision-tree ensemble model.

To evaluate feature importance, we used absolute SHAP values averaged across participants. **Figure 4** shows SHAP values for the 10 most important DCM (**Figure 4A**) and EEG (**Figure 4B**) features. By taking the average across each of the 49 classification decisions (participants), we have a relative rank of feature importance. Overall, the DCM features have higher mean absolute SHAP values compared to EEG features.

Individual DCM features have higher SHAP values that the EEG features with the same corresponding rank. For example, the most important DCM predictor is the rSPL inhibitory time constant (mean SHAP = 0.22). The most important ERP predictor is AF7 Latency score in the interference condition (mean SHAP = 0.17). These results are directly comparable, with the larger SHAP score reflecting greater feature importance.

We also compared the mean SHAP scores between the two datasets, with DCM scoring higher than ERP across all ten most important features (the SHAP value of the top DCM feature is larger than the corresponding SHAP value for the top ERP feature; the SHAP value of the second-best DCM feature is larger than the corresponding SHAP value for the second-best ERP feature, etc.). We compared the mean absolute SHAP values in a paired $t$-test using rank as the pairwise grouping. The 10 top DCM features had higher SHAP values than the 10 top EEG features [$t = 4.28$, $p = 0.002$, CI = (0.01, 0.03)].

Overall, DCM features appeared equivalent or more powerful than EEG features (higher SHAP values). They also captured separate subtypes of depressed participants better. This may relate to DCM features' ability to capture variability between individuals. **Figure 5** shows the distributions of the ten most important DCM and EEG features. Visually, DCM features show distributions with central tendencies, with areas of non-overlap between patient and control distributions. This can be explained by the Laplace approximation used to define the

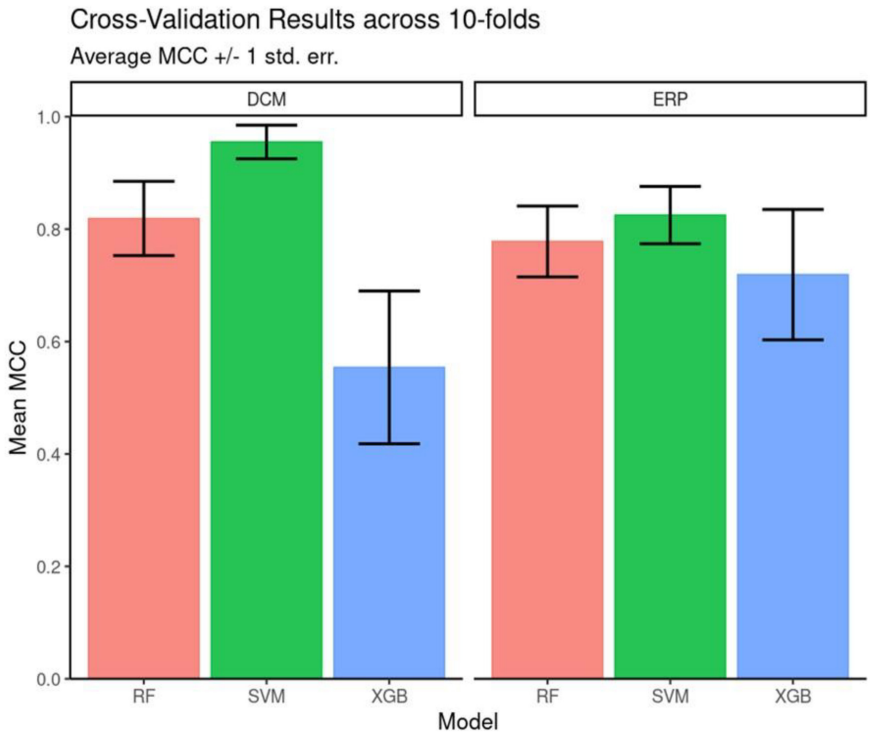


FIGURE 3
Cross validation performance. Matthew's Correlation Coefficient scores across the 10-fold cross-validation. Mean and +/− 1 standard error are reported for each model architecture. Support vector machines (SVM) classifiers performed the best for both feature sets on average across the 10 folds.

posterior densities of DCM parameters (66). They occupy less of the available numeric range. ERP features are more uniformly dispersed over the available range.

## Unsupervised clustering

To capture depressive heterogeneity, t-SNE representations were made from unlabeled data of each of the 10 most important features for DCM and ERP feature sets. *t*-SNEs were generated using a perplexity value of 25 over 2,500 iterations. These embeddings were labeled *post hoc* to determine if these features could elucidate differences between patients and controls. **Figure 6** shows the three dimensional representations of patient embeddings in blue and control embeddings in red. **Supplementary Figure 10** shows the t-SNE results for two dimensional t-SNE representations. DCM feature embeddings show clustering tendencies, while EEG feature embeddings were more dispersed throughout the lower dimensional spaces with no clear patterns.

Our patient dataset included two depression subtypes, bipolar, and unspecified depression. We asked whether we could recover this using the t-SNE spaces obtained using DCM and EEG biomarkers. We clustered the t-SNE embeddings using k-means and compared Silhouette scores for all participants and for the patient cohort only. **Figure 7** shows the mean Silhouette score with $\pm$ 1 standard error. Scores for both DCM and EEG results across 2–12 clusters ($k$) can be found in **Supplementary Table 4**. Silhouette scores decrease monotonically, suggesting that a two cluster representation is the most parsimonious solution in this small dataset. The lack of change in Silhouette scores over increasing values of $k$ for the all-participants dataset suggests that there is no clear clustering solution that separates, e.g., controls and two or more patient types. This may reflect unstructured heterogeneity in the control participants.

The DCM embeddings for the patients only at low $k$ values had the highest mean Silhouette scores [$\mu$ = 0.388, CI = (0.334, 0.441)]. Compared to the second highest Silhouette score, ERP embeddings for patients only, the DCM embeddings were significantly higher when compared using a two-sided $t$-test ($t$ = 2.17, $p$ = 0.032). Interestingly, DCM Silhouette scores decreased in a monotonic fashion (blue solid line in **Figure 7**), thus confirming the two known clusters in the patient dataset.

# Discussion

We demonstrated a proof of concept that transforming electrophysiological data to underlying biophysical parameters

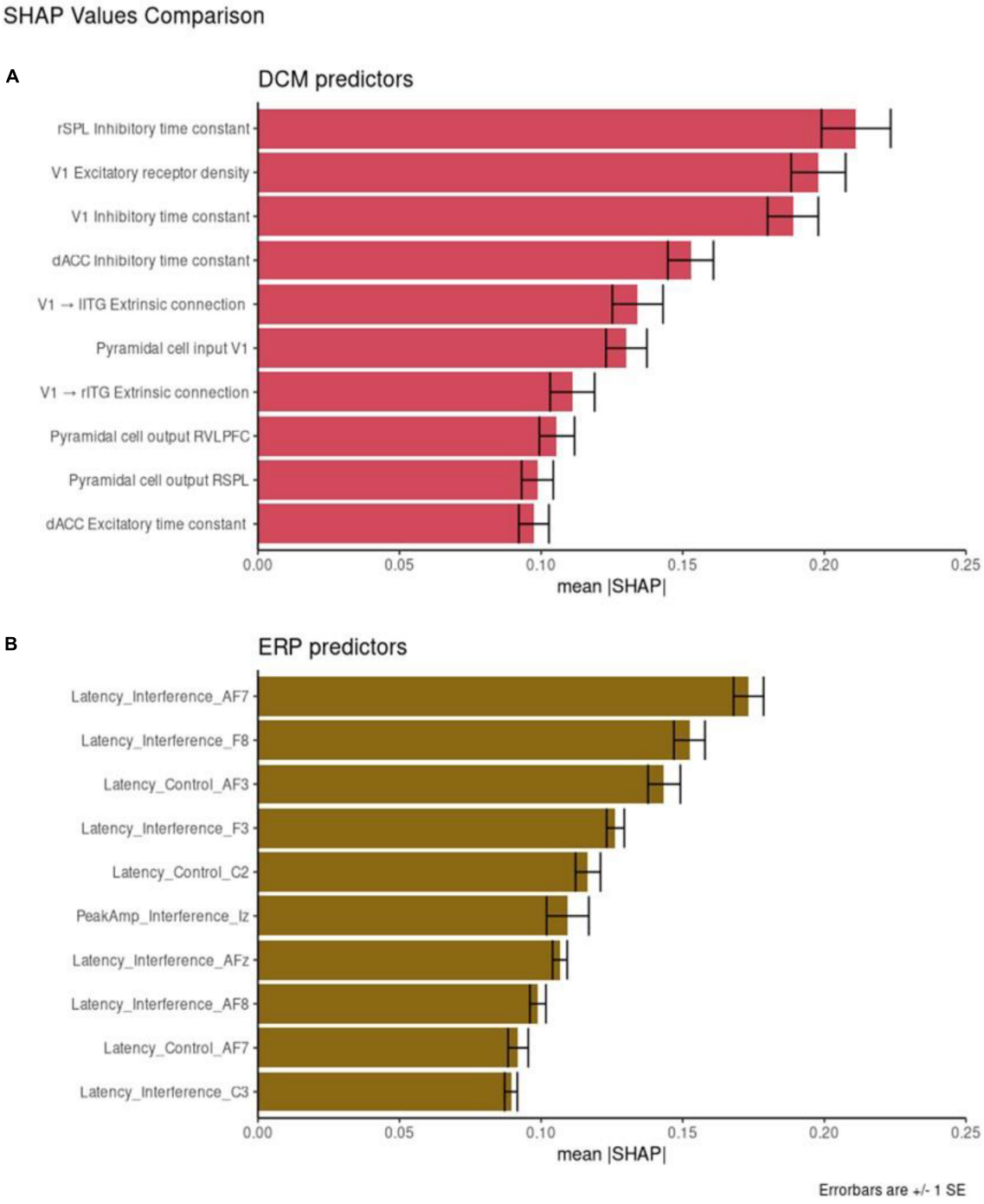


FIGURE 4
Feature importance. Shapley additive explanation (SHAP) values for the 10 best performing dynamic causal modeling (DCM) features **(A)** and event-related potentials (ERP) features **(B)**. Mean absolute SHAP values are reported with +/− 1 SE.

using DCM may reliably capture variability that correlates with clinical status. Although our dataset is small and heterogeneous, DCM features were equivalent to or outperformed raw EEG features on most classification metrics, and this held true across multiple classifier algorithms. Our results align with prior work that has successfully used DCM to identify differences between unipolar and bipolar depression (22). We considered participant-specific (first-level) analysis and group effects (second-level analysis) as two separate steps. An alternative approach could be to combine these steps into

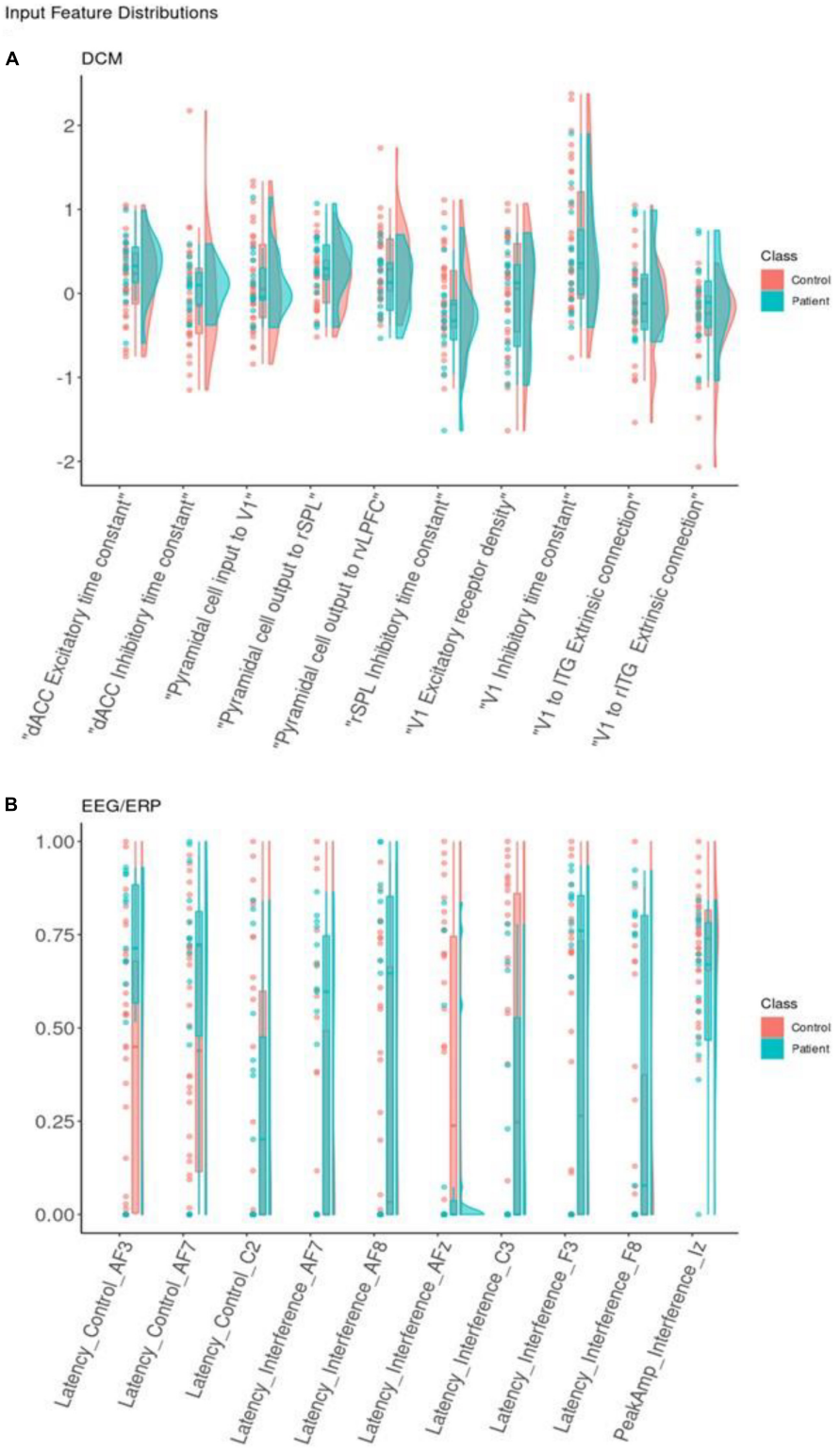


FIGURE 5
Feature distribution. Distributions for the 10 best dynamic causal modeling (DCM) input features **(A)** and (electroencephalogram/event-related potentials) (EEG/ERP) input features **(B)**. Classification features represented using dots for real observations alongside boxplots and density curves to show the shape of the distribution. Each is colored by class. Each input feature on the x-axis is shown with raw data points, boxplots, and density curves. Patients (blue) and controls (red) are shown separately. The EEG/ERP features are shown after min-max scaling to present latency and interference variables together on a comparable scale.

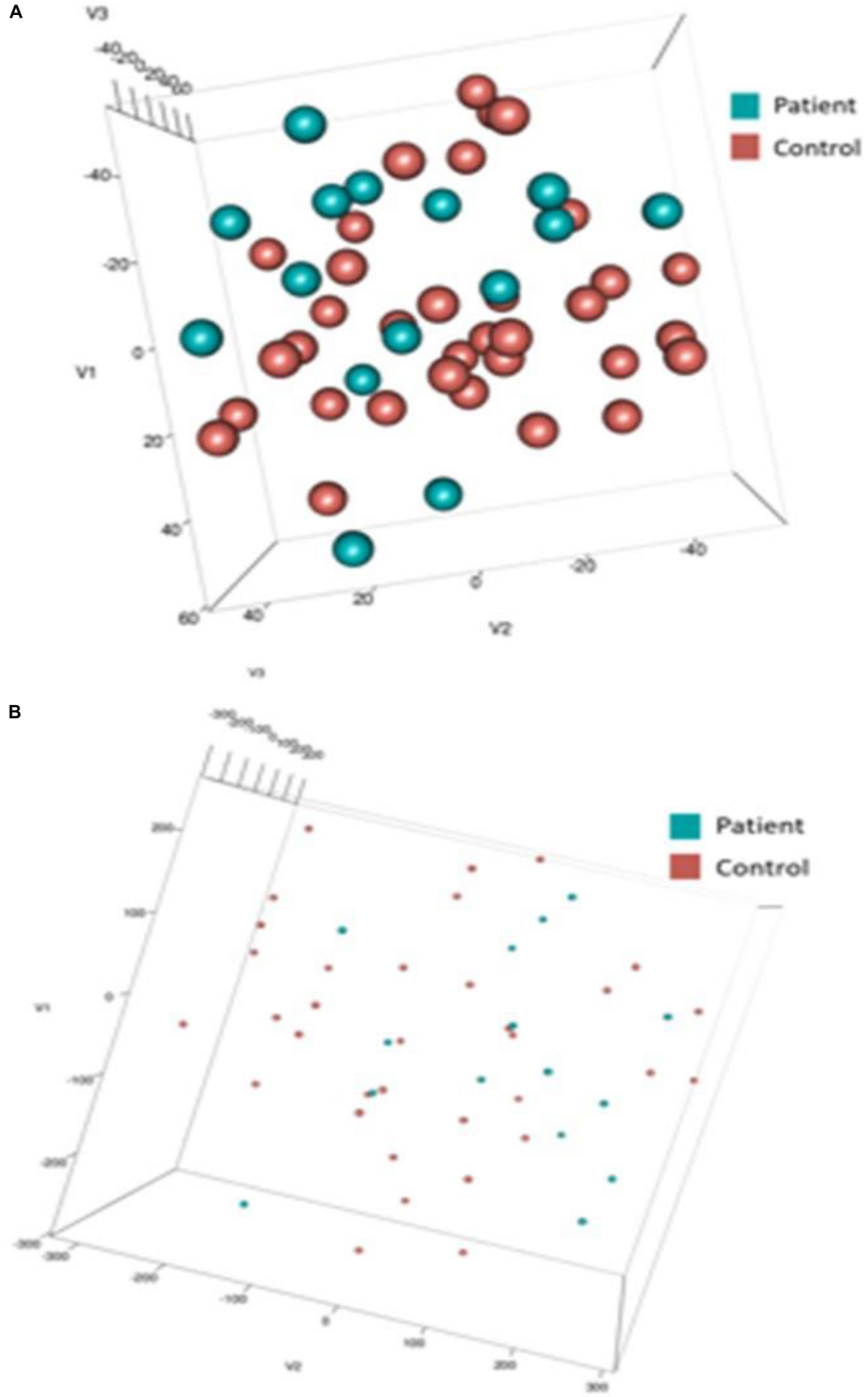


FIGURE 6
t-SNE embeddings. Three dimensional t-SNE values for top 10 input features of dynamic causal modeling (DCM) **(A)** and electroencephalogram/event-related potentials (EEG/ERP) **(B)**. *Post-hoc*, data points were colored blue (Patient) and red (Control). Note the difference in point size reflects the difference in axis scale (−/+ 40 in DCM compared with −/+ 300 in EEG).

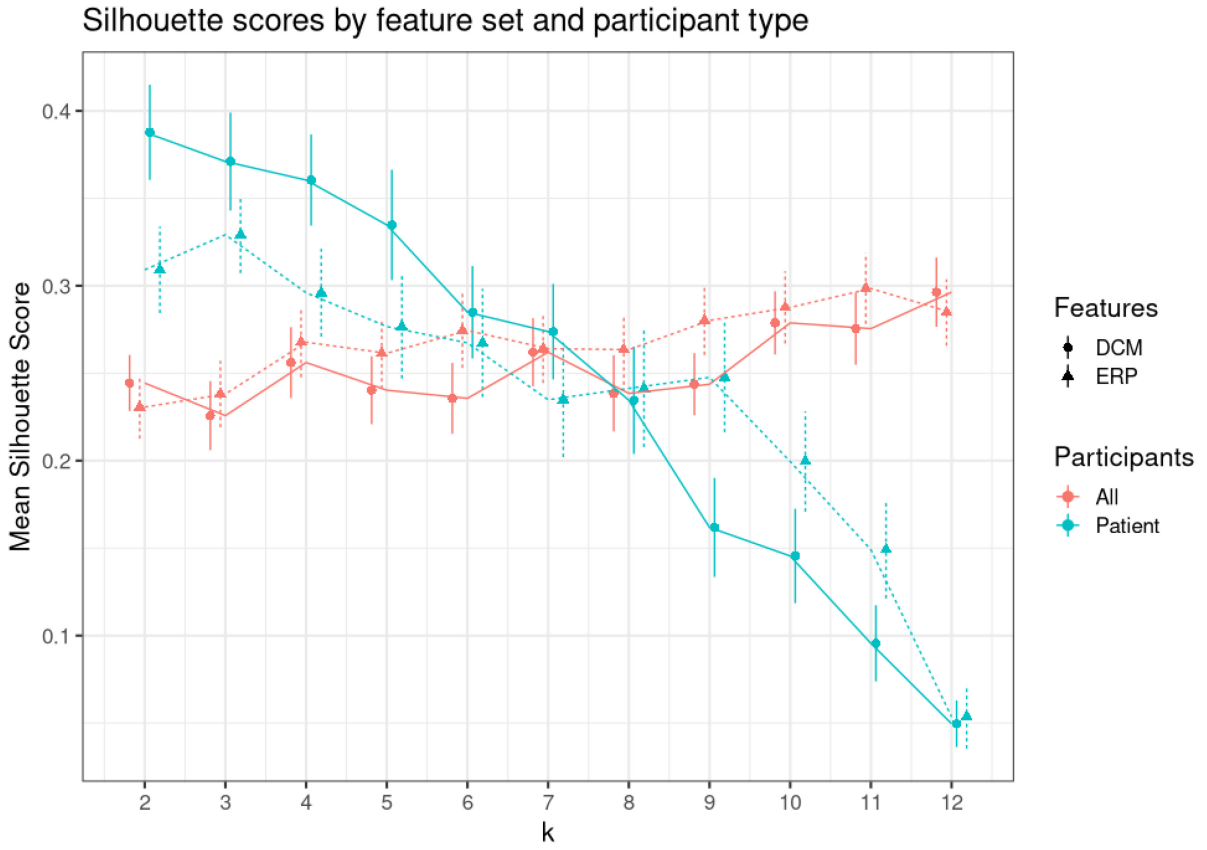


FIGURE 7
*K*-means clustering. Clustering was performed using varying numbers of *k* (from 2 to 12) to cluster (1) all participants, (2) only patients. To find the optimal *k*, we used Silhouette scores. Silhouette scores quantify the ratio of within over between cluster distance. A positive value signifies appropriate clustering (74). This figure shows the mean Silhouette scores (vertical axis) for each level of *k* (shown on the horizontal axis). Red dots show scores for patients and controls (all), while green dots show scores for patients only. Solid lines and discs depict scores for dynamic causal modeling (DCM) features, while dashed lines and triangles for electroencephalogram (EEG) features. Error bars show +/− 1 SE.

a single hierarchical model (35, 66). We will consider this elsewhere. Similarly, we used a Fixed Effects approach that could be replaced by Mixed-Effects (7, 8).

If our DCM-based approach can be replicated on a larger dataset, it may provide an intriguing avenue for personalized medicine. There is a growing set of TMS and similar tools for manipulating brain connectivity (67, 68). For any individual patient, the approach we describe permits the identification of which DCM feature(s) are driving that patient's vulnerability. Although this is still a speculative claim, it may be possible to then target and normalize those specific features. The feasibility of that approach will depend on whether these features change as a patient undergoes treatment or remain present even during euthymia (69, 70).

We found that MSIT-induced variance was explained by feedforward connectivity from primary sensory and object discrimination areas to prefrontal cortex, plus changes in intrinsic, within-region connectivity. This is consistent with current working models of cognitive control, the construct tested by MSIT. In those models, information about cognitive control demands is computed posteriorly, then fed forward to anterior structures (dlPFC which then influences motor circuits) (71, 72). Given that we explored a wide range of connectivity changes, our data-driven recovery of a known phenomenon provides some confidence that we identified known effects. At the same time, our analyses were based on prior assumptions about anatomically plausible connections. In future work, we will test these assumptions using BMS.

Dynamic causal modeling (DCM) suggested a low dimensional space of potential biomarkers (we went from 240 EEG-based to 92 biophysical features). This helps fitting prediction models on limited datasets. As an alternative to PCA (69, 73), DCM-derived parameters are also directly interpretable; e.g., DCM synaptic connectivity or intrinsic excitability can directly map to potential treatments (20). Another DCM advantage is that its biophysical models provide ways to test potential explanations of pathophysiology. This advantage can also be a limitation: for example, some prior

knowledge about cortical regions and their interactions is required. Finally, we note that DCM is not simply a biophysical model. It also involves source reconstruction (i.e., analyses in source space) while the ERP analyses are in sensor space. This source attribution adds further explainability, and may explain why the DCM parameters had higher explanatory power on this specific dataset.

We here used a small patient dataset, to present a proof of concept that biophysical biomarkers are a feasible approach to depression subtyping. That dataset is inherently limited, and our current results should not be treated as generalizable. In future work, we will analyze larger patient datasets, which are increasingly available for secondary analysis (74). We will address data leakage by using a hold-out test set of data observations to evaluate classification model performance. This will allow for better comparisons between model architectures and depression subtyping using self-supervised learning algorithms like TABnet (75).

We emphasize that the classification performance of our models should not be treated as a claim that the current pipeline carries diagnostic or clinical utility. The sample size is too small; we and others have pointed out that small-sample-derived biomarkers frequently do not generalize (14, 76, 77). Although we did perform some internal cross-validation, this sample size did not permit us to follow best-practices for preventing data leakage (78). Specifically, we performed SMOTE up-sampling on the full dataset, before conducting a training/validation split. On the other hand, the goal of this work was not to develop a reliable classifier and report its performance. The classification approach was used solely to compare the relative performance of DCM vs. ERP features (similar susceptibility to data leakage and effect size inflation). Regarding the ERP features themselves, we only found statistically significant differences in ERP latencies, not peak amplitudes. We could have used average amplitudes within an epoch, but this would not change our results, as the average and peak amplitude were highly correlated ($R$ = 0.98 for the control and $R$ = 0.92 for the interference condition). Related to this, a significant difference between DCM and the standard ERP technique is that the ERP focuses on two specific aspects of the polyphasic ERP response (peak amplitude and peak latency). By virtue of its model fitting, DCM considers the full shape/temporal evolution of the ERP. One approach to addressing this might be to perform a principal component analysis (PCA) or similar decomposition on the ERPs. One could then compare, e.g., eigenvalues of the first few PCs against the DCM derived features. In this proof of concept study, we focused on alignment with earlier work (5–8). We used the most common ERP markers, i.e., peak amplitude and latency. A more detailed exploration of potential ERP features will be considered elsewhere.

Our feature importance scoring (SHAP) results should also be interpreted with caution. With this caveat, over half the highly weighted features from the DCM analysis came from more caudal regions and included within region features. This aligns with recent results in a larger dataset, where signals in primary sensory regions were more able to classify (non)response to antidepressant treatment than were signals from higher-order cognitive/associative regions (79). On the other hand, those response predictors were generally unstable in a cross-validation analysis (80). In all, we do not make claims that our specific identified markers/clusters are new generalizable findings. Rather, they are proof of concept for a larger point: that DCM parameters can be scored and interpreted, and that subtypes might be identifiable by clustering. With a larger dataset, it would become feasible to identify robust DCM features (78). There are such recent datasets available (81, 82), particularly in depression (51, 74). In future work we will consider these new datasets together with additional covariates, like stressful life events (83), biological sex (6), and others. These can easily be combined with the DCM pipeline we have shown here.

Another caveat is that t-SNE does not preserve distances and global structure in the data. It only preserves local geometry while constructing the low dimensional embedding (84). Thus one might wonder whether it is an appropriate tool for discovering clusters in biomarker data. Indeed, we did not find unknown clusters that might reflect phenotypic or other differences. Our only point was that using DCM—not ERP—biomarkers projected onto t-SNE spaces allowed us to recover the known patient clusters in our data, bipolar and unspecified depression. This is similar to the use of the t-SNE algorithm in bioinformatics (85, 86), where t-SNE has been used to obtain hierarchical clusters and recover clusters obtained with different methods, like genetic analyses. t-SNE extensions are also useful for discovering unknown structure. There are also recent t-SNE algorithms that can address the above geometry distortion, like parametric (87) and hierarchical SNE (88). We will consider them in future work.

In summary, we have demonstrated the first proof of concept for a novel approach to identifying psychiatric biomarkers from EEG, based on converting manifest EEG signals to interpretable biophysical parameters. We demonstrated the viability of this approach against the same biomarker pipeline applied to manifest data (in this case, ERPs). If applied to larger datasets and a more robust variety of data sources, this DCM-based pipeline can be an important new approach to dissecting the heterogeneity of depression and depressive vulnerability.

## Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: https://drive.google.com/drive/folders/1X5_NJJDqyV11va5asOdLHMPski-zrJ1a?usp=sharing

and https://github.com/pinotsislab/ERP-Feature-Extraction-and-Classification/tree/main.

## Ethics statement

The studies involving human participants were reviewed and approved by Massachusetts General Hospital Institutional Review Board. The patients/participants provided their written informed consent to participate in this study.

## Author contributions

DAP and ASW: conceptualization, methodology, funding acquisition, and resources. DAP and SF: writing—original draft. All authors validation, formal analysis, investigation, data curation, writing—review and editing, and visualization.

## Funding

DAP acknowledges financial support from UKRI ES/T01279X/1. ASW acknowledges financial support from the National Institutes of Health (R21MH120785, R01MH123634, and R01EB026938), the MnDRIVE Brain Conditions initiative, and the Minnesota Medical Discovery Team on Addictions.

## Conflict of interest

Author ASW has multiple pending and granted patents in the area of biomarkers of psychiatric illness. None is licensed to any commercial entity. Author ASW receives consulting income from a company (Dandelion Science) engaged in EEG biomarker discovery.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyt.2022.938694/full#supplementary-material

## References

1. McManus, S, Bebbington P, Jenkins R, Brugha T. *Mental Health and Wellbeing in England: Adult Psychiatric Morbidity Survey 2014*. London: NHS (2016).

2. Baker C. *Mental Health Statistics for England: Prevalence, Services and Funding*. (2018). Available online at: https://researchbriefings.parliament.uk/ResearchBriefing/Summary/SN06988. (accessed February 12, 2020).

3. Widge A, Malone D, Dougherty D. Closing the loop on deep brain stimulation for treatment-resistant depression. *Front Neurosci.* (2018) 12:175. doi: 10.3389/fnins.2018.00175

4. Cuthbert B, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med.* (2013) 11:126. doi: 10.1186/1741-7015-11-126

5. Simpraga S, Alvarez-Jimenez R, Mansvelder H, Van Gerven J, Groeneveld G, Poil S, et al. EEG machine learning for accurate detection of cholinergic intervention and Alzheimer's disease. *Sci Rep.* (2017) 7:5775. doi: 10.1038/s41598-017-06165-4

6. van Dinteren R, Arns M, Kenemans L, Jongsma MLA, Kessels RPC, Fitzgerald P, et al. Utility of event-related potentials in predicting antidepressant treatment response: an iSPOT-D report. *Eur Neuropschopharmacol.* (2015) 25:1981–90. doi: 10.1016/j.euroneuro.2015.07.022

7. Khodayari-Rostamabad A, Reilly JP, Hasey GM, de Bruin H, MacCrimmon DJ. A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. *Clin Neurophysiol.* (2013) 1240:1975–85. doi: 10.1016/j.clinph.2013.04.010

8. Widge AS, Bilge TM, Montana R, Chang W, Deckersbach T, Carpenter LL, et al. Electroencephalographic biomarkers for treatment response prediction in major depressive illness: a meta-analysis. *Am J Psychiatry.* (2019) 176:44–56. doi: 10.1176/appi.ajp.2018.17121358

9. Karaaslan F, Gonul AS, Oguz A, Erdinc E, Esel E. P300 changes in major depressive disorders with and without psychotic features. *J Affect Disord.* (2003) 73:283–7. doi: 10.1016/S0165-0327(01)00477-3

10. Kawasaki T, Tanaka S, Wang J, Hokama H, Hiramatsu K. Abnormalities of P300 cortical current density in unmedicated depressed patients revealed by LORETA analysis of event-related potentials. *Psychiatry Clin Neurosci.* (2004) 58:68–75. doi: 10.1111/j.1440-1819.2004.01195.x

11. Gallinat J, Bottlender R, Juckel G, Munke-Puchner A, Stotz G, Kuss H. The loudness dependency of the auditory evoked N1/P2-component as a predictor of the acute SSRI response in depression. *Psychopharmacol.* (2000) 148:404. doi: 10.1007/s002130050070

12. Jaworska N, Blondeau C, Tessier P, Norris S, Fusee W, Blier P. Response prediction to antidepressants using scalp and source-localized loudness dependence of auditory evoked potential (LDAEP) slopes. *Prog Neuropsychopharmacol Biol Psychiatry.* (2013) 44:100. doi: 10.1016/j.pnpbp.2013.01.012

13. Juckel G, Pogarell O, Augustin H, Mulert C, Siecheneder F, Frodl T. Differential prediction of first clinical response to serotonergic and noradrenergic antidepressants using the loudness dependence of auditory evoked potentials in patients with major depressive disorder. *J Clin Psychiatry.* (2007) 68:1206. doi: 10.4088/JCP.v68n0806

14. Widge A, Bilge M, Montana R, Chang W, Rodriguez C, Deckersbach T, et al. Electroencephalographic biomarkers for treatment response prediction in major depressive illness: a meta-analysis. *Am J Psychiatry.* (2019) 176:44.

15. Roelofs CL, Krepel N, Corlier J, Carpenter LL, Fitzgerald PB, Daskalakis ZJ, et al. Individual alpha frequency proximity associated with repetitive transcranial

magnetic stimulation outcome: an independent replication study from the ICON-DB consortium. *Clin Neurophysiol.* (2021) 132:643–9. doi: 10.1016/j.clinph.2020.10.017

16. Ip C-T, Olbrich S, Ganz M, Ozenne B, öhler-Forsberg KK, Dam VH, et al. Pretreatment qEEG biomarkers for predicting pharmacological treatment outcome in major depressive disorder: independent validation from the neuropharm study. *Eur Neuropsychopharmacol.* (2021) 49:101–12. doi: 10.1016/j.euroneuro.2021.03.024

17. Pizzagalli D, Webb C, Dillon D, Tenke C, Kayser J, Goer F, et al. Pretreatment rostral anterior cingulate cortex theta activity in relation to symptom improvement in depression: a randomized clinical trial. *JAMA Psychiatry.* (2018) 75:547–54. doi: 10.1001/jamapsychiatry.2018.0252

18. Brunel N, Wang XJ. What determines the frequency of fast network oscillations with irregular neural discharges? I. Synaptic dynamics and excitation-inhibition balance. *J Neurophysiol.* (2003) 90:415. doi: 10.1152/jn.01095.2002

19. Deco G, Jirsa VK, Robinson PA, Breakspear M, Friston K. The dynamic brain: from spiking neurons to neural masses and cortical fields. *PLoS Comput Biol.* (2008) 4:e1000092. doi: 10.1371/journal.pcbi.1000092

20. Frässle S, Yao Y, Schöbi D, Aponte EA, Heinzle J, Stephan KE. Generative models for clinical applications in computational psychiatry. *Wiley Interdiscip Rev Cogn Sci.* (2018) 9:e1460. doi: 10.1002/wcs.1460

21. Graña M, Ozaeta L, Chyzhyk D. Dynamic causal modeling and machine learning for effective connectivity in auditory hallucination. *Neurocomputing.* (2019) 326-327:61–8. doi: 10.1016/j.neucom.2016.08.157

22. Almeida JR, Versace A, Mechelli A, Hassel S, Quevedo K, Kupfer DJ, et al. Abnormal amygdala-prefrontal effective connectivity to happy faces differentiates bipolar from major depression. *Biol Psychiatry.* (2009) 66:451–9. doi: 10.1016/j.biopsych.2009.03.024

23. Pinotsis D, Perry G, Litvak V, Singh KD, Friston KJ. Intersubject variability and induced gamma in the visual cortex: DCM with empirical B ayes and neural fields. *Hum Brain Mapp.* (2016) 372:4597–614. doi: 10.1002/hbm.23331

24. Broderson KH, Schofield TM, Leff AP, Ong CS, Lomakina EI, Buhmann JM, et al. Generative embedding for model-based classification of fMRI data. *PLoS Comput Biol.* (2011) 7:e1002079. doi: 10.1371/journal.pcbi.1002079

25. Díez Á, Ranlund S, Pinotsis D, Calafato S, Shaikh M, Hall M-H, et al. Abnormal frontoparietal synaptic gain mediating the P300 in patients with psychotic disorder and their unaffected relatives: frontoparietal synaptic gain and P300 in psychosis. *Hum Brain Mapp.* (2017) 38:3262–76. doi: 10.1002/hbm.23588

26. Pinotsis DA, Moran RJ, Friston KJ. Dynamic causal modeling with neural fields. *Neuroimage.* (2012) 59:1261–74. doi: 10.1016/j.neuroimage.2011.08.020

27. Moran RJ, Pinotsis DA, Friston KJ. Neural masses and fields in dynamic causal modeling. *Front Comput Neurosci.* (2013) 7:57. doi: 10.3389/fncom.2013.00057

28. Widge A, Ellard K, Paulk A, Basu I, Yousefi A, Zorowitz S, et al. Treating refractory mental illness with closed-loop brain stimulation: progress towards a patient-specific transdiagnostic approach. *Exp Neurol.* (2017) 287:361. doi: 10.1016/j.expneurol.2016.07.021

29. Bush G, Shin LM. The multi-source interference task: an fMRI task that reliably activates the cingulo-frontal-parietal cognitive/attention network. *Nat Protoc.* (2006) 1:308–13. doi: 10.1038/nprot.2006.48

30. Widge A, Zorowitz S, Basu I, Paulk A, Cash SEE, Deckersbach T, et al. Deep brain stimulation of the internal capsule enhances human cognitive control and prefrontal cortex function. *Nat Commun.* (2019) 10:1536. doi: 10.1038/s41467-019-09557-4

31. Samartin-Veiga N, González-Villar AJ, Carrillo-de-la-Peña MT. Neural correlates of cognitive dysfunction in fibromyalgia patients: reduced brain electrical activity during the execution of a cognitive control task. *Neuroimage.* (2019) 23:101817. doi: 10.1016/j.nicl.2019.101817

32. González-Villar AJ, Carrillo-de-la-Peña MT. Brain electrical activity signatures during performance of the multisource interference task. *Psychophysiology.* (2017) 54:874–81. doi: 10.1111/psyp.12843

33. Hanslmayr S, Pastötter B, Bäuml K, Gruber S, Wimber M, Klimesch W. The electrophysiological dynamics of interference during the stroop task. *J Cogn Neurosci.* (2007) 20:215–25. doi: 10.1162/jocn.2008.20020

34. Folstein JR, Van Petten C. Influence of cognitive control and mismatch on the N2 components of the ERP: a review. *Psychophysiology.* (2008) 45:152–70. doi: 10.1111/j.1469-8986.2007.00602.x

35. Friston K, Zeidman P, Litvak V. Empirical Bayes for DCM: a group inversion scheme. *Front Syst Neurosci.* (2015) 9:164. doi: 10.3389/fnsys.2015.00164

36. Pinotsis DA, Robinson P, Graben PB, Friston KJ. Neural masses and fields: modelling the dynamics of brain activity. *Front Comput Neurosci.* (2014) 8:149. doi: 10.3389/fncom.2014.00149

37. Marreiros A, Pinotsis D, Brown P, Friston K. DCM, conductance based models and clinical applications. In: Bhattacharya B, Chowdhury F editors. *Validating NeuroComputational Models of Neurological and Psychiatric Disorders.* Cham: Springer International Publishing (2015). p. 43–70. doi: 10.1007/978-3-319-20037-8_3

38. Pinotsis DA, Friston KJ. Extracting novel information from neuroimaging data using neural fields. *EPJ Nonlinear Biomed Phys.* (2014) 2:5. doi: 10.1140/epjnbp18

39. Pinotsis DA, Buschmann TJ, Miller EK. Working memory load modulates neuronal coupling. *Cereb Cortex.* (2018) 29:1670–81. doi: 10.1093/cercor/bhy065

40. Jafarian, A, Zeidman P, Litvak V, Friston K. Structure learning in coupled dynamical systems and dynamic causal modelling. *Philos Trans R Soc A.* (2019) 377:2160. doi: 10.1098/rsta.2019.0048

41. Ahmadizadeh S, Karoly PJ, Nešiæ D, Grayden DB, Cook MJ, Soudry D, et al. Bifurcation analysis of two coupled Jansen-Rit neural mass models. *PLoS One.* (2018) 13:e0192842. doi: 10.1371/journal.pone.0192842

42. Goodfellow M, Schindler K, Baier G. Intermittent spike–wave dynamics in a heterogeneous, spatially extended neural mass model. *Neuroimage.* (2011) 55:920. doi: 10.1016/j.neuroimage.2010.12.074

43. Grimbert F, Faugeras O. Bifurcation analysis of Jansen's neural mass model. *Neural Comput.* (2006) 182:3052. doi: 10.1162/neco.2006.18.12.3052

44. Basu I, Crocker B, Farnes K, Robertson M, Paulk A, Vellejo D, et al. A neural mass model to predict electrical stimulation evoked responses in human and non human primate brain. *J Neural Eng.* (2018) 15:066012. doi: 10.1088/1741-2552/aae136

45. Roberston JA, Thomas AW, Prato FS, Johansson M, Nittby H. Simultaneous fMRI and EEG during the multi-source interference task. *PLoS One.* (2014) 92:e114599. doi: 10.1371/journal.pone.0114599

46. Vasic N, Walter H, Sambataro F, Wolf RC. Aberrant functional connectivity of dorsolateral prefrontal and cingulate networks in patients with major depression during working memory processing. *Psychol Med.* (2009) 39:977–87. doi: 10.1017/S0033291708004443

47. Schlösser RGM, Wagner G, Koch K, Dahnke R, Reichenbach JR, Sauer H. Fronto-cingulate effective connectivity in major depression: a study with fMRI and dynamic causal modeling. *Neuroimage.* (2008) 43:645–55. doi: 10.1016/j.neuroimage.2008.08.002

48. Murrough J, Abdallah C, Anticevic A, Collins K, Geha P, Averill L, et al. Reduced global functional connectivity of the medial prefrontal cortex in major depressive disorder. *Hum Brain Mapp.* (2016) 37:3214–23. doi: 10.1002/hbm.23235

49. Jansen BH, Rit VG. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biol Cybern.* (1995) 73:357–66. doi: 10.1007/BF00199471

50. David O, Kiebel SJ, Harrison LM, Mattout J, Kilner JM, Friston KJ. Dynamic causal modeling of evoked responses in EEG and MEG. *Neuroimage.* (2006) 30:1255–72. doi: 10.1016/j.neuroimage.2005.10.045

51. Trivedi MH, McGrath PJ, Fava M, Parsey RV, Kurian BT, Phillips ML, et al. Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): rationale and design. *J Psychiatr Res.* (2016) 78(Suppl. C):11–23. doi: 10.1016/j.jpsychires.2016.03.001

52. Breiman L. Random forests. *Mach Learn.* (2001) 45:5. doi: 10.1023/A:1010933404324

53. Kursa MB, Rudnicki W. Feature selection with the boruta package. *J Stat Softw.* (2010) 36:1–13. doi: 10.18637/jss.v036.i11

54. Chawla N, Bowyer K, Hall L, Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. *J Artif Intell Res.* (2002) 16:321–57. doi: 10.1613/jair.953

55. Stone M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J R Stat Soc Ser B.* (1977) 39:44. doi: 10.1111/j.2517-6161.1977.tb01603.x

56. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. *Stat Survey.* (2010) 4:40–79. doi: 10.1214/09-SS054

57. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* (1995) 20:273–97. doi: 10.1007/BF00994018

58. Andreas Ziegler MNW. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw.* (2017) 77:i01.

59. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* San Francisco, CA (2016). doi: 10.1145/2939672.2939785

60. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv* [Preprint]. (2010). arXiv: 2010.16061.

61. Lundberg SLSI. A unified approach to interpreting model predictions. *arXiv* [Preprint]. (2017). arXiv:1705.07874.

62. van der Maaten GH. t-SNE/LJP. *J Mach Learn Res.* (2008) 8:2579.

63. Forgy EW. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics.* (1965) 21:768–9.

64. Lloyd, SP. Least square quantization in PCM. *Bell Tele Lab Pap.* (1957). 18:5.

65. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Comput Appl Math.* (1987) 20:53–65. doi: 10.1016/0377-0427(87)90125-7

66. Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ. Bayesian model selection for group studies. *Neuroimage.* (2009) 46:1004–17. doi: 10.1016/j.neuroimage.2009.03.025

67. Lo M, Younk R, Widge AS. Paired electrical pulse trains for controlling connectivity in emotion-related brain circuitry. *IEEE Trans Neural Syst Rehabil Eng.* (2020) 202:2721–30. doi: 10.1109/TNSRE.2020.3030714

68. Yang Y, Qiao S, Sani OG, Sedillo JI, Ferrentino B, Pesaran B, et al. Modelling and prediction of the dynamic responses of large-scale brain networks during direct electrical stimulation. *Nat Biomed Eng.* (2021) 5:324–45. doi: 10.1038/s41551-020-00666-w

69. Sani OG, Yang Y, Lee MB, Dawes HE, Chang EF, Shanechi MM. Mood variations decoded from multi-site intracranial human brain activity. *Nat Biotechnol.* (2018) 106:954–61. doi: 10.1038/nbt.4200

70. Olsen S, Basu I, Bilge MT, Kanabar A, Boggess MJ, Rockhill AP, et al. Case report of dual-site neurostimulation and chronic recording of cortico-striatal circuitry in a patient with treatment refractory obsessive compulsive disorder. *Front Hum Neurosci.* (2020) 14:569973. doi: 10.3389/fnhum.2020.569973

71. Shenhav A, Musslick S, Lieder F, Kool W, Griffiths TL, Cohen JD, et al. Toward a rational and mechanistic account of mental effort. *Annu Rev Neurosci.* (2017) 40:99–124. doi: 10.1146/annurev-neuro-072116-031526

72. Froemer R, Lin H, Wolf C, Inzlicht M, Shenhav A. When effort matters: expectations of reward and efficacy guide cognitive control allocation. *bioRxiv* [Preprint]. (2020). bioRxiv: 95935. doi: 10.1101/2020.05.14.095935

73. Wu W, Zhang Y, Jiang J, Lucas MV, Fonzo GA, Rolle CE, et al. An electroencephalographic signature predicts antidepressant response in major depression. *Nat Biotechnol.* (2020) 38:439–47. doi: 10.1038/s41587-019-0397-3

74. Williams LM, Rush AJ, Koslow SH, Wisniewski SR, Cooper NJ, Nemeroff CB, et al. International study to predict optimized treatment for depression (iSPOT-D), a randomized clinical trial: rationale and protocol. *Trials.* (2011) 12:4. doi: 10.1186/1745-6215-12-4

75. Arik SÖ, Pfister T. Tabnet: attentive interpretable tabular learning. *Proc AAAI Conf Artif Intell.* (2021) 35:6679–87. doi: 10.1609/aaai.v35i8.16826

76. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *Neuroimage.* (2017) 145:137–65. doi: 10.1016/j.neuroimage.2016.02.079

77. Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, et al. Towards reproducible brain-wide association studies. *bioRxiv* [Preprint]. (2020). bioRxiv:020.08.21.257758. doi: 10.1101/2020.08.21.257758

78. Grzenda A, Kraguljac N, McDonald WM, Nemeroff C, Torous J, Alpert J, et al. Evaluating the machine learning literature: a primer and user's guide for psychiatrists. *Am J Psychiatry.* (2021) 178:715–29. doi: 10.1176/appi.ajp.2020.20030250

79. Rolle CE, Fonzo GA, Wu W, Toll R, Jha MK, Cooper C, et al. Cortical connectivity moderators of antidepressant vs placebo treatment response in major depressive disorder: secondary analysis of a randomized clinical trial. *JAMA Psychiatry.* (2020) 77:397–408. doi: 10.1001/jamapsychiatry.2019.3867

80. Grzenda A, Widge AS. Electroencephalographic biomarkers for predicting antidepressant response: new methods, old questions. *JAMA Psychiatry.* (2020) 77:347–8. doi: 10.1001/jamapsychiatry.2019.3749

81. Poldrack RA, Gorgolewski KJ. Making big data open: data sharing in neuroimaging. *Nat Neurosci.* (2014) 171:1510–7. doi: 10.1038/nn.3818

82. Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci.* (2017) 20:365–77. doi: 10.1038/nn.4478

83. Horesh N, Klomek AB, Apter A. Stressful life events and major depressive disorders. *Psychiatry Res.* (2008) 160:192–9. doi: 10.1016/j.psychres.2007.06.008

84. Wattenberg M, Viégas F, Johnson I. How to use t-SNE effectively. *Distill.* (2016) 1:e2. doi: 10.23915/distill.00002

85. Taskesen E, Reinders MJ. 2D representation of transcriptomes by t-SNE exposes relatedness between human tissues. *PLoS One.* (2016) 11:e0149853. doi: 10.1371/journal.pone.0149853

86. Cieslak MC, Castelfranco AM, Roncalli V, Lenz PH, Hartline DK. t-Distributed stochastic neighbor embedding (t-SNE): a tool for eco-physiological transcriptomic analysis. *Mar Genomics.* (2020) 51:100723. doi: 10.1016/j.margen.2019.100723

87. Van Der Maaten L. Learning a parametric embedding by preserving local structure. *Artif Intell Stat.* (2009) 5:384–91.

88. Pezzotti N, Höllt T, Lelieveldt B, Eisemann E, Vilanova A. Hierarchical stochastic neighbor embedding. *Comput Graph Forum.* (2016) 35:21–30. doi: 10.1111/cgf.12878

89. Wendling F, Benquet P, Bartolomei F, Jirsa V. Computational models of epileptiform activity. *J Neurosci Methods.* (2016) 260:233. doi: 10.1016/j.jneumeth.2015.03.027

90. Ruffini G, Sanchez-Todo R, Dubreuil L, Salvador R, Pinotsis D, Miller EK, et al. P118 A biophysically realistic laminar neural mass modeling framework for transcranial current stimulation. *Clin Neurophysiol.* (2020) 131:e78. doi: 10.1016/j.clinph.2019.12.229

91. Oestreich LK, Randeniya R, Garrido MI. Structural connectivity facilitates functional connectivity of auditory prediction error generation within a fronto-temporal network. *bioRxiv* [Preprint]. (2018): doi: 10.1101/365072

92. Ranlund S, Adams RA, Díez Á, Constante M, Dutt A, Hall MH, et al. Impaired prefrontal synaptic gain in people with psychosis and their relatives during the mismatch negativity. *Hum Brain Mapp.* (2016) 37:351–65. doi: 10.1002/hbm.23035

93. Kiebel S, David O, Friston K. Dynamic causal modelling of evoked responses in EEG/MEG with lead-field parameterization. *Neuroimage.* (2006) 30:1273–84. doi: 10.1016/j.neuroimage.2005.12.055

94. Kiebel S, Daunizeau J, Phillips C, Friston K. Variational bayesian inversion ofthe equivalent current dipole model in EEG/MEG. *Neuroimage.* (2008) 39:728–41. doi: 10.1016/j.neuroimage.2007.09.005

95. Bonaiuto J, Rossiter H, Meyer S, Adams N, Little S, Callaghan MF, et al. Non-invasive laminar inference with MEG: comparison of methods and source inversion algorithms. *Neuroimage.* (2018) 167:372. doi: 10.1016/j.neuroimage.2017.11.068

96. Hvitfeldt E. *themis: Extra Recipes Steps for Dealing with Unbalanced Data.* (2020). Available online at: https://CRAN.R-project.org/package=themis (accessed July 02, 2022).

97. Zhao Y, Wong ZSY, Tsui KL. A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. *J Healthc Eng.* (2018) 2018:6275435. doi: 10.1155/2018/6275435

98. Abeysinghe W, Hung CC, Bechikh S, Wang X, Rattani A. Clustering algorithms on imbalanced data using the SMOTE technique for image segmentation. In: *Proceedings of the 2018 Conference on Research in Adaptive and Convergent Systems - RACS*. Honolulu, HI (2018). doi: 10.1145/3264746.3264774

99. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics.* (2020) 21:6. doi: 10.1186/s12864-019-6413-7