Accurate and Interpretable Prediction of Antidepressant Treatment Response from Receptor-informed Neuroimaging

Hanna M. Tolle,[1,*] Andrea I Luppi,[2,3,4] Timothy Lawn,[5] Leor Roseman,[6] David Nutt,[7] Robin L. Carhart-Harris,[8,9] and Pedro A. M. Mediano[1,10,†]

[1] *Department of Computing, Imperial College London*
[2] *Centre for Eudaimonia and Human Flourishing, Department of Psychiatry, University of Oxford*
[3] *Montreal Neurological Institute, McGill University*
[4] *Division of Information Engineering and St John's College, University of Cambridge*
[5] *Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital and Harvard Medical School*
[6] *Department of Psychology, University of Exeter*
[7] *Division of Psychiatry, Imperial College London*
[8] *Departments of Neurology, Psychiatry and Behavioral Sciences, Weill Institute for Neurosciences, University of California San Francisco*
[9] *Centre for Psychedelic Research, Imperial College London*
[10] *Division of Psychology and Language Sciences, University College London*

Conventional antidepressants show moderate efficacy in treating major depressive disorder. Psychedelic-assisted therapy holds promise, yet individual responses vary, underscoring the need for predictive tools to guide treatment selection. Here, we present `graphTRIP`(graph-based Treatment Response Interpretability and Prediction) – a geometric deep learning architecture that enables three advances: 1) accurate prediction of post-treatment depression severity using only pretreatment clinical and neuroimaging data; 2) identification of robust, patient-specific biomarkers; and 3) causal analysis of treatment effects and underlying mechanisms. Trained on data from a clinical trial comparing psilocybin and escitalopram (NCT03429075), `graphTRIP` achieves strong predictive accuracy ($r = 0.75$, $p < 10^{-8}$), and generalises both to an independent dataset and across brain atlases. The model links better outcomes to reduced functional coupling within serotonin systems, and broader serotonergic integration with sensory-motor networks. Finally, causal analysis reveals a group-level advantage of psilocybin over escitalopram, but also identifies individuals with specific stress-related neuromodulatory profiles who may benefit more from escitalopram. Overall, this work advances precision medicine and biomarker discovery in depression.

Keywords: precision medicine, geometric deep learning, depression, interpretability, psychedelic therapy

INTRODUCTION

Major depressive disorder (MDD) is a debilitating psychiatric disorder that places a serious burden on individuals, their families, and society. Already a leading cause of disability worldwide [1], its incidence has continued to rise in recent years [2, 3], highlighting the urgent need for a more effective treatment strategy.

Selective serotonin reuptake inhibitors (SSRIs) are the most widely prescribed antidepressant drugs [4]. These compounds act by inhibiting the reuptake of serotonin (5-HT) from the synaptic cleft through blockade of the serotonin transporter (5-HTT, a.k.a. SERT) [4]. While SSRIs are among the most effective treatments currently available, less than 40% of patients achieve remission after the first treatment course, and approximately one in three fails to remit even after multiple successive treatment trials with different SSRIs [5, 6].

The psychedelic compound psilocybin has recently emerged as a promising alternative to conventional antidepressants [7], with some evidence indicating efficacy even in SSRI-resistant patients [8, 9]. Like most psychedelic drugs, psilocybin primarily acts as an agonist at the serotonin receptors 5-HT2A and 5-HT1A [10]. Notably, in contrast to SSRIs, which require continuous daily use, psilocybin treatment typically involves only one or two therapist-guided drug sessions in addition to psychological therapy [11].

However, like all treatments, psychedelic treatments come with associated risks. Individual responses to psychedelics vary widely, with one study reporting increased suicidal ideation in certain individuals [9]. This variability likely reflects both the underlying heterogeneity of MDD [12] and the marked individual differences in response to psychedelics [13]. Thus, to safely deploy psychedelic treatments and improve MDD prognosis worldwide, we need a means to predict how an individual patient will respond to a given treatment, allowing for a more targeted approach and reducing the prolonged trial-and-error process in antidepressant prescribing.

One promising approach is to predict treatment outcomes from pre-treatment neuroimaging data. Because brain function is assumed to emerge from complex interactions between specialised regions, it is natural to model the brain as a network, or brain graph, where nodes represent brain regions (defined by a brain atlas) and edges

* h.tolle23@imperial.ac.uk
† p.mediano@imperial.ac.uk

capture anatomical or functional connectivity [14]. Functional connectivity (FC) refers to statistical dependencies – typically correlations – between brain regional activity, and its disruption has been consistently linked to psychiatric disorders, including MDD [15–17], making it a particularly relevant feature for antidepressant response prediction. Furthermore, recent approaches such as REACT (Receptor-Enriched Analysis of functional Connectivity by Targets) enable the integration of fMRI data with normative maps of molecular targets, derived from PET imaging [18–21], providing features of neuromodulatory activity that have proven valuable for studying neurological and psychiatric disorders [22, 23], as well as the acute effects of psychedelics [24]. Together, these developments position molecularly informed brain graphs as a clinically relevant basis for predicting antidepressant treatment outcomes.

Indeed, previous studies using machine learning (ML) have achieved notable successes in predicting antidepressant treatment outcome from brain graph features [25–27], yet several key challenges have so far hindered clinical translation. For instance, prior approaches have largely focused on binary classification of treatment response, or remission, because sample sizes of available datasets are often too small for accurate prediction of post-treatment depression scores [25, 27, 28]. Furthermore, few models have been validated on independent datasets, and those that have, show limited generalisation performance [25]. Moreover, conventional ML architectures require fixed input sizes, which constrains the analysis to a specific brain atlas. However, the choice of atlas can substantially alter findings, and no universally accepted standard exists [29, 30].

Another key barrier to clinical translation is the lack of interpretability – most classical ML models operate as "black boxes," offering little insight into the basis of their predictions. In a clinical context, this is unacceptable. To ensure safe and reliable decisions, it is crucial to understand why a specific treatment outcome was predicted. Importantly, interpretability must be available at the individual patient level to verify that each decision is based on valid, biologically-relevant features. Although there is a broad and fast-moving literature on interpretability for large language models [31, 32], its application to clinical ML has thus far remained limited [33]. Closing this gap is essential, not only for safety, but also for identifying biomarkers of treatment responsiveness and advancing our understanding of antidepressant treatment.

Beyond interpretability, a further critical challenge lies in estimating *causal* treatment effects. Predictive models can identify features associated with good or poor response, but they cannot reliably determine whether one treatment would have worked better than another for a given patient. To enable data-driven treatment selection, we must therefore go beyond prediction and incorporate causal inference methods that can distinguish shared biomarkers of treatment responsiveness from treatment-specific moderators.

Here, we present a geometric deep learning (GDL) approach for predicting antidepressant treatment response from pre-treatment clinical and fMRI data in patients treated with escitalopram or psilocybin. GDL is uniquely suited to learning from brain graphs, enabling more powerful and biologically-informed predictions than conventional ML. Our novel architecture, dubbed `graphTRIP`, addresses key roadblocks to clinical translation in a fundamentally new way: 1) it directly predicts post-treatment depression scores (QIDS) with high accuracy; 2) it generalises to an independent dataset of SSRI-resistant patients, treated with psilocybin; 3) it flexibly adapts to different brain atlases, maintaining significant predictions without retraining; 4) it enables rich interpretability analyses – particularly through our novel method GRAIL (Gradient Alignment for Interpreting Latent-variable models), which quantifies learned associations between treatment outcome and any brain-graph biomarkers of interest. Finally, we extend our model within a causal inference framework [34], enabling us to estimate the expected difference in outcome under psilocybin versus escitalopram for each patient, and discern treatment-specific moderators.

Leveraging state-of-the-art GDL and causal inference, our approach advances the field of antidepressant response prediction, offering a robust tool for both biomarker discovery and clinical decision-making.

RESULTS

The primary dataset [35] in our study included 42 MDD patients who participated in a double-blind randomised controlled trial (DB-RCT) with two treatment arms: psilocybin ($N = 22$) and escitalopram ($N = 20$) (Fig. 1a). To assess the generalisability of our model, we used an independent dataset [8] consisting of 16 patients with treatment-resistance depression (TRD), who received psilocybin in an open-label trial (Fig. 1b). In both datasets, we used baseline resting-state fMRI scans and clinical data (QIDS and BDI depression scores, and a binary indicator of prior SSRI use) and the drug condition (psilocybin or escitalopram) as input features to our model. The primary prediction target was post-treatment QIDS.

Our model, called `graphTRIP` (graph-based Treatment Response Interpretability and Prediction), combines a variational graph autoencoder (VGAE) with a multilayer perceptron (MLP) (Fig. 1c). The VGAE learns a latent representation of brain graphs. The MLP predicts post-treatment QIDS. Brain graphs were constructed from baseline fMRI data, with edges defined by thresholded FC ($|FC| > 0.5$). The threshold was chosen to approximate the density of typical structural connectomes, though performance generalised to alternative thresholding choices (Supp. Fig. 7). Node features encoded REACT values for 5-HT1A, 5-HT2A, and 5-HTT, denoted as R5-HT1A, R5-HT2A, and R5-HTT to distin-

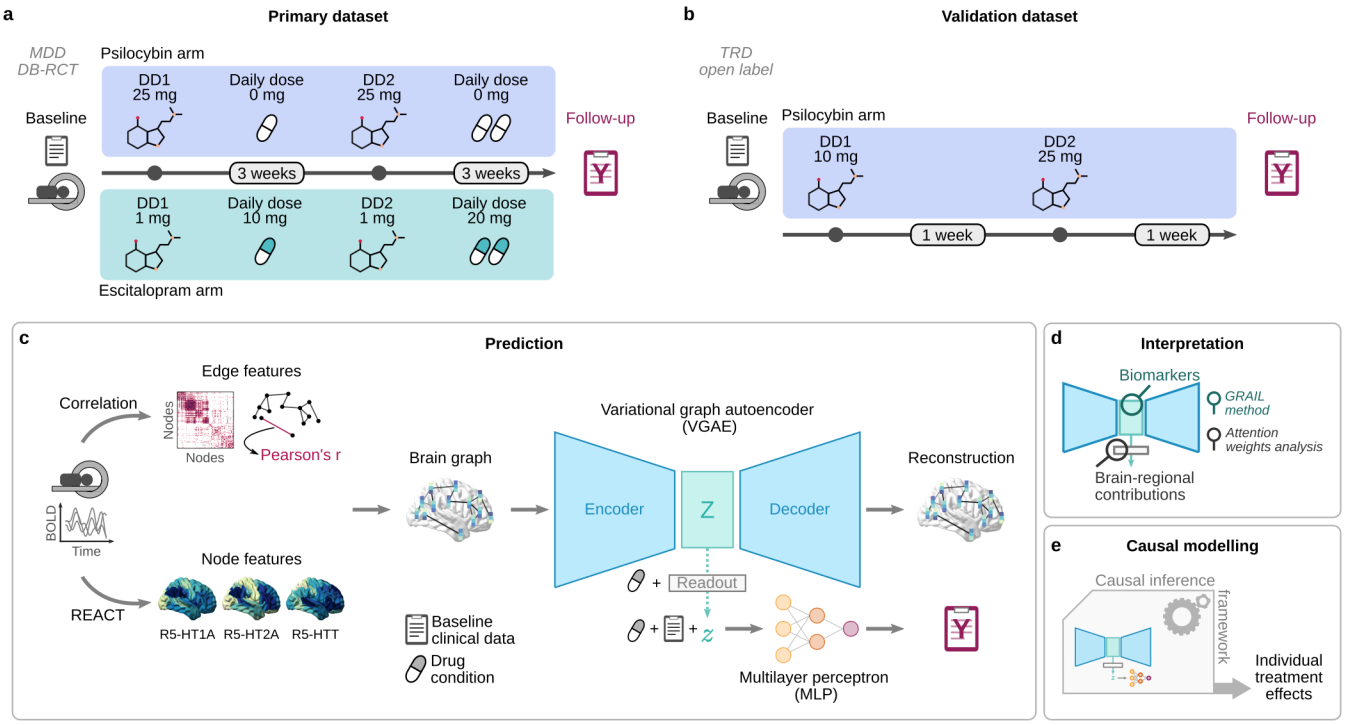


FIG. 1. **Overview of study design and methodological framework.** **a**, Primary dataset: 42 MDD patients from a double-blind RCT, comparing escitalopram ($N = 20$) and psilocybin ($N = 22$) treatment. Escitalopram and psilocybin groups, respectively, received inactive and high doses of psilocybin on each dosing day (DD1, DD2), conducted in a safe environment with psychological support. Escitalopram or placebo was administered daily for three weeks after each session. Depression severity was assessed using QIDS three weeks after DD2. **b**, Validation dataset: 16 TRD patients received low (DD1) and high (DD2) doses of psilocybin in an open-label trial. Depression severity was assessed using QIDS one week after DD2. **c**, Brain graphs were constructed from baseline fMRI, with node features encoding serotonin system REACT maps and edges reflecting functional connectivity. A variational graph autoencoder (VGAE) learned latent graph embeddings, which were passed to a multilayer perceptron (MLP) for predicting post-treatment QIDS. **d**, Two complementary interpretability analyses reveal the predictive contributions of brain regions and biomarkers. **e** We extend our model within a causal inference framework to estimate individual treatment effects and identify treatment-specific moderators.

guish them from the corresponding normative receptor densities. REACT provides subject-specific estimates of how strongly each brain region's activity is functionally coupled with a molecular target of interest. The VGAE encoder maps each brain region to a latent space, and the decoder reconstructs the original graph from the resulting matrix of latent node vectors, Z. A readout layer then aggregates Z into a graph-level representation vector $\boldsymbol{z}$, conditioned on the treatment. This vector, combined with the drug condition and baseline clinical data, serves as input to the MLP.

We applied two analyses to probe model predictions: 1) analysis of the VGAE readout to assess the predictive contributions of brain regions, and 2) analysis of the latent space using our GRAIL method to identify biomarkers of treatment responsiveness (Fig. 1d).

Moving beyond predictive modelling, we further extended `graphTRIP` within a causal inference framework to estimate individual treatment effects and identify treatment-specific moderators (Fig. 1e).

Accurate, robust predictions of treatment response

`graphTRIP` achieved a strong correlation between true and predicted post-treatment QIDS scores ($r = 0.7047$, $p < 1.9 \times 10^{-7}$; Fig. 2a) and showed good reconstruction performance (Figure 2c-d). In contrast, a control MLP, trained only on the clinical data and the drug condition (i.e., without neuroimaging data) performed substantially worse ($r = 0.3947$, $p < 9.7 \times 10^{-3}$; Fig. 2a). Further analysis suggests that the control MLP primarily predicted the mean QIDS scores for each drug condition, which was higher in the escitalopram group [35]. After controlling for the drug condition, the partial correlation between true and predicted scores remained significant for `graphTRIP` ($r = 0.6607$, $p < 1.9 \times 10^{-6}$), but not for the control MLP ($r = 0.2663$, $p = 0.0883$; Fig. 2b). Reducing the neuroimaging data using PCA or t-SNE also failed to produce meaningful predictions, highlighting the advantage of `graphTRIP`'s VGAE (Supp. Fig. 8).

To assess the contribution of latent brain-graph features in the `graphTRIP` model, we performed permuta-

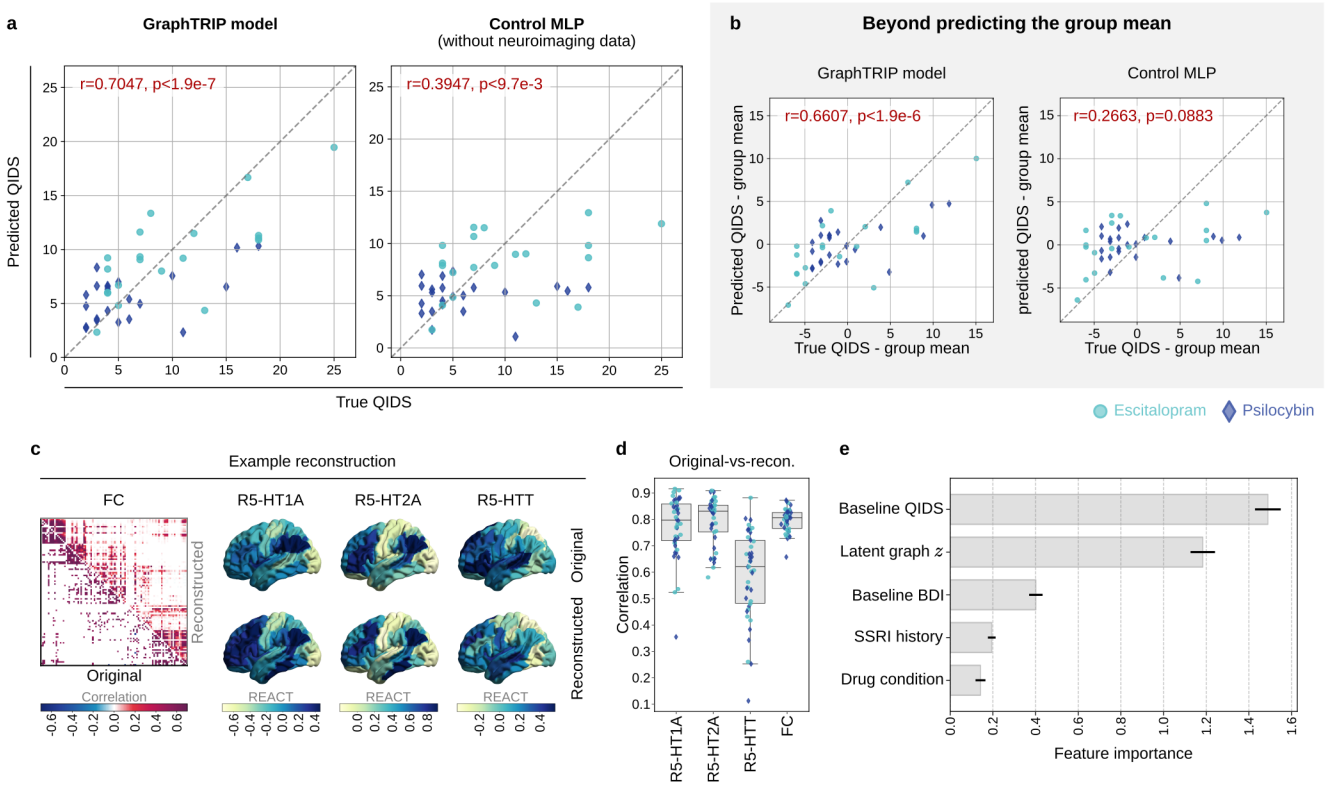


FIG. 2. **Reconstruction and prediction performance of the `graphTRIP` model.** **a**, `graphTRIP` significantly predicts post-treatment QIDS, outperforming a control MLP trained only on clinical data and drug condition, without latent neuroimaging features ($\boldsymbol{z}$). **b**, After controlling for the difference in post-treatment QIDS between drug conditions, the partial correlation between true and predicted scores remains significant for the `graphTRIP` model, but not for the control MLP. **c**, Example VGAE reconstructions of FC and REACT node features. **d**, Correlations between original and reconstructed FC and node features across all patients. **e**, Permutation importance analysis, showing the increase in mean absolute error (MAE) caused by shuffling individual features across patients. Intuitively, shuffling important features results in greater prediction error. Bars indicate the mean increase in MAE across 50 random permutations; error bars denote standard error.

tion importance analysis, measuring the impact of feature shuffling on prediction error (mean absolute error) across 50 repetitions. Higher increases in error indicate greater feature importance. As expected, baseline QIDS was the most important predictor (Fig. 2e). The latent brain-graph representation $\boldsymbol{z}$, ranked second.

We additionally trained a `graphTRIP` model to predict post-treatment BDI scores. We achieved significant predictions ($r = 0.5490$, $p < 1.7 \times 10^{-4}$; Supp. Fig. 9), confirming the robustness of our approach.

Generalisation across brain atlases

We evaluated `graphTRIP`'s ability to generalise across brain parcellations by testing it on atlases different from the one used for training (Schaefer 100 [36]). On Schaefer 200, which has twice as many parcels, `graphTRIP` produced highly accurate reconstructions and maintained significant prediction performance ($r = 0.5633$, $p < 1.0 \times 10^{-4}$; Fig. 3a). We further replicated this result on the AAL atlas [37], which differs fundamentally from Schaefer and includes subcortical regions, where `graphTRIP` also generalised ($r = 0.5205$, $p < 4.1 \times 10^{-4}$; Supp. Fig. 10).

Once a parcellation is selected, all brain graphs have a consistent number of nodes and direct anatomical correspondence across subjects. Exploiting this, we trained an atlas-bound model on the full, unthresholded Schaefer 100 FC matrices. Like `graphTRIP`, it includes a representation learning module (analogous to the VGAE) and an MLP for QIDS prediction. However, unlike `graphTRIP`, it processes entire FC columns as node features and does not incorporate edges, enabling it to handle the full FC, at the trade-off of being restricted to a fixed parcellation.

We trained a hybrid predictor combining the strengths of both models by concatenating their latent representations. Specifically, after pretraining each model separately, we froze their representation learning modules and trained a new MLP on the combined latent features, along with baseline clinical data and the drug condition (Fig. 3b). This hybrid approach yielded the best prediction performance ($r = 0.7507$, $p < 1.0 \times 10^{-8}$; Fig. 3c), suggesting that the models capture comple-

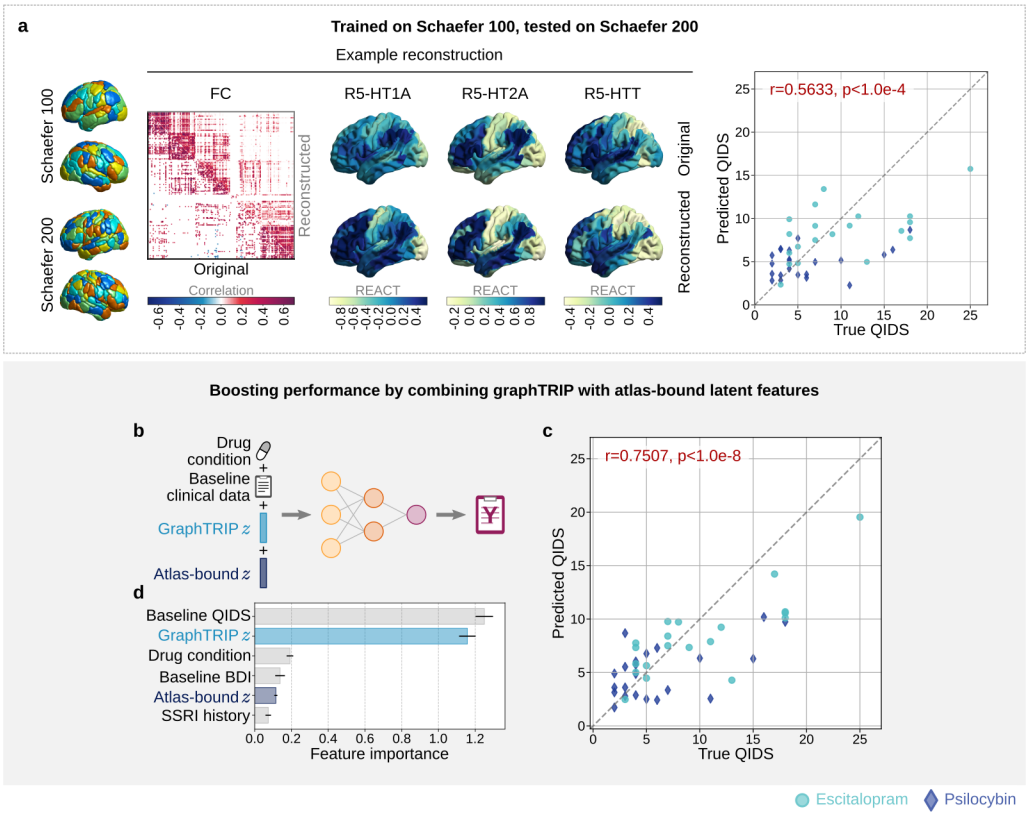


FIG. 3. **Generalisation across brain parcellations and boosting performance with a hybrid model.** **a**, The `graphTRIP` model, trained on Schaefer 100, generalises to Schaefer 200 while maintaining strong reconstruction and prediction performance. The panel shows the Schaefer 200 atlas (left), FC and node feature reconstructions for an example subject (centre), and the MLP's prediction performance (right). **b**, Schematic of the hybrid model, where a new MLP is trained on drug condition, baseline clinical data, and the concatenated latent representations of the separately pretrained atlas-bound and `graphTRIP` models. **c**, The hybrid model achieves the best prediction performance. **d**, Permutation importance analysis shows that `graphTRIP`'s latent features contribute significantly more to predictions than those of the atlas-bound model.

mentary information. Importance analysis revealed that `graphTRIP`'s latent features were substantially more influential (Fig. 3d), suggesting that the atlas-bound model acts as an auxiliary component, offering a performance boost if committing to a parcellation is acceptable.

Validation in an independent dataset

We evaluated `graphTRIP` on an independent dataset that differed from the primary dataset in key aspects (Fig. 1a), including 1) a single treatment arm (psilocybin); 2) patients with treatment-resistant depression (TRD); 3) a different psilocybin treatment protocol; and 4) a different follow-up assessment time point.

Fine-tuning `graphTRIP` on the validation dataset required addressing some of the dataset differences. Since all patients in the validation dataset had TRD and received psilocybin, SSRI history and drug condition inputs were constant and hence uninformative for training. Thus, we pretrained a modified `graphTRIP` model on the primary dataset, removing drug condition and SSRI history from the MLP (Fig. 4a), with only a small drop in prediction accuracy ($r = 0.6393$, $p < 5.2 \times 10^{-6}$). The pretrained model was then transferred, freezing VGAE weights and fine-tuning only the MLP on the new dataset (see Methods). Note that since pretraining involved 6-fold cross-validation (CV), we obtained six pretrained models. We fine-tuned each of these models separately, and averaged final predictions across models.

Even without fine-tuning, the pretrained VGAEs maintained high reconstruction accuracy on the validation dataset (Fig. 4b). Crucially, fine-tuned models achieved a significant correlation between true scores and mean predictions (Fig. 4c). Notably, given the small validation sample ($N = 16$), training from scratch proved infeasible, whereas pretraining and fine-tuning consistently enabled successful transfer (Fig. 4d).

Model attention reflects cortical hierarchy

`graphTRIP`'s attention module provides a natural way to assess regional contributions to predictions: since the

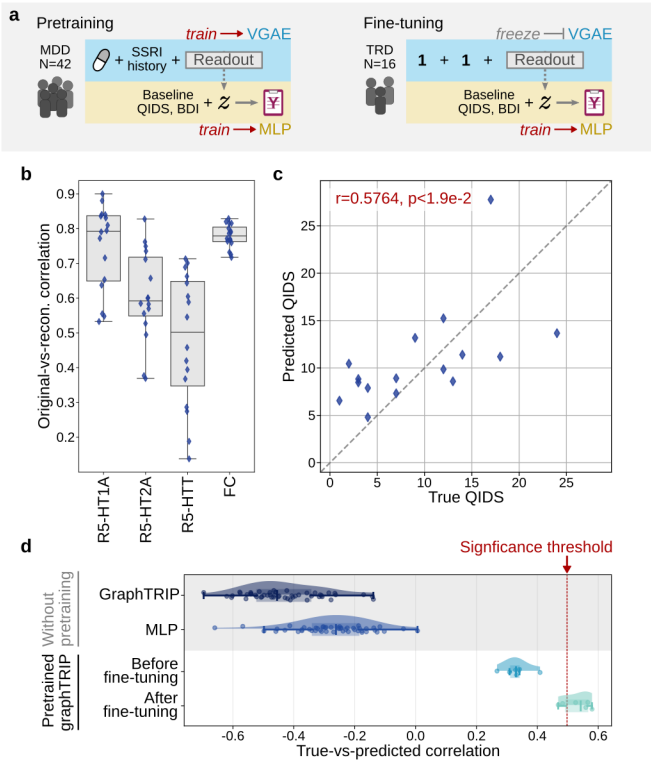


FIG. 4. **Transfer learning enables generalisation to an independent dataset. a**, A modified `graphTRIP` model was pretrained on the primary dataset, excluding from the MLP variables not available in the validation dataset. The pretrained model was then fine-tuned on the validation dataset, training only the MLP and freezing the VGAE. **b**, Correlations between original and reconstructed FC and node features across all patients. **c**, True vs. predicted QIDS scores using the mean prediction across six fine-tuned models. **d**, Training 50 randomly initialised `graphTRIP` models, or 50 MLPs trained on demographic and clinical data, fails due to the small validation sample size. In contrast, all six pretrained models produce positive true-vs-predicted correlations before fine-tuning, and most achieve significance after fine-tuning.

final prediction is based on a weighted average of latent node vectors $\boldsymbol{z}_i$, regions with higher attention weight contribute more strongly to the prediction (Fig. 5a). More specifically, attention weights are computed based on $\boldsymbol{z}_i$ and the drug condition, allowing us to estimate each region's predictive effect at both individual and treatment-group levels. Given the high consistency of attention patterns across CV-fold models (Supp. Fig. 11a), we averaged attention weights across folds for each patient in all subsequent analyses.

First, for each patient, we computed regional attention weights by conditioning the readout on the patient's actual treatment, yielding one attention vector per patient. Averaging across patients revealed a clear population-wide pattern: the model attends most to sensory-motor (SMN) and visual (VIS) regions, and least to association cortices (Fig. 5a). This spatial pattern aligns with the unimodal-transmodal cortical gradient, known to reflect key anatomical and functional differences [20, 38–41]. Indeed, a linear model predicting regional attention weights from unimodal-transmodal axis assignments and ten normative molecular-target density maps, derived from open-access datasets [20, 41, 42] (see Methods), achieved a strong correlation with observed values ($r = 0.8782$, $p < 10^{-4}$ against 1000 spatial autocorrelation-preserving nulls [43]). Dominance analysis confirmed the unimodal-transmodal axis as the strongest predictor, followed by serotonin receptor densities (Fig. 5b).

Second, to assess treatment-specific effects, we computed two attention vectors for each patient by conditioning the model separately on psilocybin and escitalopram. We found a significant shift in attention from unimodal (SMN, VIS) to transmodal (DMN, FPN, LIM) regions when the model predicted outcomes under psilocybin compared to escitalopram (Fig. 5d). This suggests that the model relies more heavily on association cortices when estimating psilocybin response.

Biomarkers of treatment responsiveness

Beyond attention weights, we developed GRAIL to systematically link treatment response predictions to brain graph-derived biomarkers (Fig. 5f). This method exploits the dual structure of `graphTRIP` – a VGAE for learning brain graph representations and an MLP for predicting treatment outcome. For any differentiable biomarker of interest (b), GRAIL computes the alignment (cosine similarity) between the gradient of the model's prediction ($\hat{y}$) and the gradient of b in latent space. Intuitively, these gradients describe how small changes in the latent brain representation affect the prediction $\hat{y}$ and the biomarker b, respectively. Thus, their alignment provides a direct, analytical measure of learned associations: positive alignment indicates that higher biomarker values are associated with worse treatment outcomes (i.e., higher post-treatment QIDS), and negative alignment suggests the opposite. Zero alignment suggests no learned association.

We applied GRAIL to `graphTRIP` and assessed learned associations for a range of candidate biomarkers (Tab. I), chosen to span two biologically and clinically relevant axes: large-scale brain networks (i.e., resting-state networks; RSNs [44]) and neuromodulatory systems. Specifically, we included ten normative molecular-target density maps available from open-access datasets [20, 42], covering major neurotransmitter systems implicated in antidepressant treatment [12]: serotonin (5-HT), dopamine (DA), noradrenaline (NA), acetylcholine (ACh). Candidate biomarkers included i) mean brain graph features (i.e., FC or REACT maps) within each RSN, and ii) spatial correlations between each brain graph feature and the normative molecular target maps. This broad screening strategy leverages GRAIL's ability to discover biomarkers in a data-driven fashion.

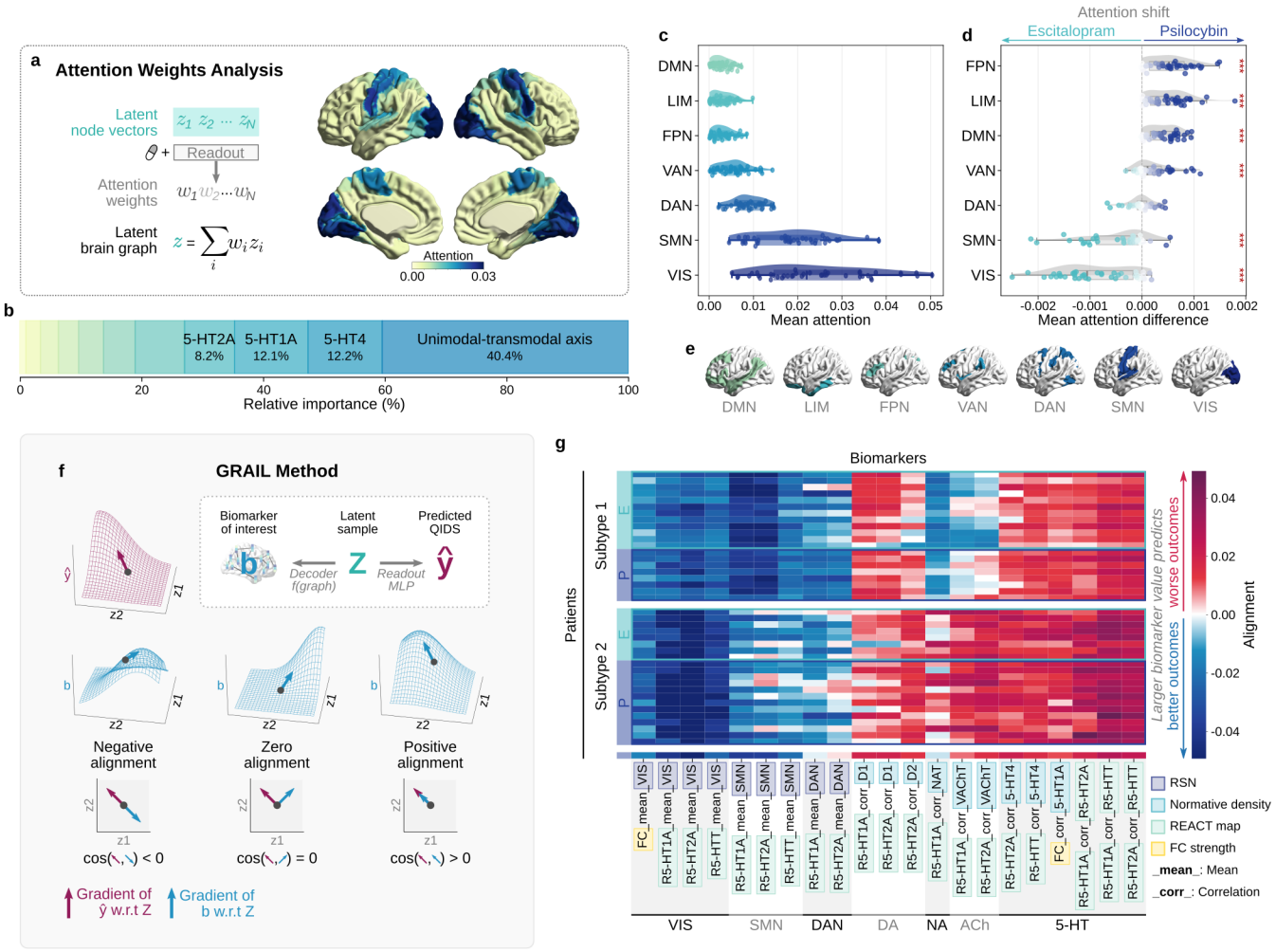


FIG. 5. **Interpretability analysis reveals unique role of unimodal brain regions and population subtypes.** **a**, Population-mean attention weights from the VGAE readout reveal a structured pattern. **b**, Relative contributions of normative molecular target densities and unimodal-transmodal axis assignments to a linear model for explaining population-mean attention weights ($r = 0.8782$, $p < 10^{-4}$). **c**, Mean attention weights within each resting-state network (RSN) for each patient, showing highest values in sensory-motor networks. **d**, Attention shifts from VIS and SMN regions to FPN, DMN, and LIM regions when the model predicts treatment response under psilocybin versus escitalopram. Dots represent per-patient RSN mean attention differences (psilocybin minus escitalopram); asterisks indicate significant shifts after FDR correction ('*': $p < 0.05$, '**': $p < 0.01$, '***': $p < 0.001$). **e**, Resting-state networks. **f**, GRAIL method overview. By computing latent gradients of predicted outcomes ($\hat{y}$) and candidate biomarkers (b), GRAIL quantifies learned associations between biomarkers and treatment response. **g**, GRAIL results across patients (rows), treated with psilocybin (P) or escitalopram (E), showing universal and subtype-specific biomarkers of treatment responsiveness. Biomarker labels are defined in Tab. I.

For each patient and candidate biomarker, we computed gradient alignments for 100 latent brain-graph samples. Results were highly consistent across both latent samples and CV-fold models (Supp. Fig. 11b,c). Thus, each patient's alignment pattern was derived by averaging across samples and folds. Some biomarkers displayed distinct patterns, suggesting the presence of patient subtypes. To identify these, we clustered patients based on the similarity of their mean alignment profiles using the Louvain algorithm for community detection, revealing two treatment responsiveness subtypes, with one patient left unassigned due to an intermediate profile. We focused all following analyses on biomarkers with absolute mean alignment above the 75th percentile.

Among the strongest biomarkers with consistent positive alignment were the correlations between R5-HT1A, R5-HT2A, and R5-HTT REACT maps, as well as correlations of R5-HT2A and R5-HTT with normative 5-HT4 density (Fig. 5g). These results suggest that decreased functional coupling within serotonin systems is broadly associated with better treatment outcomes. Additionally, weaker FC of regions with higher normative 5-HT1A density was also linked to improved response.

We identified two distinct treatment response subtypes

Description	Example
FC modularity based on Louvain or RSN partition	modularity, modularity_rsn
Mean FC of all edges attached to an RSN	FC_mean_VIS
Mean REACT node feature in an RSN	R5-HT1A_mean_VIS, R5-HT2A_mean_SMN
Correlation of regional FC with 10 target* densities	FC_corr_D2, FC_corr_5-HT1A
Correlation of REACT node features with 7 target* densities	R5-HT2A_corr_5-HT4, R5-HTT_corr_D2
Pairwise correlations of REACT node features	R5-HT1A_corr_R5-HT2A, R5-HT2A_corr_R5-HTT

TABLE I. **Candidate biomarkers tested with GRAIL**. *The 10 targets were normative maps of serotonin targets 5-HT1A, 5-HT2A, 5-HTT, 5-HT1B, 5-HT4; dopamine receptors D1, D2, and transporter DAT; noradrenaline transporter NAT; and vesicular acetylcholine transporter VAChT. Densities were derived from open-access PET scan data of healthy subjects [20, 42]. Note: we excluded correlations between REACT node features and molecular maps for 5-HT1A, 5-HT2A, 5-HTT, because these maps were used to compute the REACT features themselves, making such relationships difficult to interpret.

based on biomarker alignment patterns (Fig. 5g). Subtype 1 was especially marked by strong negative alignment of SMN biomarkers, and negative alignment of noradrenaline transporter (NAT) and vesicular acetylcholine transporter (VAChT) biomarkers. In contrast, Subtype 2 featured strong negative alignment of VIS biomarkers, milder SMN effects, and positive alignment of NAT and VAChT biomarkers.

Although the two subtypes did not differ significantly in treatment allocation (Subtype 1: 60% escitalopram; Subtype 2: 62% psilocybin; $\chi^2 = 1.188$, $p = 0.2757$), several biomarkers showed nominally significant alignment differences between drug conditions, including VIS and VAChT biomarkers, as well as correlations among the serotonin REACT maps (Supp. Fig. 12). While these effects did not survive FDR correction, the pattern suggests that `graphTRIP` learned both shared biomarkers, with similar predictive effect across treatments, and treatment-specific biomarkers.

Despite consistent patterns, alignment magnitudes remained low (approx. -0.05 to 0.05 on a scale from -1 to 1) (Fig. 5g), suggesting that no single biomarker alone predicts treatment response. This highlights the need for ML approaches like `graphTRIP`, which can capture global, high-order patterns of brain organisation [45].

Notably, PCA of the alignment matrix revealed that the first two principal components captured 71% of the variance, and their patient loadings correlated significantly with treatment outcome (Supp. Fig. 13a). This indicates that, while individual biomarkers lack strong predictive power, each patient's full GRAIL profile captures clinically meaningful variation. Furthermore, group-averaged alignment values showed a borderline significant correlation ($r = 0.2440$, $p = 0.0520$) with the outcome-predictive strength of each biomarker, as estimated by the direct correlation between outcome and brain-graph derived biomarker values (Supp. Fig. 13b). This confirms that GRAIL identifies meaningful biomarkers, while also capturing higher-order dependencies that go beyond simple univariate associations.

Estimating individual treatment effects

Identifying the most effective treatment for a given patient relies on estimating causal treatment effects. This can be achieved by comparing outcomes between treatment groups that are statistically equivalent in all other covariates (e.g., baseline clinical and brain-graph data).

RCTs, like the one underlying our primary dataset, are designed to approximate such conditions. However, small-sample effects such as dropouts can introduce subtle biases. Indeed, we found that treatment assignment could be predicted from pre-treatment data, primarily due to lower – though non-significant – baseline QIDS scores in the psilocybin group (Supp. Fig. 14a). As a result, `graphTRIP` cannot reliably infer causal effects. To address this, we developed `X-graphTRIP`, which extends the `graphTRIP` architecture within the X-learner causal inference framework [34]. Together, these models provide complementary insights: `graphTRIP` predicts treatment outcome and identifies treatment responsiveness biomarkers, while `X-graphTRIP` estimates individual treatment effects (ITEs) and reveals treatment-specific moderators.

In the X-learner framework (Fig. 6a), two so-called T-learners are first trained independently on patients from each treatment group. These T-learners are then used to compute pseudo-ITE labels by comparing observed outcomes with counterfactual predictions. Finally, an X-learner – `X-graphTRIP` – is trained on the full dataset to predict these ITEs. The idea is that the X-learner can leverage the entire sample to learn a smooth, generalisable ITE function from the imputed pseudo-labels. To ensure unconfoundedness, we removed pre-treatment QIDS from the input features of all models (T- and X-learners), and defined ITEs in terms of the expected difference in QIDS change from baseline under psilocybin versus escitalopram. Negative ITEs indicate that psilocybin is expected to lead to a greater reduction in QIDS compared to escitalopram, and vice versa for positive values. We confirmed that treatment assignment was not predictable from `X-graphTRIP`'s latent brain-graph representations (Supp. Fig. 14b), confirming that the model masks treatment-associated confounds and satis-

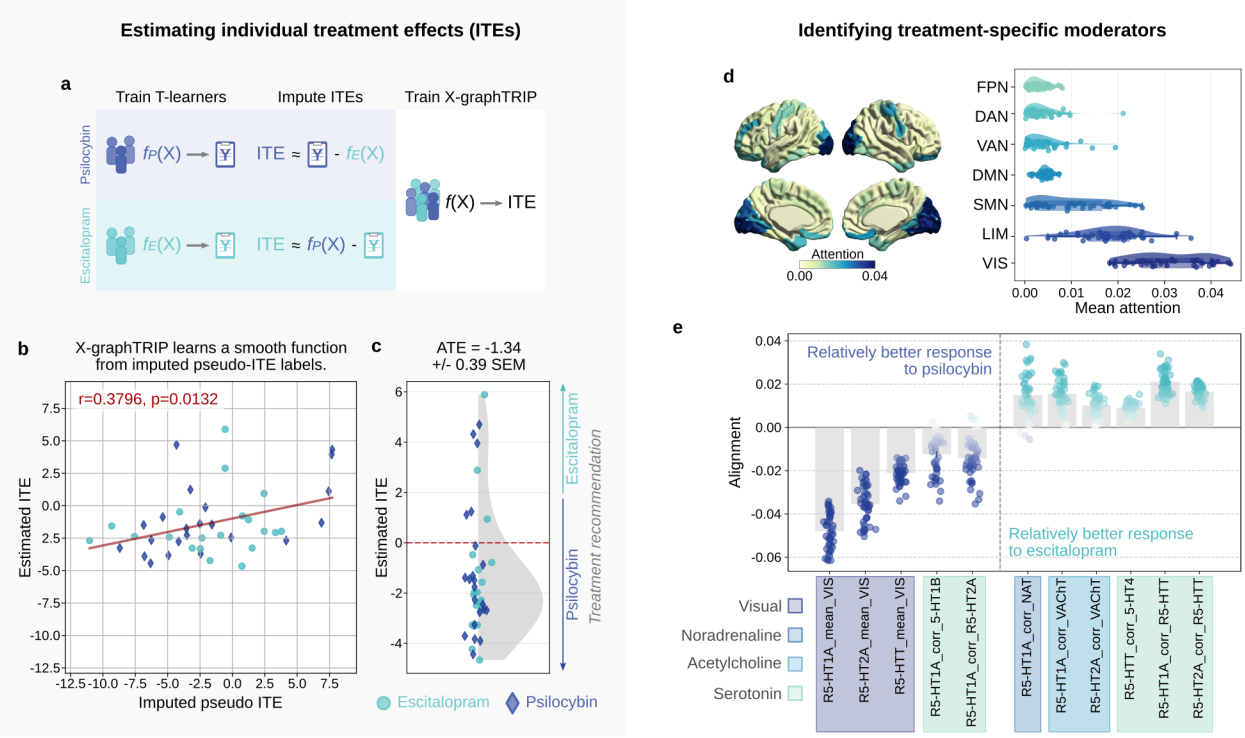


FIG. 6. **Causal inference via `X-graphTRIP` predicts individual treatment effects (ITEs) and reveals treatment-specific moderators.** **a**, Schematic of the X-learner framework [34], used to extend `graphTRIP` with causal inference, resulting in `X-graphTRIP`. **b**, `X-graphTRIP` learns a smooth ITE function, with predicted ITEs significantly correlated with pseudo-labels. **c**, Most patients are predicted to benefit more from psilocybin than escitalopram, with an average treatment effect (ATE) of -1.34 ± 0.39 SEM. **d**, Performance-weighted mean attention weights are highest in VIS and LIM regions. **e**, Performance-weighted mean alignments of robust biomarkers for all patients. Models that learn these biomarkers generalise significantly better, suggesting they reflect true treatment-specific moderators.

fies causal identifiability assumptions [46].

`X-graphTRIP` successfully learned a smooth ITE function that significantly correlated with the pseudo-labels ($r = 0.38$, $p = 0.0132$; Fig. 6b). On average, psilocybin was predicted to yield better outcomes than escitalopram, with an estimated average treatment effect (ATE) of -1.34 ± 0.39 (SEM; Fig. 6c). However, several patients were predicted to respond better to escitalopram, emphasising the importance of personalised treatment.

To identify neural biomarkers associated with treatment effect, we applied our attention-weight and GRAIL analyses to `X-graphTRIP`. However, while the resulting interpretability patterns learned by most CV-fold models were broadly consistent, some variability remained across folds (Supp. Fig. 16), suggesting that the underlying signal may be weak, noisy, or distributed across many overlapping features – a common challenge in low-sample causal inference settings with imputed labels. To isolate consistently informative features, we developed a robustness analysis based on the assumption that meaningful patterns should improve model generalisation. We trained a large set of CV-fold models with varying train-test splits and focused our analyses on the performance-weighted average attention weights and gradient alignments from models with significant test performance (defined by the Spearman correlation between ITE predictions and pseudo-labels). For GRAIL, we additionally applied a conservative filtering procedure based on partial least squares (PLS) regression to identify biomarkers reliably associated with generalisation across folds and patients (see Methods).

Models with significant generalisation performance attended most strongly to VIS regions, as well as LIM areas, particularly the orbitofrontal cortex and temporal poles (Fig. 6d). GRAIL analysis further identified several treatment-specific moderators (Fig. 6e). Relatively better response to psilocybin was linked to increased serotonin REACT maps in VIS regions and stronger functional coupling between 5-HT1A and 5-HT2A signalling. In contrast, relatively better response to escitalopram was associated with enhanced serotonin signalling in NAT- and VAChT-rich regions, and stronger 5-HT1A–5-HTT coupling. Please note that these relative biomarker effects should be interpreted in the context of an overall group-level advantage for psilocybin.

DISCUSSION

Predicting individual response to antidepressant treatment remains a major challenge in psychiatry, with implications for developing more effective, personalised interventions. Here, we introduced `graphTRIP`, a geometric deep learning (GDL) model that predicts post-treatment depression severity from pre-treatment neuroimaging and clinical data. Our model demonstrated strong prediction performance, generalised across brain parcellations and datasets, and provided insights into treatment responsiveness through rich interpretability analyses. Moreover, we delevoped `X-graphTRIP`, a causal inference extension of `graphTRIP`, to estimate individual treatment effects and identify treatment-specific moderators.

`X-graphTRIP` estimated an average treatment effect favouring psilocybin over escitalopram at the primary endpoint of three weeks post-treatment. This finding is consistent with previous analyses of the same dataset [35] and several other studies [10]. However, some patients were predicted to respond better to escitalopram. This underscores the importance of personalised treatment, and raises the question of what neurobiological mechanisms drive differential responsiveness to each drug.

Shared mechanisms of treatment responsiveness

Our analysis identified a number of brain-graph biomarkers that were consistently associated with treatment outcome across both treatments, possibly pointing to shared mechanisms of responsiveness. Specifically, better response was linked to reduced functional coupling within serotonin systems, and stronger coupling between these systems and sensory-motor regions. These findings were corroborated by attention weight analysis, which assigned significantly higher importance to VIS and SMN regions across both treatment arms. This pattern suggests that enhanced cross-talk between serotonin-rich cortices and serotonin-sparse sensory-motor regions may facilitate therapeutic effects, potentially by supporting broader integration of serotonin signalling.

Notably, a recent study showed that the predictive value of RSN connectivity varies across post-treatment time points [25]. At three weeks – the primary endpoint in our dataset – VIS connectivity was the strongest predictor, followed by auditory and SMN connectivity. In contrast, DMN and FPN connectivity performed at chance level. This suggests that the relevance of sensory-motor networks may be specific to early treatment response. Future work could extend `graphTRIP` to predict long-term treatment outcomes.

Treatment-specific moderators

While `graphTRIP` highlighted the general predictive role of VIS regions, `X-graphTRIP` further refined this picture by identifying serotonergic signalling in this network as a key moderator of psilocybin response. The mechanisms underlying psilocybin's antidepressant effects remain unclear. However, one of the more robust findings is that the acute subjective experience – particularly so-called "mystical-type experiences" – is predictive of long-term clinical benefit across conditions such as depression, end-of-life anxiety, and addiction [47–51]. Although 5-HT2A receptors are most densely expressed in association cortices such as the DMN and FPN, parts of the VIS also show high 5-HT2A density and have been implicated in the vivid visual phenomena characteristic of the acute psychedelic state [52–54]. In this context, our results may suggest that stronger VIS–serotonin coupling predisposes patients to a more intense psychedelic experience, potentially amplifying its therapeutic impact.

Furthermore, `X-graphTRIP` also identified stronger correlation between R5-HT1A and R5-HT2A REACT maps as a biomarker of relatively better response to psilocybin. However, `graphTRIP` linked greater 5-HT1A–5-HT2A coupling to poorer treatment outcomes overall. Together, these results imply that the biomarker likely reflects reduced responsiveness to escitalopram, rather than enhanced responsiveness to psilocybin. A plausible explanation for this pattern is a disrupted neuromodulatory state in which inhibitory control via 5-HT1A is compromised, leading to heightened excitability and co-activation of 5-HT1A- and 5-HT2A-rich cortices. This interpretation aligns with evidence that under chronic stress, inhibitory 5-HT1A signaling may downregulate while the sensitivity of excitatory 5-HT2A receptors increases [55–57]. Since 5-HT2A receptors are predominantly expressed by excitatory pyramidal neurons in transmodal cortices – most of which also express 5-HT1A [58] – this shift in the balance between 5-HT1A and 5-HT2A signalling could increase cortical excitability in response to serotonin. Notably, while the mechanisms of SSRIs like escitalopram remain incompletely understood, a leading hypothesis is that their antidepressant effects are facilitated by elevated cortical serotonin levels that suppress pyramidal cell firing via 5-HT1A-mediated inhibition [4, 59]. In this context, our findings may support the idea that intact 5-HT1A-mediated inhibition is critical for SSRI efficacy.

Finally, `X-graphTRIP` identified greater coupling between serotonin systems and other neuromodulatory systems – particularly the cholinergic and noradrenergic systems – as predictors of relatively better response to escitalopram. Interestingly, in `graphTRIP` we observed a trend whereby serotonin–VAChT coupling predicted better outcomes under escitalopram, but worse outcomes under psilocybin, suggesting that this biomarker has opposing predictive effects for the two drugs.

One plausible conjecture is that greater serotonin-system coupling with VAChT- and NAT-rich regions reflects elevated cholinergic and noradrenergic tone, indicative of a more reactive stress system. Both systems are tightly linked to stress responsivity in depres-

sion. The noradrenergic system, via the locus coeruleus, directly activates the hypothalamic–pituitary–adrenal (HPA) axis – a neuroendocrine pathway frequently dysregulated in MDD [60]. Meanwhile, central cholinergic (ACh) tone has been associated with depressive and anxiety symptoms [61]. For instance, pharmacological inhibition of ACh breakdown can induce depressive-like behaviour in both animals and humans [62, 63]. Conversely, VAChT knockdown mice, exhibiting reduced cholinergic tone, show attenuated depressive-like behaviour and elevated striatal serotonin levels [64]. Several SSRIs are also known to inhibit nicotinic acetylcholine receptors (nAChRs), and nicotinic modulators have been shown to enhance SSRI efficacy [61]. In this context, elevated serotonin-system coupling with NAT- and VAChT-rich regions may reflect an overactive stress-related neuromodulatory state that escitalopram can stabilise more effectively than psilocybin.

This hypothesis aligns with a theoretical model that proposes distinct stress-coping mechanisms for SSRIs and psychedelics [59]. According to this framework, SSRIs promote emotional blunting and stress buffering via post-synaptic 5-HT1A-mediated inhibition, whereas psychedelics foster emotional openness and psychological flexibility via 5-HT2A-mediated cortical excitation and enhanced synaptic plasticity. Individuals with heightened stress reactivity may benefit more from SSRIs, as psychedelics could be destabilising or counterproductive. Viewed through this lens, our results suggest that cholinergic and noradrenergic tone may serve as a moderator of antidepressant efficacy, influencing the extent to which patients benefit from stress-buffering versus emotionally amplifying treatment strategies.

Limitations and future work

One potential source of confusion lies in the use of two separate models to predict treatment outcome (`graphTRIP`) and causal treatment effects (`X-graphTRIP`). This distinction arises from fundamental modeling limitations: since treatment assignment is partially predictable from pre-treatment data, plain `graphTRIP` violates the identifiability assumption required for causal inference [34]. Conceptually, outcome prediction involves learning the two main effects of treatment and brain-graph features, whereas ITE estimation requires capturing the *interaction* between treatment and brain-graph features – that is, treatment-specific moderators. Our results suggest that `graphTRIP` captures the main effects well: after controlling for treatment, the correlation between true and predicted outcomes was slightly reduced but remained significant, indicating that both treatment and brain-graph effects were learned. However, `graphTRIP` likely captures treatment-specific interactions only partially. Nevertheless, `graphTRIP` remains clinically valuable: it identifies robust predictors of overall treatment response, highlighting substantial shared mechanisms underlying responsiveness to both escitalopram and psilocybin. Moreover, its predictions can inform decisions about care intensity, safety considerations, and treatment escalation. Future work should explore unified causal inference frameworks, such as CFR-Net [46] or Dragonnet [65], which can – in principle – estimate both outcome and treatment effects within a single model. The `graphTRIP` architecture can be readily integrated into such frameworks, similar to how we integrated it into the X-learner framework [34]. While these alternative frameworks tend to be more data-hungry, they offer a promising path forward for future studies with larger datasets.

The GRAIL method also has certain constraints. First, it evaluates individual biomarkers, while treatment response likely depends on complex interactions. Second, GRAIL does not account for redundancy among correlated biomarkers; when two candidate biomarkers are highly correlated, both may show similar alignment patterns, but the method does not indicate which one is more explanatory. This limitation is particularly relevant given the substantial correlations among several normative molecular target maps used in our analysis (Supp. Fig. 17b). Finally, GRAIL identifies statistical relationships, not causal mechanisms. An exciting future direction involves integrating brain connectivity and molecular chemoarchitecture with generative whole-brain models – biophysically informed simulations of brain activity [66–71]. By using `graphTRIP` to predict treatment response from simulated brain dynamics and backpropagating through the whole-brain model, we could systematically screen for plausible neurophysiological mechanisms underlying treatment response.

Finally, we used normative molecular target maps derived from independent PET datasets, acquired in healthy individuals. While this is standard practice in molecular-enriched neuroimaging analyses [19], it is important to note that these maps were obtained from independent cohorts using distinct imaging methods. Although all data were transformed to standard space, such differences may introduce variability in resolution. Nonetheless, extensive work has demonstrated that this approach yields valid and clinically informative insights [18, 19, 22–24].

CONCLUSION

In this work, we introduced `graphTRIP`, a geometric deep learning approach for predicting antidepressant treatment outcomes, offering both high predictive accuracy and interpretability. `graphTRIP` overcomes key limitations of conventional ML approaches, generalises across brain atlases, and enables patient-specific biomarker discovery. Our analysis highlights the role of global serotonergic integration with serotonin-sparse regions of the sensory-motor networks in treatment responsiveness. Additionally, we estimated causal treatment

effects and identified drug-specific response biomarkers, particularly involving serotonin-system interactions with stress-regulating neuromodulatory systems. Future work could scale our approach to larger datasets and integrate whole-brain models to move beyond biomarker-based to mechanistic explanations of treatment responsiveness. Overall, these advances mark a step toward data-driven, personalised treatment selection, bringing ML models closer to clinical translation.

METHODS

Datasets

We used data from two clinical trials conducted at the Imperial Clinical Research Facility and approved by relevant UK regulatory bodies. All participants provided written informed consent. The main dataset consisted of a double-blind randomised controlled trial (DB-RCT, clinicaltrials.gov: NCT03429075) comparing psilocybin and escitalopram for the treatment of major depressive disorder (MDD). An additional validation dataset was derived from an open-label trial of psilocybin treatment (gtr.ukri.org: MR/J00460X/1). Detailed trial protocols and clinical outcomes have been previously published [72].

Participants

Eligible participants in both trials were diagnosed with unipolar MDD (Hamilton Depression Rating Scale score ≥ 16). The DB-RCT included 59 patients, randomly assigned to either psilocybin (n = 30) or escitalopram (n = 29) treatment. The final imaging sample comprised 22 patients in the psilocybin arm (mean age = 44.5 ± 11.0 years; 8 female) and 20 in the escitalopram arm (mean age = 40.9 ± 10.1 years; 6 female). The open-label trial included 19 patients with treatment-resistant depression (TRD), defined as having failed to respond to multiple courses of antidepressant treatment (mean = 4.6 ± 2.6 past medications). Out of the initial sample, 16 patients were retained for analysis after excluding three due to excessive head motion (mean age = 42.75 ± 10.15 years; 4 female). Exclusion criteria for both trials included a history of psychosis, significant medical conditions, serious suicide attempts, pregnancy, and MRI contraindications. The DB-RCT additionally excluded patients with contraindications for SSRIs or prior escitalopram use.

Treatments

In both trials, all participants underwent a pre-treatment baseline session, involving clinical assessment and resting-state fMRI. The psilocybin arm received a 25 mg dose on Day 1 followed by a second identical dose 3 weeks later, along with daily placebo capsules for six weeks. The escitalopram arm received a negligible 1 mg psilocybin dose on Day 1, followed by 10 mg of escitalopram daily for the first 3 weeks, increased to 20 mg daily thereafter. In the open-label trial, patients received two psilocybin doses (10 mg and 25 mg, one week apart).

Clinical outcome measures

Depression severity was assessed using both the Quick Inventory of Depressive Symptomatology (QIDS) and the Beck Depression Inventory (BDI-1A) in both trials, at baseline and post-treatment. In the DB-RCT, QIDS was the primary outcome measure, with post-treatment assessments conducted 3 weeks after dosing day 2. BDI served as a secondary outcome measure in this trial. Thus, we used QIDS as the default prediction target of our model. The open-label trial used BDI as the primary outcome measure, with post-treatment assessments occurring 1 week after dosing day 2. In the open-label dataset, additional pre-treatment scores from the HAMD and LOT-R scales were available, which we included as input features when training models (simple MLP or `graphTRIP`) from scratch without pretraining.

fMRI data acquisition and preprocessing

Resting-state fMRI data were acquired using a 3T Siemens Tim Trio scanner with T2*-weighted echo-planar imaging. In the DB-RCT, scans included 480 volumes in approximately 10 min (TR = 1,250 ms; TE = 30 ms; 44 axial slices; spatial resolution = 3 mm isotropic; flip angle = 70 degrees; bandwidth = 2,232 Hz per pixel; and GRAPPA acceleration = 2). In the open-label trial, scans included 280 volumes in approximately 8 min (TR = 2,000 ms; TE = 31 ms; 36 axial slices; flip angle = 80 degrees; bandwidth = 2,298 Hz per pixel; and GRAPPA acceleration = 2). We used the Schaefer et al. [36] brain atlas with 100 parcels as the default atlas, and also conducted analyses using the Schaefer atlas with 200 parcels and the Automated Anatomical Labeling (AAL) atlas [37]. Preprocessing was performed using a custom pipeline integrating FSL, AFNI, Freesurfer, and ANTs, following standard procedures: motion correction, spatial smoothing, band-pass filtering (0.01–0.08 Hz), and nuisance regression. Volumes with $> 20\%$ framewise displacement > 0.5 mm were excluded. Full details of the preprocessing steps can be found in [72].

Functional connectivity

BOLD time series were parcellated and z-scored using the Schaefer atlases with 100 or 200 parcels [36], and the AAL atlas [37]. Functional connectivity (FC) was

then computed as the Pearson correlation coefficient between the z-scored time series of each pair of brain regions, resulting in an $N \times N$ FC matrix for each participant and parcellation. FC matrices were thresholded to retain only values with $|r| > 0.5$, which served as edge attributes in the main model of our study. This threshold was chosen to yield network densities comparable to typical structural connectomes (mean density $\rho = 0.16$, SD $= 0.05$). The atlas-bound model, in contrast, used the non-thresholded FC matrices as input.

Molecular target maps

Positron emission tomography (PET) maps of regional neuroreceptor and transporter densities were obtained from publicly available datasets. Serotonin receptor (5-HT1A, 5-HT2A) and serotonin transporter (5-HTT) maps were downloaded from Beliveau et al. [42] (https://nru.dk/index.php/allcategories/category/90-nru-serotonin-atlas-and-clustering). The maps were resampled to 2 mm MNI152 space, cerebellar voxels were excluded, and density values were normalised to a range of 0–1.

Additionally, we used PET maps for seven other molecular targets to explain the regional attention weight patterns and compute interpretable biomarkers in the gradient alignment analysis. Specifically, we included density maps for serotonin receptors (5-HT4, 5-HT1B), dopamine receptors (D1, D2) and transporter (DAT), noradrenaline transporter (NAT), and vesicular acetylcholine transporter (VAChT). These data were obtained from Hansen et al. [20].

REACT maps

The REACT maps, which served as the node attributes of our main model, were computed for each participant using the voxelwise PET maps of 5-HT1A, 5-HT2A, and 5-HTT densities from Beliveau et al. [42]. The analysis was performed with the `react-fmri` Python toolbox [18], following the protocol described in Lawn et al. [19].

Specifically, we used the `react_masks` command to generate masks that ensured that all voxels included in the subsequent analysis had valid PET data for all molecular targets, BOLD data for all participants, and were located within grey matter. Subsequently, we computed the REACT maps with the `react` command, which involves two sequential linear regressions for each voxel and molecular target. First, a voxel-level spatial regression of each molecular density map (5-HT1A, 5-HT2A, and 5-HTT) against the BOLD values at each time point yields a time series capturing the dominant fluctuations within each molecular system over time. In the second step, these molecular time series are regressed against the BOLD time series of each voxel, producing estimates of the coupling between each voxel's activity and the broader activity associated with each molecular system. The resulting voxelwise REACT time series were parcellated to obtain one time series for each brain region, participant, and atlas.

Model and training configurations

Full hyperparameter configurations, including architectural details and training parameters are provided in Supplementary Tables II-VII. These cover all models evaluated in this study. Addtionally, the complete codebase and all configuration files are available at: https://github.com/Imperial-MIND-lab/graphTRIP.

`graphTRIP` model

The main model in our study, called `graphTRIP` model, consists of a variational graph autoencoder (VGAE) with graph attention layers and a downstream multi-layer perceptron (MLP). The VGAE is trained to learn informative latent node representations by reconstructing the original input graph, while the MLP predicts post-treatment depression severity based on these latent representations and additional clinical data.

VGAE inputs. The input to the VGAE is a brain graph, where nodes represent brain regions defined by a brain atlas and edges represent FC between regions. Each node has a 6-dimensional feature vector: three elements encode the REACT values for serotonin receptors 5-HT1A, 5-HT2A, and the serotonin transporter 5-HTT, and the remaining three encode the 3D spatial coordinates of the brain region in MNI space. Edges exist between regions if the absolute FC value exceeds 0.5, and edge attributes are the original (non-thresholded) FC values. Self-connections were set to 1.

VGAE architecture. The VGAE consists of an encoder, a decoder, and a readout layer. The encoder consists of six GATv2Conv layers [73] with skip connections between consecutive layers of consistent width, followed by layer normalisation and dropout after each hidden layer. A final dense linear layer outputs two vectors per node: the latent means μ and log variances $\log \sigma^2$. Node features are updated iteratively in each GATv2Conv layer as

$$x_i' = x_i + \sum_{j \in \mathcal{N}(i) \cup \{i\}} \alpha_{i,j} W_t x_j \ . \tag{1}$$

Here, $\mathcal{N}(i)$ denotes the set of neighbors of node i, $\alpha_{i,j}$ are the attention coefficients, and W_t is a learnable weight matrix. The term x_i+ implements the skip connection, which is used in all layers with consistent input and output dimensions – specifically, between all layers except the transition from layer 0 to layer 1. The attention coefficients $\alpha_{i,j}$ – not to be confused with the attention

weights of the VGAE readout – are computed as

$$\alpha_{i,j} = \frac{\exp\left(a^T \varphi(W_s x_i + W_t x_j + W_e e_{i,j})\right)}{\sum_{k \in \mathcal{N}(i) \cup \{i\}} \exp\left(a^T \varphi(W_s x_i + W_t x_k + W_e e_{i,k})\right)} , \tag{2}$$

where φ is the LeakyReLU activation function, W_s, W_t, and W_e are learnable weight matrices for source nodes, target nodes, and edge attributes, respectively, and a^T is a learnable vector determining the relevance of features for attention.

Latent space sampling. Latent node representations z_i are sampled as

$$z_i = \mu_i + \sigma_i \odot \epsilon , \tag{3}$$

where $\epsilon \sim \mathcal{N}(0, I)$ and $\sigma_i = \exp(0.5 \cdot \log \sigma^2)$. These samples serve as input to the decoders and the readout.

Decoders. The VGAE has three decoders:

- **Node decoder:** A 3-layer MLP with LeakyReLU activation and dropout, and a linear output layer to reconstruct node features from z_i. Note: only REACT node features were reconstructed, as the 3D spatial coordinates of brain regions are normative and do not contain patient-specific information.
- **Edge index decoder:** A 3-layer MLP with LeakyReLU activation and dropout, and a sigmoid output to predict the probability of edges between nodes.
- **Edge attribute decoder:** A 3-layer MLP with LeakyReLU activation and dropout, and a tanh output to predict FC values in the range $[-1, 1]$.

The reconstruction of the thresholded FC matrix is obtained by elementwise multiplication of the edge-index and edge-attribute decoder outputs.

Readout layer. The VGAE includes an attention-based mean pooling layer to generate a single graph representation vector z from the latent node vectors z_i,

$$z = \frac{1}{N} \sum_i w_i z_i , \tag{4}$$

where N is the number of nodes, and w_i are attention weights computed as

$$s_i = \Phi(z_i \| d) , \tag{5}$$

$$\tilde{s}_i = s_i - \max_j s_j , \tag{6}$$

$$w_i = \frac{\exp(\tilde{s}_i)}{\sum_{j=1}^{n} \exp(\tilde{s}_j)} . \tag{7}$$

Here, d is a scalar encoding the drug condition (1 for psilocybin, -1 for escitalopram) and Φ is a 1-layer MLP with ReLU activation. The subtraction of the maximum score in Equation (6) implements the log-sum-exp trick, a standard technique to improve the numerical stability of the softmax computation by preventing potential overflow in the exponentials.

MLP for prediction. The input to the downstream MLP consists of a concatenation of the graph representation vector z and clinical data: baseline BDI and QIDS scores (pre-treatment depression scores), a binary indicator of prior SSRI use (0 for no, 1 for yes), and the drug condition d (-1 for escitalopram, 1 for psilocybin). The MLP has four layers with LeakyReLU activation and dropout in the hidden layers, and a linear output layer to predict post-treatment depression scores (QIDS).

Atlas-bound model

In contrast to the `graphTRIP` model, the atlas-bound model employs a variational autoencoder (VAE) and learns a latent representation of graphs rather than nodes. This model processes the full un-thresholded functional connectivity (FC) matrix, making it dependent on a specific brain atlas, as the dimension of the FC matrix determines the size of the input feature vectors.

VAE inputs. Each node's feature vector consists of all FC edges connected to that node, along with an integer encoding of the brain region's identity. This model does not use REACT features. The FC matrix is not thresholded, allowing the model to leverage the full connectivity information.

VAE architecture. The encoder consists of a 4-layer MLP with LeakyReLU activation and dropout, shared across nodes, which transforms each node's feature vector into a lower-dimensional node embedding vector. The node embeddings for all nodes in a graph are concatenated to form a single graph-level feature vector. This vector is then passed through a dense linear layer that outputs two vectors: the latent means μ and log variances $\log \sigma^2$. Latent graph representation vectors z are sampled from the latent distribution analogously to Eq. (3).

Decoder. The decoder is a 3-layer MLP with LeakyReLU activation and dropout, designed to reconstruct the FC matrix. The first hidden layer increases the dimensionality stepwise from the latent dimension to match the number of upper triangular edges in the FC matrix. The output layer uses a tanh activation function, ensuring that predicted FC values lie within the range $[-1, 1]$.

Hybrid model

To construct the hybrid model, we first trained the `graphTRIP` and atlas-bound models separately on the primary dataset. We then froze the weights of their respective representation learning modules (VGAE and VAE) and trained a new MLP on the concatenated latent readouts from both models, combined with the pre-treatment clinical data and drug condition.

X-graphTRIP model

To estimate individual treatment effects (ITEs), we extended the graphTRIP architecture within a causal inference framework based on the X-learner [34]. The X-learner is a two-step approach: first, counterfactual outcomes are estimated using separate models trained on each treatment group (T-learners); second, pseudo-ITE labels are computed from these counterfactuals and used to train a new model (X-learner) that predicts ITEs from pre-treatment covariates (i.e., clinical and fMRI data).

T-learners. We trained two T-learners, each using the graphTRIP architecture with minor modifications: One model was trained exclusively on psilocybin-treated patients, the other on escitalopram-treated patients. To avoid encoding treatment information, the drug condition was excluded from the MLP input, and the VGAE readout was no longer conditioned on treatment. Instead, the readout employed a simplified attention-based mean pooling, in which attention weights were computed from latent node vectors using a linear scoring function (i.e., Φ in Eq. 7 was replaced by a single linear layer). Additionally, we also excluded pre-treatment QIDS from the MLP inputs due to its predictive association with treatment assignment. To account for this omission, the T-learners were trained to predict the change in QIDS from pre- to post-treatment instead of post-treatment QIDS.

Each T-learner was trained on all patients from its respective treatment group and evaluated on the patients from the other treatment group. We selected T-learners that achieved the highest correlation between observed outcomes and predicted counterfactuals during evaluation (Supp. Fig. 15). This selection was motivated by the expectation of substantial shared variance across treatment conditions, as suggested by prior results with the main graphTRIP model.

Computing pseudo-ITE labels. From the trained T-learners, we estimated ITEs for individual patients i by combining observed outcomes with counterfactual predictions as

$$\tilde{D}_i^1 = \Delta Y_i^1 - \hat{\mu}_0(X_i^1), \quad \tilde{D}_i^0 = \hat{\mu}_1(X_i^0) - \Delta Y_i^0 \ . \tag{8}$$

Here, $\tilde{D}_i^1$ is the estimated treatment effect for psilocybin-treated patients, and $\tilde{D}_i^0$ for escitalopram-treated patients. ΔY_i^t denotes the observed QIDS change under treatment t. Conversely, $\hat{\mu}_{t'}$ denotes the counterfactual outcome with treatment $t \neq t'$. These counterfactuals are generated with the T-learner that was trained on patients who actually received treatment t'. Finally, X_i^t are the pre-treatment covariates.

X-learner. Finally, we trained a new graphTRIP model, called X-graphTRIP, on patients from both treatment conditions to predict the pseudo-ITE labels from their pre-treatment covariates. The goal of the X-learner is to smooth the noisy pseudo-labels into a robust and generalisable estimate of the ITE function.

Treatment classifiers

To assess whether treatment assignment was predictable from pre-treatment data – which would indicate a violation of the unconfoundedness assumption necessary for causal inference [34, 65] – we trained classifiers to distinguish between psilocybin- and escitalopram-treated patients. First, we trained a graphTRIP-type VGAE on the full dataset using only pre-treatment clinical and fMRI data. The readout was not conditioned on treatment, and the downstream MLP was replaced with a logistic regression head (using the same architecture as the graphTRIP MLP, but with a sigmoid output). Second, to test whether the latent representations learned by X-graphTRIP encoded treatment information, we trained a new logistic regression head on the pre-trained, frozen VGAE outputs of X-graphTRIP.

Model training

All models were implemented in PyTorch and trained using gradient descent with the ADAM optimizer and a learning rate of 0.001. The VGAE (or VAE, in the case of the atlas-bound model) and the downstream MLP were trained jointly to optimise both reconstruction accuracy and QIDS prediction accuracy. The loss function was a weighted combination of the VGAE/VAE loss and the MLP loss,

$$\mathcal{L} = \alpha \mathcal{L}_{\text{VGAE}} + (1-\alpha)\mathcal{L}_{\text{MLP}} \ , \tag{9}$$

where α is a weighting factor, usually set to 0.5.

The loss for the VGAE consists of a reconstruction loss, a Kullback-Leibler (KL) divergence loss, and an L2 regularisation term,

$$\mathcal{L}_{\text{VGAE}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{reg}} \tag{10}$$

$$= \sum_{i=1}^{N} \|x_i - \hat{x}_i\|_2^2 + \sum_{(i,j)\in\mathcal{E}} \|e_{ij} - \hat{e}_{ij}\|_2^2 \tag{11}$$

$$+ \mathbb{E}_{q(z|X)}\left[\text{KL}(q(z|X)\|p(z))\right] \tag{12}$$

$$+ \lambda \|W_{\text{VGAE}}\|_2^2 \ , \tag{13}$$

where x_i and $\hat{x}_i$ are the input and reconstructed node features, respectively, e_{ij} and $\hat{e}_{ij}$ are the input and reconstructed edge attributes (FC values), N is the number of nodes, and $\mathcal{E}$ is the set of unique (i.e., upper triangular) edges. The KL divergence term minimises the difference between the learned latent distribution $\mathcal{N}(\mu, \sigma^2)$ and the prior $\mathcal{N}(0, I)$. The regularisation term $\lambda\|W_{\text{VGAE}}\|_2^2$ prevents overfitting.

The MLP loss combines the prediction error with an L2 regularisation term,

$$\mathcal{L}_{\text{MLP}} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{reg}} \tag{14}$$

$$= \sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2^2 + \lambda \|W_{\text{MLP}}\|_2^2 \ , \tag{15}$$

where y_i and $\hat{y}_i$ are the true and predicted QIDS scores for the i-th patient, respectively, n is the batch size, and $\lambda\|W_{\text{MLP}}\|_2^2$ is the L2 regularisation term for the MLP weights.

To improve robustness we augmented our data during training by sampling multiple latent vectors per patient and computing a loss for each sample. All models were trained for a fixed number of epochs and using k-fold cross-validation (CV). Unless stated otherwise, predictions and reconstructions for each patient were generated using the CV-fold model in which that patient was part of the test set – that is, the model had not seen the patient during training. Full training configurations for each model are provided in the supplementary tables.

Fine-tuning

To transfer the `graphTRIP` model to the validation dataset ($n = 16$), we first pre-trained the model on the main dataset using only clinical variables that were available in both datasets (baseline QIDS and BDI scores, and drug condition). After pre-training, the weights of the VGAE were frozen, and only the downstream MLP was fine-tuned on the validation dataset for 100 epochs.

The validation dataset differed from the main dataset in several aspects. For instance, it included patients with TRD rather than MDD, and the prediction target was the QIDS score one week post-treatment instead of three weeks. Given these differences, perfect transfer was not expected. However, we reasoned that a positive correlation between the pre-trained model's predictions and the true labels would indicate successful transfer.

Following this rationale, the total loss was defined as the combination of 1) mean squared error loss with L_2 regularisation, and 2) a correlation-based loss,

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda\|W\|^2 \ , \tag{16}$$

$$\mathcal{L}_{\text{corr}} = -\frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}} \ , \tag{17}$$

where n is the batch size, y_i and $\hat{y}_i$ are the true and predicted post-treatment QIDS scores, $\bar{y}$ and $\bar{\hat{y}}$ are their respective means, W are the MLP weights, and λ is the regularisation coefficient.

To balance the influence of both terms during optimisation, we normalised each loss component by its exponential running average. At training step t, the running averages $\mathcal{A}_{\text{MSE}}^{(t)}$ and $\mathcal{A}_{\text{corr}}^{(t)}$ were updated as

$$\mathcal{A}_{\text{MSE}}^{(t)} = \beta\mathcal{A}_{\text{MSE}}^{(t-1)} + (1-\beta)\mathcal{L}_{\text{MSE}}^{(t)} \ , \tag{18}$$

$$\mathcal{A}_{\text{corr}}^{(t)} = \beta\mathcal{A}_{\text{corr}}^{(t-1)} + (1-\beta)\mathcal{L}_{\text{corr}}^{(t)} \ , \tag{19}$$

where $\beta = 0.9$ is the leak rate. The normalised loss used for optimisation was

$$\mathcal{L}_{\text{MLP}}^{(t)} = \rho\frac{\mathcal{L}_{\text{corr}}^{(t)}}{\mathcal{A}_{\text{corr}}^{(t)}} + (1-\rho)\frac{\mathcal{L}_{\text{MSE}}^{(t)}}{\mathcal{A}_{\text{MSE}}^{(t)}} \ , \tag{20}$$

where ρ was set to 0.4. This dynamic normalisation ensured stable training despite the different scales of the correlation and MSE components.

The same learning rate as in pre-training (lr = 0.001) was used. Data augmentation was implemented by sampling three latent vectors per patient for each batch and summing the loss across samples.

Gradient Alignment for Interpreting Latent-variable models

The GRAIL method was developed to estimate the association between specific biomarkers and treatment response. Positive alignment indicates that higher values of a biomarker are associated with higher predicted post-treatment depression scores (i.e., lower treatment responsiveness), while negative alignment suggests the opposite. An alignment of zero indicates no association between the biomarker and treatment responsiveness.

This method is enabled by the two-part architecture of the `graphTRIP` model, which includes a VGAE for reconstruction and an MLP for regression. The VGAE encoder maps input brain graphs to latent representations z, which serve as inputs to 1) the VGAE readout and the MLP (predicting post-treatment QIDS scores, $\hat{y}$), and 2) the VGAE decoder (reconstructing brain graphs). Biomarkers of interest, such as the mean FC within a specific RSN, are computed from the reconstructed brain graphs. The only requirement for a biomarker to be tested is that it can be computed in a differentiable manner from the reconstructed graph.

By deriving both $\hat{y}$ and a biomarker b as differentiable functions of z, we can compute the gradients $\nabla_z\hat{y}$ and $\nabla_z b$. These gradients indicate the directions in the latent space that most strongly increase $\hat{y}$ and b, respectively. If the gradients point in similar directions, it implies that increasing the biomarker value would also increase the predicted QIDS score, indicating positive alignment. Conversely, opposite directions imply negative alignment.

To quantify alignment, we computed the cosine similarity between the normalised gradient vectors:

$$\text{Alignment}(b) = \frac{\nabla_z\hat{y}\cdot\nabla_z b}{\|\nabla_z\hat{y}\|\|\nabla_z b\|} \ . \tag{21}$$

The alignment score ranges from -1 (maximum negative alignment) to 1 (maximum positive alignment), with 0 indicating no alignment.

We computed the gradient alignment for a number of biomarkers as follows. For each patient, we sampled 100 latent vectors from a Gaussian distribution with a mean equal to the latent mean vector of the patient's original

brain graph and a standard deviation of 2. For each latent sample, we then:

1. Computed the predicted QIDS score $\hat{y}$ and derived the gradient $\nabla_z \hat{y}$.
2. Reconstructed the brain graph using the VGAE decoder and computed the biomarkers of interest.
3. Derived the gradient $\nabla_z b$ for each biomarker b.
4. Calculated the alignment score as the cosine similarity between $\nabla_z \hat{y}$ and $\nabla_z b$.

This approach enabled us to systematically assess the alignment of a wide range of biomarkers, providing an interpretable link between model predictions and biologically meaningful factors.

Permutation importance analysis

To quantify the contribution of each input feature to the treatment outcome predictions, we conducted permutation importance analysis. Individual features, including pre-treatment depression scores, prior SSRI use, treatment condition, and the latent brain-graph representation, were shuffled independently across patients in the test set, using 50 random permutations per feature. All other features were held constant. Feature importance was defined as the difference between the model's mean absolute error (MAE) on the permuted input and the MAE on the original data. Larger increases in MAE indicate greater predictive importance. For the latent brain representation, the entire vector $\boldsymbol{z}$ was permuted as a unit.

PLS-based Robustness Analysis

To identify robust biomarkers, we developed a new method based on partial least squares (PLS) regression. This method identifies gradient alignment patterns that correlate with model generalisation (test performance) across cross-validation (CV) folds, building on the idea that meaningful biomarkers improve test performance while spurious ones degrade it.

First, to enhance statistical power, we trained a larger number of models with overlapping test folds. Specifically, test-fold assignments were circularly shifted by one index $N = 41$ times until the original assignments were recovered. For `X-graphTRIP`, trained with 7-fold CV, this yielded $41 \times 7 = 287$ models. Each model's test performance was quantified as the Spearman correlation (ρ) between predicted and true labels on its respective test fold. Models with $\rho \leq 0.3$ were excluded to reduce noise.

For each model and patient, we computed the gradient alignments for all candidate biomarkers, resulting in a matrix per patient (rows: CV models; columns: candidate biomarkers). We then performed a PLS regression for each patient to extract the direction (first PLS component) in biomarker space that most strongly correlated with model performance across folds. To enable group-level interpretation, we aligned the sign of each patient's PLS component (multiplying by -1 if the Pearson correlation between component scores and model performance was negative) and stacked the aligned components across patients.

Finally, due to the signed nature of gradient alignment values, additional filtering was required to resolve ambiguity where a positive PLS weight could reflect either a beneficial or detrimental alignment. To identify robust, performance-enhancing biomarkers, we implemented a conservative three-step filtering procedure:

1. **Consistency of Influence:** We performed a one-sample t-test across patients on the horizontally stacked PLS components for each biomarker. Biomarkers with FDR-corrected $p < 0.05$ were retained, indicating a consistent association between their PLS-component weights and generalisation performance.
2. **Directionality of Benefit:** For each biomarker, we compared the sign of its average PLS weight with the sign of its weighted mean gradient alignment (computed by averaging CV-fold alignment vectors, weighted by each model's test performance). Only biomarkers with matching signs were retained, ensuring they positively contributed to generalisation.
3. **Necessity for Performance:** We performed a one-sample t-test across patients on the weighted mean gradient alignment values for each biomarker. Biomarkers were retained if their alignment values were consistently non-zero (FDR-corrected $p < 0.05$), indicating they were reliably learned by generalising models and were necessary for strong predictive performance.

DATA AVAILABILITY

All requests for raw or processed data and study materials will be reviewed by R.L.C.-H., the chief investigator of both clinical trials from which the datasets were obtained. The receptor density maps from in vivo PET are available at https://github.com/netneurolab/hansen_receptors.

CODE AVAILABILITY

The code used to implement and run all analyses in this study is publicly available at: https://github.com/Imperial-MIND-lab/graphTRIP

ACKNOWLEDGMENTS

H.M.T. is supported by the Doctoral Teaching Scholarship of the Department of Computing, Imperial College London. A.I.L. acknowledges support from St John's College, Cambridge; and a Wellcome Early Career Award (grant number 226924/Z/23/Z). We are grateful to Lewis J. Ng, whose prior work on predictive modelling for this dataset provided valuable context and inspiration for this project. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC-BY) license to any Author Accepted Manuscript version arising from this submission.

[1] E. J. Bromet, L. H. Andrade, I. H. Hwang, N. A. Sampson, J. Alonso, G. de Girolamo, R. de Graaf, K. Demyttenaere, C. Hu, N. Iwata, A. N. Karam, J. Kaur, S. Kostyuchenko, J. P. Lépine, D. Levinson, H. Matschinger, M. E. M. Mora, M. A. O. Browne, J. A. Posada-Villa, M. C. Viana, D. R. Williams, and R. C. Kessler, BMC Medicine **9**, 90 (2011).

[2] S. Shorey, E. D. Ng, and C. H. J. Wong, The British journal of clinical psychology (2021).

[3] L. Cui, S. Li, S. Wang, X. Wu, Y. Liu, W. Yu, Y. Wang, Y. Tang, M. Xia, and B. Li, Signal Transduction and Targeted Therapy **9** (2024).

[4] T. Sharp and H. Collins, Current Topics in Behavioral Neurosciences , 21–47 (2023).

[5] A. Rush, American Journal of Psychiatry **163**, 1905 (2006).

[6] A. Cipriani, T. Furukawa, G. Salanti, J. R. Geddes, J. P. T. Higgins, R. Churchill, N. Watanabe, A. Nakagawa, I. M. Omori, H. McGuire, M. Tansella, and C. Barbui, The Lancet **373**, 746 (2009).

[7] S. Haikazian, D. C. Chen-Li, D. E. Johnson, F. Fancy, A. Levinta, M. I. Husain, R. B. Mansur, R. S. McIntyre, and J. D. Rosenblat, Psychiatry Research **329**, 115531 (2023).

[8] R. L. Carhart-Harris, L. Roseman, M. Bolstridge, L. Demetriou, J. N. Pannekoek, M. B. Wall, M. A. Tanner, M. Kaelen, J. McGonigle, K. Murphy, R. Leech, H. V. Curran, and D. J. Nutt, Scientific Reports **7** (2017).

[9] G. M. Goodwin, S. T. Aaronson, O. Alvarez, P. C. Arden, A. Baker, J. C. Bennett, C. I. V. Bird, R. E. Blom, C. Brennan, D. Brusch, L. Burke, K. Campbell-Coker, R. Carhart-Harris, J. Cattell, A. Daniel, C. Debattista, B. W. Dunlop, K. Eisen, D. Feifel, M. Forbes, H. M. Haumann, D. J. Hellerstein, A. I. Hoppe, M. I. Husain, L. A. Jelen, J. Kamphuis, J. Kawasaki, J. R. Kelly, R. E. Key, R. Kishon, S. K. Peck, G. Knight, M. H. B. Koolen, M. Lean, R. W. Licht, J. L. Maples-Keller, J. Mars, L. Marwood, M. C. McElhiney, T. Miller, A. Mirow, S. Mistry, T. Mletzko-Crowe, L. N. Modlin, R. E. Nielsen, E. M. Nielson, S. R. Offerhaus, V. O'Keane, T. Páleníček, D. J. Printz, M. C. Rademaker, A. van Reemst, F. Reinholdt, D. Repantis, J. J. H. Rucker, S. Rudow, S. Ruffell, A. J. Rush, R. A. Schoevers, M. Seynaeve, S. Shao, J. C. Soares, M. Somers, S. C. Stansfield, D. Sterling, A. Strockis, J. Tsai, L. Visser, M. M. Wahba, S. Williams, A. H. Young, P. Ywema, S. Zisook, and E. Malievskaia, The New England journal of medicine **387 18**, 1637 (2022).

[10] T. Watford and N. Masood, Clinical Psychopharmacology and Neuroscience **22**, 2–12 (2023).

[11] R. L. Carhart-Harris and G. M. Goodwin, Neuropsychopharmacology **42**, 2105–2113 (2017).

[12] C. Otte, S. M. Gold, B. W. Penninx, C. M. Pariante, A. Etkin, M. Fava, D. C. Mohr, and A. F. Schatzberg, Nature Reviews Disease Primers **2** (2016).

[13] F. Moujaes, K. H. Preller, J. L. Ji, J. D. Murray, L. Berkovitch, F. X. Vollenweider, and A. Anticevic, Biological Psychiatry **93**, 1061 (2023).

[14] E. T. Bullmore and O. Sporns, Nature Reviews Neuroscience **10**, 186 (2009).

[15] M. P. van den Heuvel and O. Sporns, Nature Reviews Neuroscience **20**, 435 (2019).

[16] Y. Liu, Y. Chen, X. Liang, D. Li, Y. Zheng, H. Zhang, Y. Cui, J. Chen, J. Liu, and S. Qiu, Frontiers in Neurology **11** (2020).

[17] R. H. Kaiser, J. R. Andrews-Hanna, T. D. Wager, and D. A. Pizzagalli, JAMA psychiatry **72 6**, 603 (2015).

[18] O. Dipasquale, P. Selvaggi, M. Veronese, A. S. Gabay, F. E. Turkheimer, and M. A. Mehta, Neuroimage **195**, 252 (2019).

[19] T. Lawn, M. A. Howard, F. E. Turkheimer, B. Mišić, G. Deco, D. F. D. Martins, and O. Dipasquale, Neuroscience and Biobehavioral Reviews **150**, 105193 (2023).

[20] J. Y. Hansen, G. Shafiei, R. D. Markello, K. Smart, S. M. L. Cox, Y. Wu, J.-D. Gallezot, É. Aumont, S. Servaes, S. G. Scala, J. M. DuBois, G. Wainstein, G. Bezgin, T. Funck, T. W. Schmitz, R. N. Spreng, J.-P. Soucy, S. Baillet, S. Guimond, J. Hietala, M.-A. Bedard, M. Leyton, E. Kobayashi, P. Rosa-Neto, N. Palomero-Gallagher, J. M. Shine, R. E. Carson, L. Tuominen, A. Dagher, and B. Mišić, Nature Neuroscience **25**, 1569 (2022).

[21] A. I. Luppi, J. Y. Hansen, R. Adapa, R. L. Carhart-Harris, L. Roseman, C. Timmermann, D. Golkowski, A. Ranft, R. Ilg, D. Jordan, *et al.*, Science advances **9**, eadf8332 (2023).

[22] T. Lawn, A. Giacomel, D. Martins, M. Veronese, M. Howard, F. E. Turkheimer, and O. Dipasquale, Communications Biology **7** (2024).

[23] D. Martins, M. Veronese, F. E. Turkheimer, M. A. Howard, S. C. Williams, and O. Dipasquale, Brain Communications **4** (2021), 10.1093/braincomms/fcab302.

[24] T. Lawn, O. Dipasquale, A. Vamvakas, I. Tsougos, M. A. Mehta, and M. A. Howard, Psychopharmacology **239**, 1797 (2022).

[25] D. Copa, D. Erritzoe, B. Giribaldi, D. Nutt, R. Carhart-Harris, and E. Tagliazucchi, Journal of Affective Disorders **353**, 60–69 (2024).

[26] A. Escrichs, Y. Sanz Perl, P. M. Fisher, N. Martínez-Molina, E. G-Guzman, V. G. Frokjaer, M. L. Kringelbach, G. M. Knudsen, and G. Deco, Molecular Psychiatry **30**, 1069–1079 (2024).

[27] G. Deco, Y. S. Perl, S. Johnson, N. Bourke, R. L. Carhart-Harris, and M. L. Kringelbach, Nature Mental Health (2024).

[28] A. M. Chekroud, R. Zotti, Z. Shehzad, R. Gueorguieva, M. K. Johnson, M. H. Trivedi, T. D. Cannon, J. H. Krystal, and P. R. Corlett, The lancet. Psychiatry **3 3**, 243 (2016).

[29] S. B. Eickhoff, B. T. T. Yeo, and S. Genon, Nature Reviews Neuroscience **19**, 672 (2018).

[30] A. I. Luppi, H. M. Gellersen, Z.-Q. Liu, A. R. Peattie, A. E. Manktelow, R. Adapa, A. M. Owen, L. Naci, D. K. Menon, S. I. Dimitriadis, *et al.*, Nature Communications **15**, 4745 (2024).

[31] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan, Transformer Circuits Thread (2024).

[32] J. Lindsey, W. Gurnee, E. Ameisen, B. Chen, A. Pearce, N. L. Turner, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. B. Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, and J. Batson, Transformer Circuits Thread (2025).

[33] K. Lekadir, A. F. Frangi, A. R. Porras, B. Glocker, C. Cintas, C. P. Langlotz, E. Weicken, F. W. Asselbergs, F. Prior, G. S. Collins, *et al.*, BMJ **388** (2025).

[34] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, Proceedings of the National Academy of Sciences of the United States of America **116**, 4156 (2019).

[35] R. L. Carhart-Harris, B. Giribaldi, R. Watts, M. Baker-Jones, A. Murphy-Beiner, R. Murphy, J. Martell, A. Blemings, D. Erritzoe, and D. J. Nutt, The New England journal of medicine **384 15**, 1402 (2021).

[36] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. T. Yeo, Cerebral Cortex (2018).

[37] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, NeuroImage **15**, 273 (2002).

[38] D. S. Margulies, S. S. Ghosh, S. S. Ghosh, A. Goulas, M. Falkiewicz, J. M. Huntenburg, G. Langs, G. Langs, G. Bezgin, S. B. Eickhoff, F. X. Castellanos, M. Petrides, E. Jefferies, and J. Smallwood, Proceedings of the National Academy of Sciences **113**, 12574 (2016).

[39] A. Goulas, J. P. Changeux, K. Wagstyl, K. Amunts, N. Palomero-Gallagher, and C. C. Hilgetag, Proceedings of the National Academy of Sciences of the United States of America **118** (2020).

[40] A. I. Luppi, P. A. M. Mediano, F. E. Rosas, N. Holland, T. D. Fryer, J. T. O'Brien, J. B. Rowe, D. K. Menon, D. Bor, and E. A. Stamatakis, Nature Neuroscience **25**, 771 (2022).

[41] V. J. Sydnor, B. Larsen, D. S. Bassett, A. Alexander-Bloch, D. A. Fair, C. Liston, A. P. Mackey, M. P. Milham, A. Pines, D. R. Roalf, *et al.*, Neuron **109**, 2820 (2021).

[42] V. Beliveau, M. Ganz, L. Feng, B. Ozenne, L. Højgaard, P. M. Fisher, C. Svarer, D. N. Greve, and G. M. Knudsen, The Journal of Neuroscience **37**, 120 (2017).

[43] F. Váša and B. Mišić, Nature Reviews Neuroscience **23**, 493 (2022).

[44] B. T. T. Yeo, F. M. Krienen, J. Sepulcre, J. Sepulcre, M. R. Sabuncu, M. R. Sabuncu, D. Lashkari, M. O. Hollinshead, M. O. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. Polimeni, B. Fischl, B. Fischl, H. Liu, and R. L. Buckner, Journal of neurophysiology **106 3**, 1125 (2011).

[45] A. I. Luppi, F. E. Rosas, P. A. Mediano, D. K. Menon, and E. A. Stamatakis, Trends in Cognitive Sciences **28**, 352 (2024).

[46] U. Shalit, F. D. Johansson, and D. A. Sontag, ArXiv **abs/1606.03976** (2017).

[47] R. R. Griffiths, M. W. Johnson, M. A. Carducci, A. Umbricht, W. A. Richards, B. D. Richards, M. P. Cosimano, and M. A. Klinedinst, Journal of Psychopharmacology (Oxford, England) **30**, 1181 (2016).

[48] L. Roseman, D. J. Nutt, and R. L. Carhart-Harris, Frontiers in Pharmacology **8** (2018).

[49] S. Ross, A. Bossis, J. Guss, G. Agin-Liebes, T. Malone, B. Cohen, S. E. Mennenga, A. Belser, K. Kalliontzi, J. Babb, and et al., Journal of Psychopharmacology **30**, 1165–1180 (2016).

[50] M. P. Bogenschutz, A. A. Forcehimes, J. A. Pommy, C. E. Wilcox, P. Barbosa, and R. J. Strassman, Journal of Psychopharmacology **29**, 289 (2015).

[51] M. W. Johnson, A. Garcia-Romeu, and R. R. Griffiths, The American Journal of Drug and Alcohol Abuse **43**, 55–60 (2016).

[52] D. E. Nichols, Pharmacological Reviews **68**, 264 (2016).

[53] G. Ballentine, S. F. Friedman, and D. Bzdok, Science Advances **8** (2021).

[54] K. H. Preller, J. B. Burt, J. L. Ji, C. H. Schleifer, B. D. Adkinson, P. Stämpfli, E. Seifritz, G. Repovs, J. H. Krystal, J. D. Murray, *et al.*, Elife **7**, e35082 (2018).

[55] O. Berton, S. Aguerre, A. Sarrieau, P. Mormède, and F. Chaouloff, Neuroscience **82**, 147 (1998).

[56] M. Benekareddy, N. M. Goodfellow, E. K. Lambe, and V. A. Vaidya, The Journal of Neuroscience **30**, 12138 (2010).

[57] J. F. Lopez, I. Liberzon, D. M. Vázquez, E. A. Young, and S. J. Watson, Biological Psychiatry **45**, 934 (1999).

[58] M. Amargós-Bosch, A. Bortolozzi, M. V. Puig, J. Serrats, A. Adell, P. Celada, M. Toth, G. Mengod, and F. Artigas, Cerebral cortex **14 3**, 281 (2004).

[59] R. Carhart-Harris and D. J. Nutt, Journal of Psychopharmacology (Oxford, England) **31**, 1091 (2017).

[60] C. Tsigos and G. P. Chrousos, Journal of psychosomatic research **53 4**, 865 (2002).

[61] H. R. Arias, K. M. Targowska-Duda, J. García-Colunga, and M. O. Ortells, Molecules **26** (2021).

[62] Y. S. Mineur, A. Obayemi, M. B. Wigestrand, G. M. Fote, C. A. Calarco, A. M. Li, and M. R. Picciotto, Proceedings of the National Academy of Sciences **110**, 3573–3578 (2013).

[63] S. C. Risch, R. M. Cohen, D. S. Janowsky, N. H. Kalin, and D. L. Murphy, Science **209 4464**, 1545 (1980).

[64] M. Pádua-Reis, N. S. S. Aquino, V. E. M. Oliveira, R. E. Szawka, M. A. M. Prado, V. F. Prado, and G. S. Pereira, Behavioural Brain Research **330**, 127 (2017).

[65] C. Shi, D. M. Blei, and V. Veitch, Neural Information Processing Systems (2019).

[66] R. Cofré, R. Herzog, P. A. M. Mediano, J. I. Piccinini, F. E. Rosas, Y. S. Perl, and E. Tagliazucchi, Brain Sciences **10** (2020).

[67] G. Deco, J. Cruzat, J. Cabral, G. M. Knudsen, R. L. Carhart-Harris, P. C. Whybrow, N. K. Logothetis, and M. L. Kringelbach, Current biology **28**, 3065 (2018).

[68] M. L. Kringelbach, J. Cruzat, J. Cabral, G. M. Knudsen, R. Carhart-Harris, P. C. Whybrow, N. K. Logothetis, and G. Deco, Proceedings of the National Academy of Sciences **117**, 9566 (2020).

[69] J. B. Burt, K. H. Preller, M. Demirtas, J. L. Ji, J. H. Krystal, F. X. Vollenweider, A. Anticevic, and J. D. Murray, Elife **10**, e69320 (2021).

[70] A. I. Luppi, S. P. Singleton, J. Y. Hansen, K. W. Jamison, D. Bzdok, A. Kuceyeski, R. F. Betzel, and B. Misic, Nature Biomedical Engineering **8**, 1142 (2024).

[71] A. I. Luppi, F. I. Milisav, L. E. Suarez, G. Shafiei, J. Vohryzek, Y. Sanz Perl, H. Ali, F. E. Rosas, P. A. Mediano, B. Misic, *et al.*, bioRxiv , 2025 (2025).

[72] R. E. Daws, C. B. Timmermann, B. Giribaldi, J. D. Sexton, M. B. Wall, D. Erritzoe, L. Roseman, D. J. Nutt, and R. L. Carhart-Harris, Nature Medicine **28**, 844 (2022).

[73] S. Brody, U. Alon, and E. Yahav, ArXiv **abs/2105.14491** (2021).

SUPPLEMENTARY

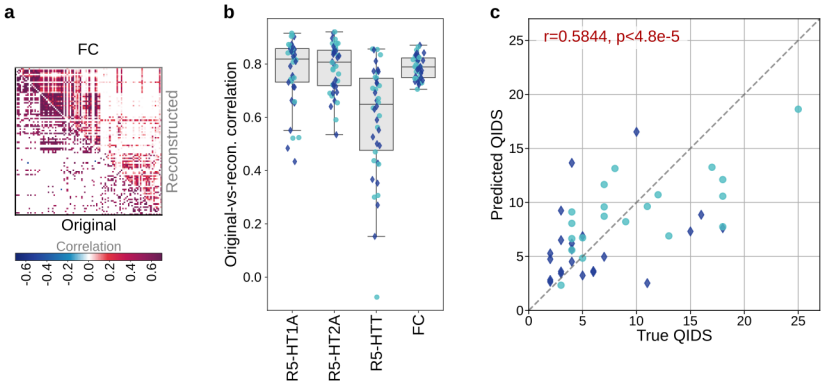


FIG. 7. `graphTRIP` **generalises across alternative edge definitions.** `graphTRIP` was trained on brain graphs constructed using a fixed functional connectivity (FC) threshold ($|\mathrm{FC}| > 0.5$), and tested on graphs generated by applying a fixed connection density threshold ($\rho = 0.20$). **a**, Original versus reconstructed FC of an example subject. **b**, Reconstruction accuracy of FC and node features across all patients. **c**, QIDS predictions remain significant, indicating robust generalisation.

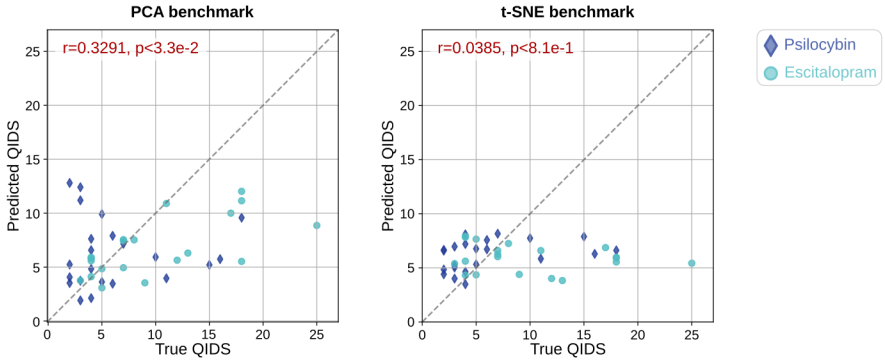


FIG. 8. **Benchmark models using dimensionality reduction of flattened brain-graph data.** True versus predicted post-treatment QIDS scores from MLPs trained on brain-graph data reduced via PCA (left) and t-SNE (right). In both cases, brain graphs were constructed identically to those used in `graphTRIP` training, except that FC matrices were not thresholded to ensure homogeneous input across patients. For each patient, node features and the upper triangle of the unthresholded FC matrix were flattened and concatenated into a single vector. The vectorised data of all patients was then reduced using either PCA (to 32 components) or t-SNE (perplexity = 30). Both MLPs were trained using 6-fold cross-validation.

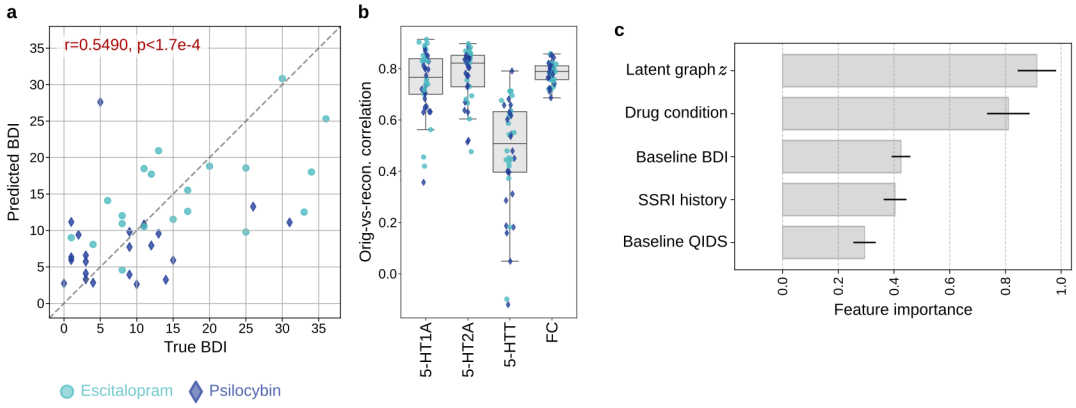


FIG. 9. **Predicting post-treatment BDI scores with `graphTRIP`.** A separate `graphTRIP` model was trained to predict post-treatment BDI scores, using the same configuration and training parameters as the main model predicting post-treatment QIDS. This analysis demonstrates the flexibility of our pipeline for alternative clinical outcomes. **a**, True versus predicted BDI scores show a significant correlation. **b**, Correlations between original and reconstructed edge and node features confirm accurate graph reconstruction by the VGAE. **c**, Permutation importance analysis reveals that latent brain-graph features are the most influential predictors of BDI scores. Bars indicate the mean increase in mean absolute error (MAE) across 50 random permutations when a given feature is shuffled across patients; error bars denote standard error.

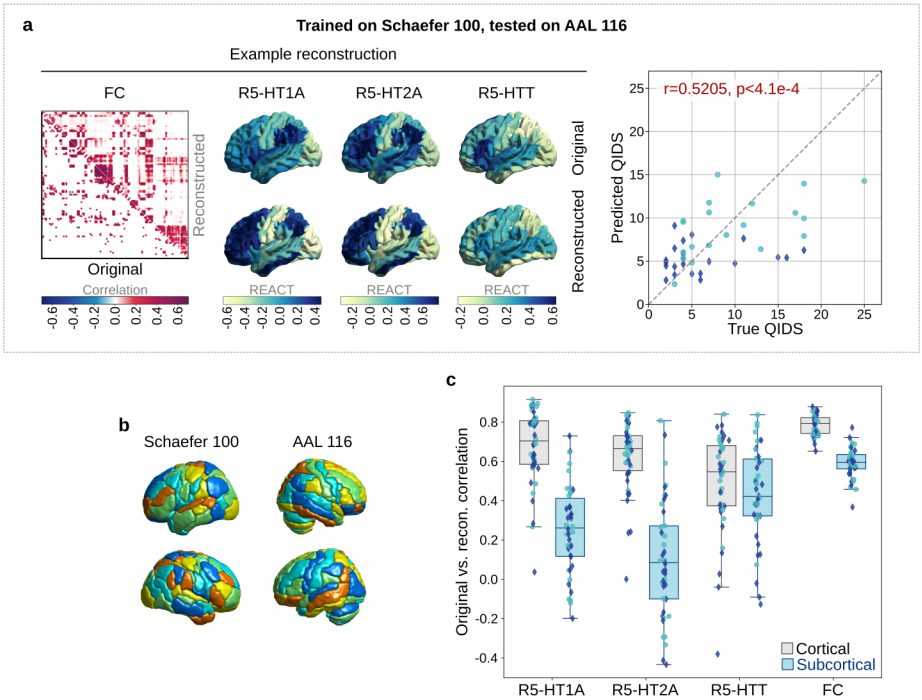


FIG. 10. `graphTRIP` **generalises to brain graphs defined using the AAL atlas.** **a**, `graphTRIP`, trained on brain graphs constructed with the Schaefer 100 atlas, generalises to graphs using the AAL atlas, maintaining strong reconstruction and prediction performance. The panel shows example reconstructions of FC edges (left) and node features (center), and QIDS prediction performance (right). **b**, Brain renderings of the Schaefer 100 and AAL 116 atlases, comprising 100 and 116 brain regions, respectively. **c**, Correlations between original and reconstructed edge and node features for all patients. The AAL atlas includes subcortical regions not present in the Schaefer atlas. To assess model performance separately for these, reconstruction correlations are shown for cortical and subcortical regions. As expected, reconstruction is slightly weaker for subcortical regions, which the model never encountered during training, but it remains largely significant.

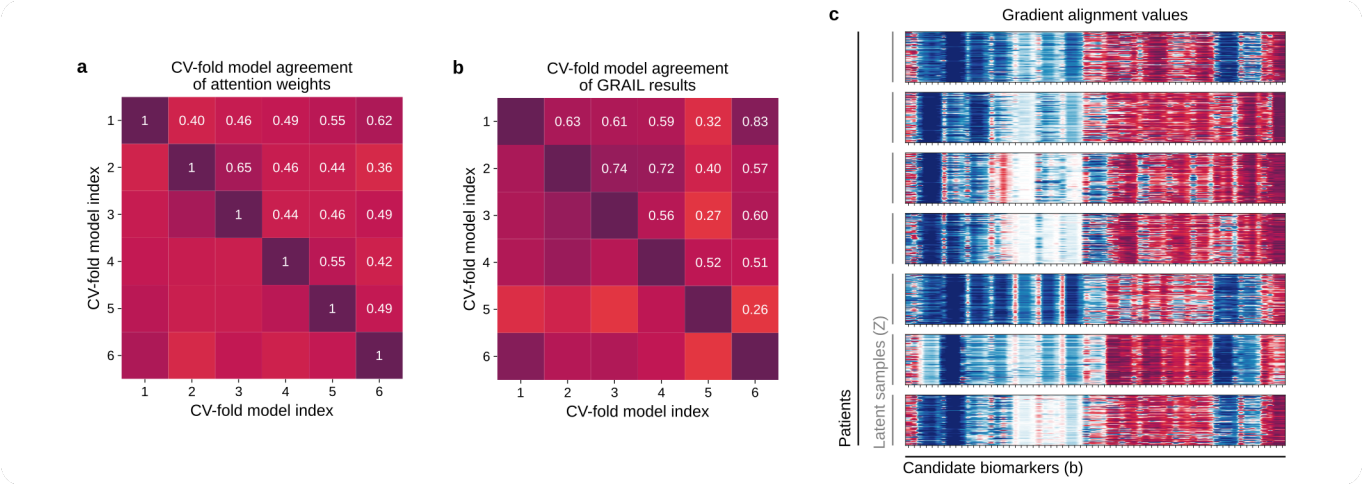


FIG. 11. **Cross-validation model agreement in attention and gradient alignment outputs.** `graphTRIP` was trained using 6-fold cross-validation (CV), resulting in six separate models. This allows us to compute attention weights and gradient alignment patterns for each patient with six independently trained models. **a**, Mean pairwise correlations of regional attention weights across CV-fold models. For each patient, attention weights from every pair of models were correlated. The heatmap shows the group-averaged correlation values across patients. All correlations are significant, indicating highly consistent model behaviour across folds. **b**, Same analysis as in (a), applied to gradient alignment patterns. Again, strong and significant correlations confirm that fold-specific variability is minimal. **c**, Gradient alignments for 100 latent samples from 7 example patients. Latent samples were drawn from a Gaussian distribution centered on the patient's mean latent brain-graph representation (with standard deviation $\sigma = 2$). Alignment patterns are highly consistent across samples. In all main analyses, we used the average across CV-fold models and latent samples.

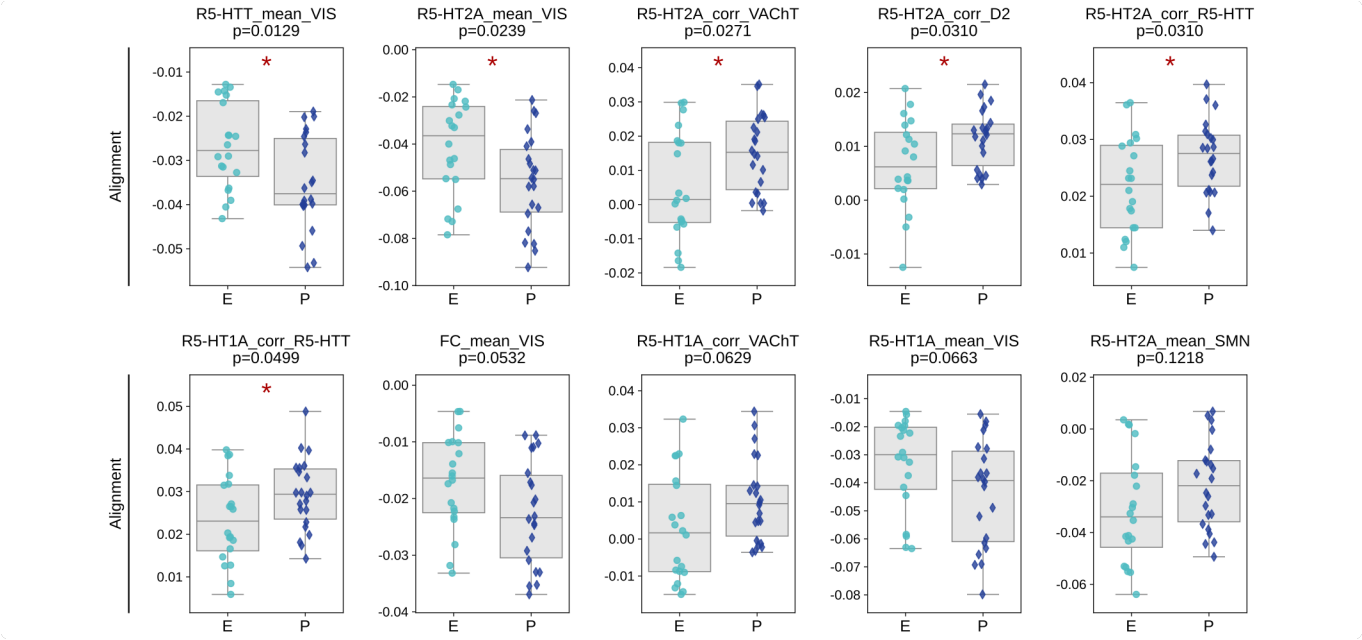


FIG. 12. **Biomarkers showing differential association with treatment outcome across drug conditions.** Boxplots show gradient alignment values for the top 10 biomarkers identified by `graphTRIP` that exhibit the most significant differences between escitalopram (E) and psilocybin (P) treatment groups. Each dot represents a single patient, and the y-axis indicates the alignment value of the corresponding biomarker. These differences suggest that the predictive role of specific biomarkers may be modulated by the drug condition, and that `graphTRIP` captured these treatment-brain interactions to some extent.

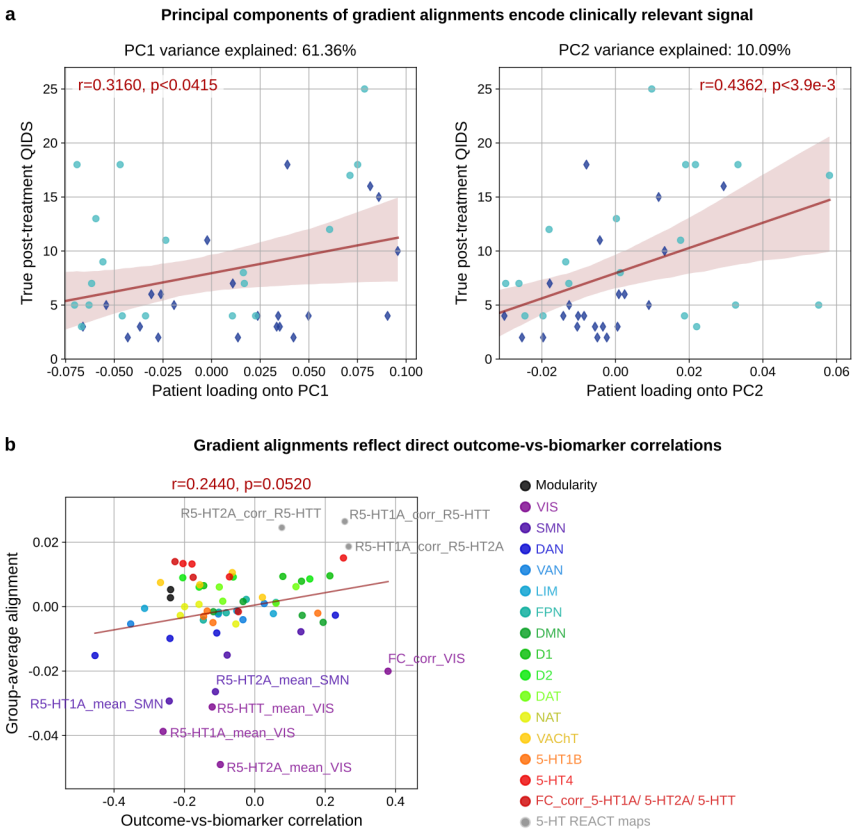


FIG. 13. **Validation of GRAIL biomarker relevance.** **a**, Principal component analysis (PCA) was performed on the GRAIL matrix, where rows correspond to patients and columns to candidate biomarkers (i.e., gradient alignment values). The first two principal components (PCs) accounted for 71% of the total variance in alignment patterns, and the patient loadings onto these PCs significantly correlated with treatment outcome. This indicates that the latent space encodes clinically meaningful variation, and that this information is reflected in the GRAIL profiles. **b**, To assess the correspondence between GRAIL-derived alignment values and direct feature–outcome associations, we computed, for each biomarker, the Pearson correlation between its values (derived from the patient brain graphs) and treatment outcome. We then correlated these univariate correlation values across biomarkers with their group-averaged gradient alignment values. The resulting borderline significant correlation ($r = 0.244$, $p = 0.052$) suggests that GRAIL captures biomarkers with predictive relevance, while also encoding dependencies that extend beyond univariate effects. Notably, gradient alignments are patient-specific, whereas feature–outcome correlations are computed at the group level, so a perfect correspondence is not expected.

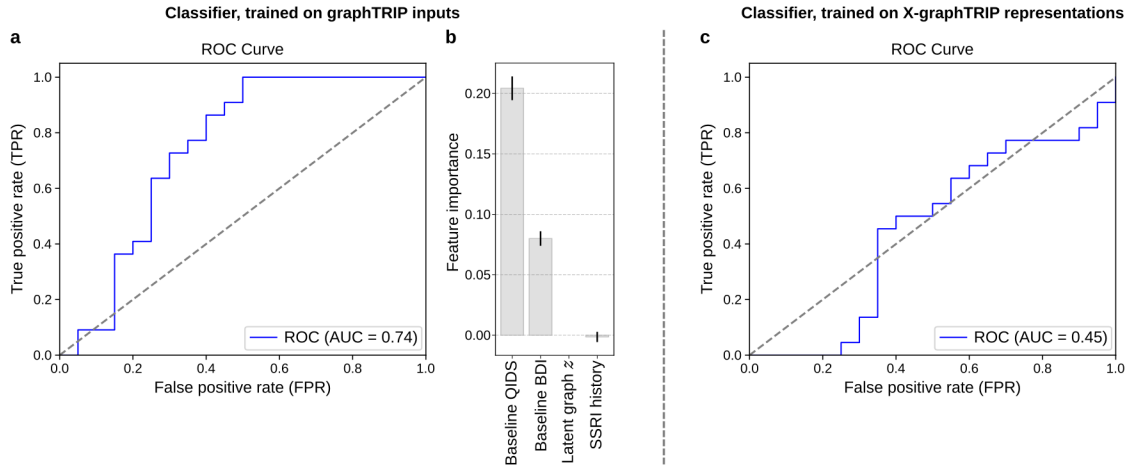


FIG. 14. **Treatment condition can be predicted from pre-treatment brain graphs and clinical data, but not from the latent representations of `X-graphTRIP`. a**, A classifier based on the `graphTRIP` architecture, but using a logistic regression output layer, was trained to predict treatment condition from brain graphs and pre-treatment clinical data (QIDS, BDI, prior SSRI use). The confusion matrix (left) shows that the model correctly classified most patients (decision threshold = 0.5). The raincloud plot (center) indicates a significant difference in predicted class probabilities between escitalopram and psilocybin groups (Mann–Whitney U test), and the ROC curve (right) confirms robust above-chance performance across decision thresholds. **b**, A second classifier was trained on the latent representations from the pre-trained VGAE of `X-graphTRIP` (with weights frozen), combined with pre-treatment clinical data. Here, the model failed to distinguish between treatment arms: it consistently predicted escitalopram (confusion matrix, left), and showed poor discriminative performance (ROC curve, right), with predicted probabilities tightly clustered around 0.45 (raincloud plot, center). This suggests that treatment condition was not encoded in the VGAE latent space.

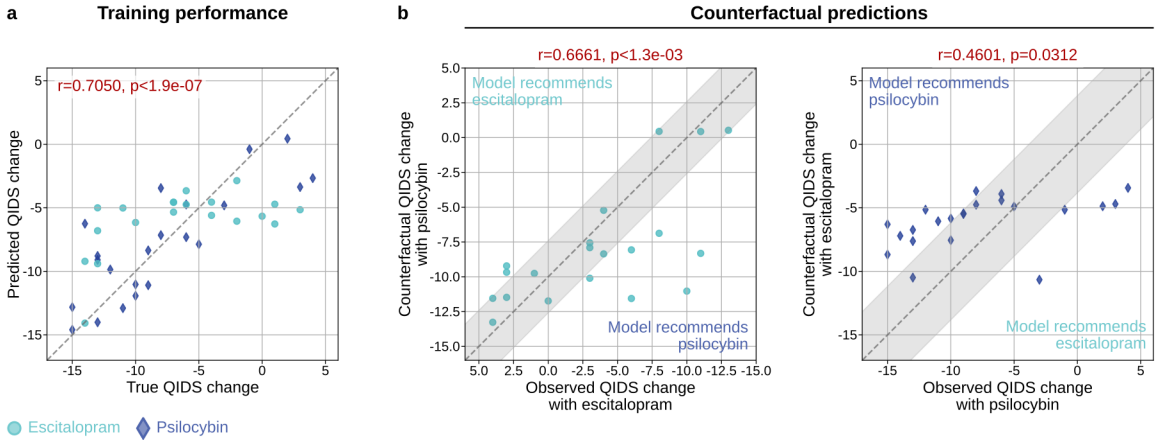


FIG. 15. **T-learner model fit and counterfactual predictions. a**, Model fit of the T-learners. Two separate `graphTRIP`-type models (T-learners) were trained independently on patients from each treatment condition. This panel shows true vs predicted post-treatment QIDS scores for each model on its own training data. Predictions from both models are combined into one plot. The high correlation indicates good within-treatment model fit. **b**, Counterfactual predictions. Left: true QIDS scores of escitalopram-treated patients plotted against their counterfactual predictions under psilocybin, generated by the psilocybin-trained T-learner. Points below the identity line indicate patients expected to respond better to psilocybin, and vice versa. The grey band shows the mean absolute error (MAE) of the psilocybin-trained T-learner at convergence, providing a reference for prediction uncertainty. Right: same analysis for psilocybin-treated patients, with counterfactual predictions under escitalopram from the escitalopram-trained T-learner.

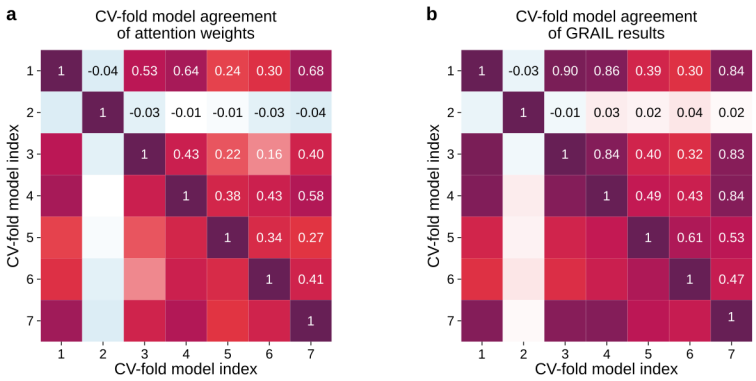


FIG. 16. **Cross-validation model disagreement in attention weights and gradient alignments for `X-graphTRIP`.** `X-graphTRIP` was trained using 7-fold cross-validation (CV), producing seven models. This figure evaluates the consistency of interpretability outputs across folds, following the same procedure as in Supp. Fig. 11. **a**, Pairwise correlations of regional attention weights across CV-fold models. For each patient, attention weights from all model pairs were correlated; the heatmap displays the average correlation values across patients. Only 13% of correlations are statistically significant, indicating considerable variability. **b**, Same analysis applied to gradient alignment patterns. Here, 20% of correlations reached significance, again reflecting substantial cross-model disagreement.

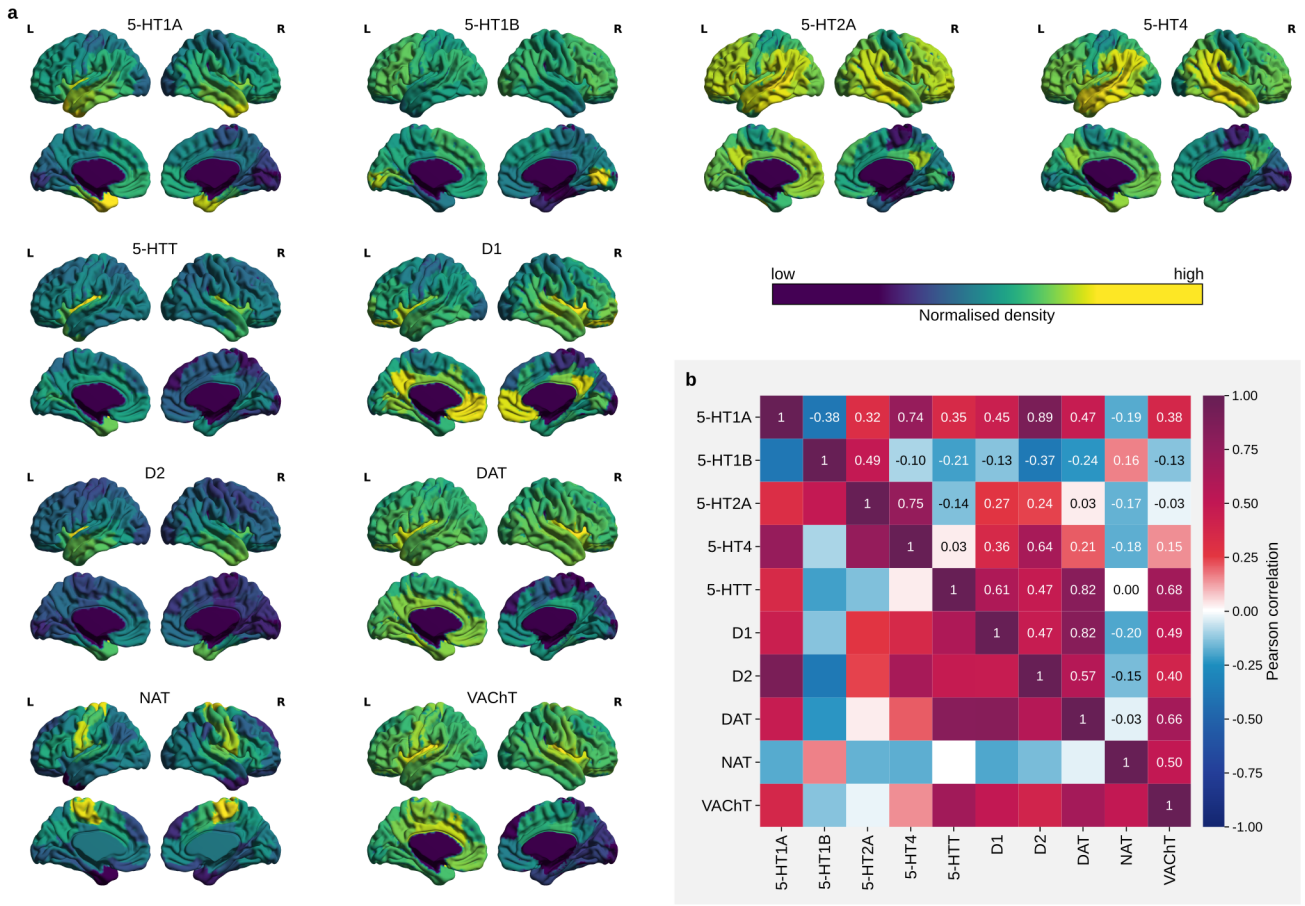


FIG. 17. **Normative molecular target distributions and cross-correlations.** **a**, Normative molecular target maps used in this study. Each map is individually normalised; the colour scale therefore indicates relative density within a given target map and should not be compared across targets. **b**, Cross-correlation matrix showing spatial similarity between normative target maps. Each row and column represents a molecular target, and matrix entries indicate Pearson correlation coefficients between pairs of target distributions.

Training configuration	
Learning rate	0.001
Number of epochs	300
Number of latent samples	3
Alpha	0.5
Batch size	7
Number of cross-validation folds	6
MLP configuration	
Regularisation strength	0.01
Dropout probability	0.25
MLP architecture	Input: 68; Layers: [64 → 64 → 64 → 1] with LeakyReLU (slope=0.01) and Dropout(0.25) between layers
VGAE configuration	
VGAE latent dimension	64
VGAE regularisation strength	0.01
Encoder	GATv2Conv layers: (1× 6→16, 2 heads), (4× 32→16, 2 heads), (1× 32→64, 1 head); LeakyReLU (0.01); LayerNorm(32); final layer: Linear(64 → 128)
Readout layer	AttentionNet: Linear(65→32) → ReLU → Linear(32→1)
Edge decoder	Linear(128→32) → LeakyReLU → Dropout(0.25) → Linear(32→32) → LeakyReLU → Dropout(0.25) → Linear(32→1); Activation: Tanh
Edge index decoder	Same as Edge decoder; Activation: Sigmoid
Node decoder	Linear(64→32) → LeakyReLU → Dropout(0.25) → Linear(32→32) → LeakyReLU → Dropout(0.25) → Linear(32→3)

TABLE II. **Model and training configuration for `graphTRIP`**. The same parameters were used for training `graphTRIP` to predict post-treatment BDI instead of QIDS.

Training configuration	
Learning rate	0.001
Number of epochs	150
Number of latent samples	0
Alpha	0.5
Batch size	7
Number of cross-validation folds	6
MLP configuration	
Regularisation strength	0.01
Dropout probability	0.25
MLP architecture	Input: 68; Layers: [64 → 64 → 64 → 1] with LeakyReLU (slope=0.01) and Dropout(0.25) between layers
VGAE configuration	
VGAE latent dimension	64
VGAE regularisation strength	0.01
Encoder	Node embedding MLP: [101 → 256 → 256 → 256 → 64] with LeakyReLU (slope=0.01) and Dropout(0.25) between layers; then concatenate node embeddings → Linear(6400→256) → LeakyReLU → Dropout(0.25) → Linear(256→128)
Decoder	Linear(64→256) → LeakyReLU → Dropout(0.25) → Linear(256→256) → LeakyReLU → Dropout(0.25) → Linear(256→4950); Activation: Tanh

TABLE III. **Model and training configuration for the atlas-bound model**.

Training configuration	
Learning rate	0.001
Number of epochs	150
Number of latent samples	0
Alpha	0.5
Batch size	7
Number of cross-validation folds	6
MLP configuration	
Regularisation strength	0.01
Dropout probability	0.25
MLP architecture	Input: 132; Layers: [128 → 128 → 128 → 1] with LeakyReLU (slope=0.01) and Dropout(0.5) between layers

TABLE IV. **Model and training configuration for the hybrid model**.

Training configuration	
Learning rate	0.001
Number of epochs	300
Number of latent samples	3
Alpha	0.5
Batch size	7
Number of cross-validation folds	6
MLP configuration	
Regularisation strength	0.01
Dropout probability	0.25
MLP architecture	Input: 67; Layers: [64 → 64 → 64 → 1] with LeakyReLU (slope=0.01) and Dropout(0.25) between layers; final activation: Sigmoid
VGAE configuration	
VGAE latent dimension	64
VGAE regularisation strength	0.01
Encoder	GATv2Conv layers: (1× 6→16, 2 heads), (4× 32→16, 2 heads), (1× 32→64, 1 head); LeakyReLU (0.01); LayerNorm(32); final layer: Linear(64 → 128)
Readout layer	GlobalAttentionPooling with attention gate: Linear(64 → 1)
Edge decoder	Linear(128→32) → LeakyReLU → Dropout(0.25) → Linear(32→32) → LeakyReLU → Dropout(0.25) → Linear(32→1); Activation: Tanh
Edge index decoder	Same as Edge decoder; Activation: Sigmoid
Node decoder	Linear(64→32) → LeakyReLU → Dropout(0.25) → Linear(32→32) → LeakyReLU → Dropout(0.25) → Linear(32→3)

TABLE V. **Model and training configuration for the drug classifier, trained on `graphTRIP`'s inputs**.

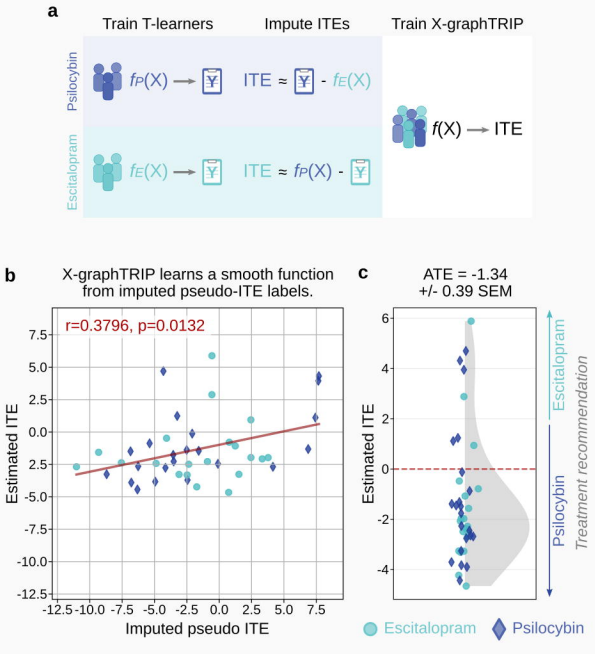
Training configuration	
Learning rate	0.001
Number of epochs	300
Number of latent samples	escitalopram T-learner: 3; psilocybin T-learner: 5
Alpha	0.5
Batch size	escitalopram T-learner: 10; psilocybin T-learner: 11
MLP configuration	
Regularisation strength	0.01
Dropout probability	0.25
MLP architecture	Input: 66; Layers: [64 → 64 → 64 → 1] with LeakyReLU (slope=0.01) and Dropout(0.25) between layers
VGAE configuration	
VGAE latent dimension	64
VGAE regularisation strength	0.01
Encoder	GATv2Conv layers: (1× 6→16, 2 heads), (4× 32→16, 2 heads), (1× 32→64, 1 head); LeakyReLU (0.01); LayerNorm(32); final layer: Linear(64 → 128)
Readout layer	GlobalAttentionPooling with attention gate: Linear(64 → 1)
Edge decoder	Linear(128→32) → LeakyReLU → Dropout(0.25) → Linear(32→32) → LeakyReLU → Dropout(0.25) → Linear(32→1); Activation: Tanh
Edge index decoder	Same as Edge decoder; Activation: Sigmoid
Node decoder	Linear(64→32) → LeakyReLU → Dropout(0.25) → Linear(32→32) → LeakyReLU → Dropout(0.25) → Linear(32→3)

TABLE VI. **Model and training configuration for escitalopram and psilocybin T-learners**. Only batch size and number of latent samples differ between the two models.

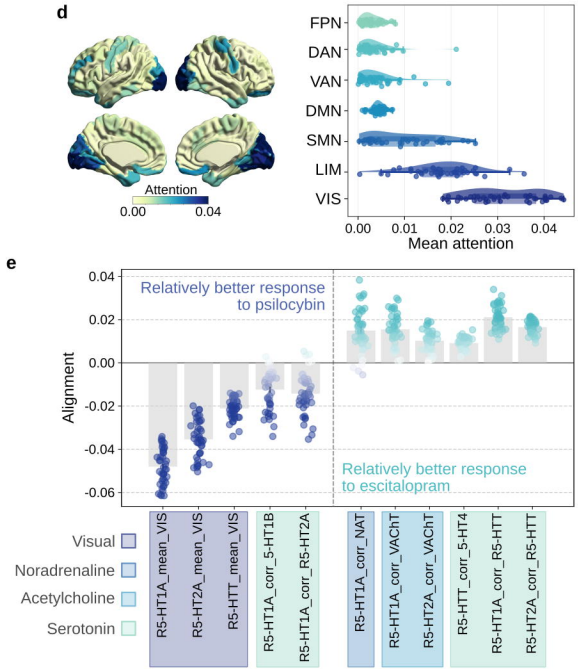
Training configuration	
Learning rate	0.001
Number of epochs	300
Number of latent samples	1
Alpha	0.3
Batch size	18
Number of cross-validation folds	7
MLP configuration	
Regularisation strength	0.01
Dropout probability	0.25
MLP architecture	Input: 66; Layers: [64 → 64 → 64 → 1] with LeakyReLU (slope=0.01) and Dropout(0.25) between layers
VGAE configuration	
VGAE latent dimension	64
VGAE regularisation strength	0.01
Encoder	GATv2Conv layers: (1× 6→16, 2 heads), (4× 32→16, 2 heads), (1× 32→64, 1 head); LeakyReLU (0.01); LayerNorm(32); final layer: Linear(64 → 128)
Readout layer	GlobalAttentionPooling with attention gate: Linear(64 → 1)
Edge decoder	Linear(128→32) → LeakyReLU → Dropout(0.25) → Linear(32→32) → LeakyReLU → Dropout(0.25) → Linear(32→1); Activation: Tanh
Edge index decoder	Same as Edge decoder; Activation: Sigmoid
Node decoder	Linear(64→32) → LeakyReLU → Dropout(0.25) → Linear(32→32) → LeakyReLU → Dropout(0.25) → Linear(32→3)

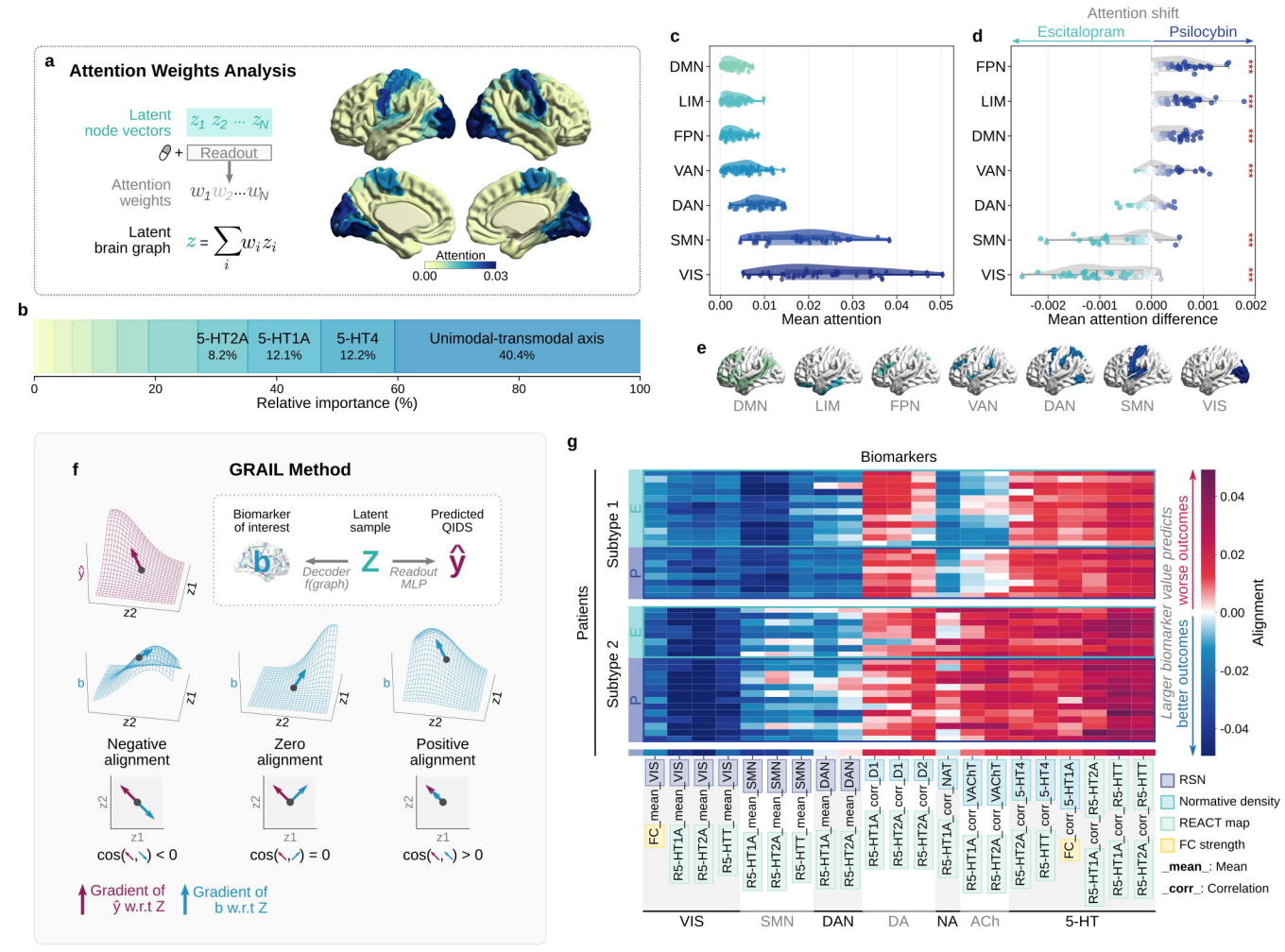
TABLE VII. **Model and training configuration for `X-graphTRIP`**.

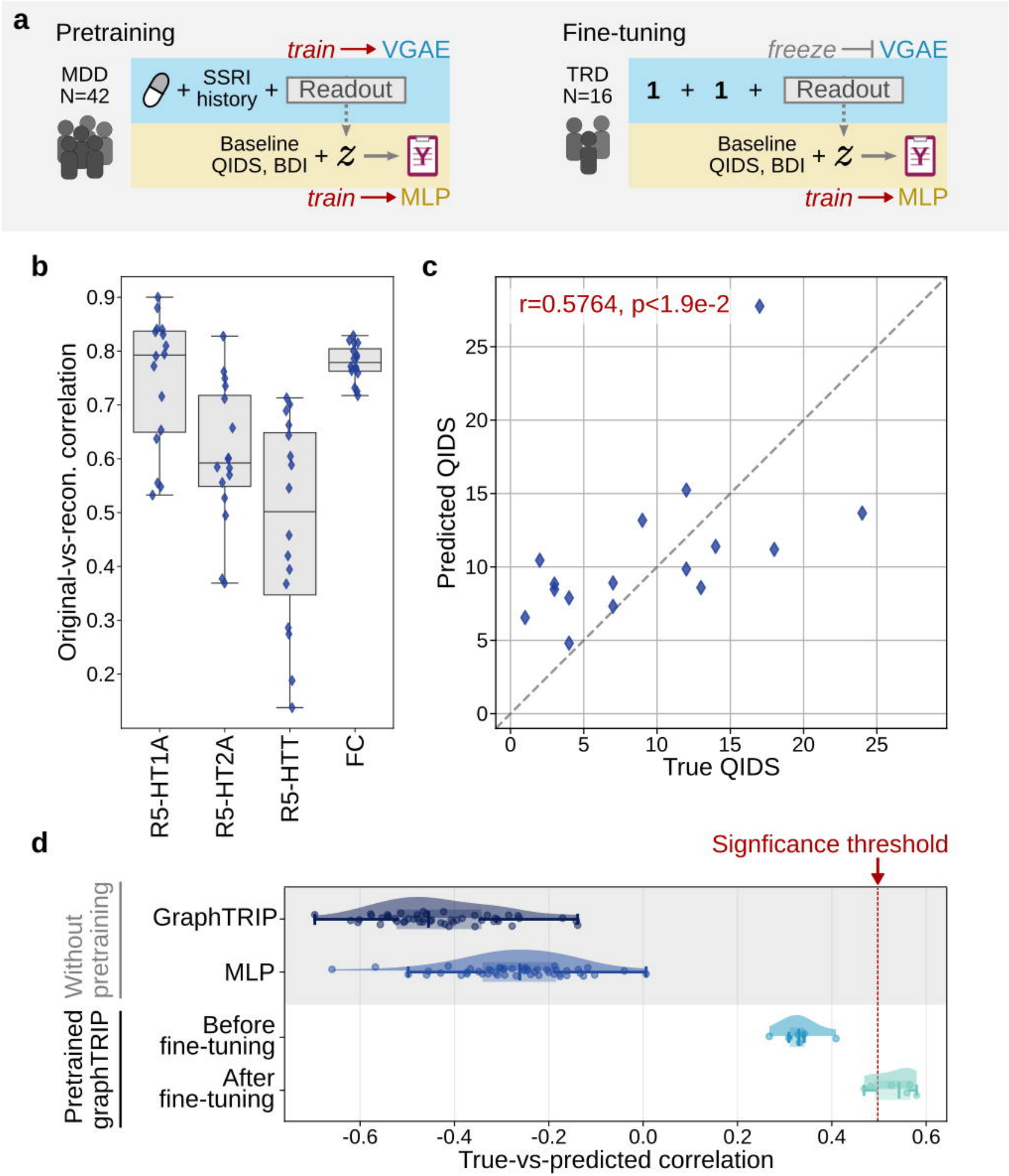
Estimating individual treatment effects (ITEs)

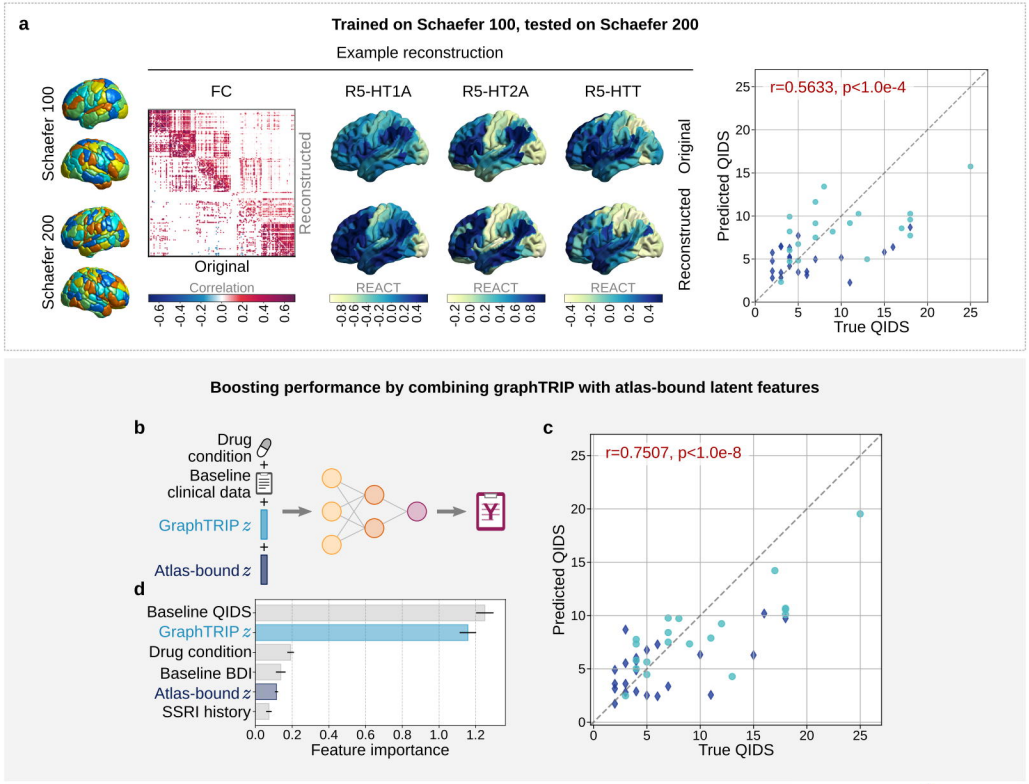


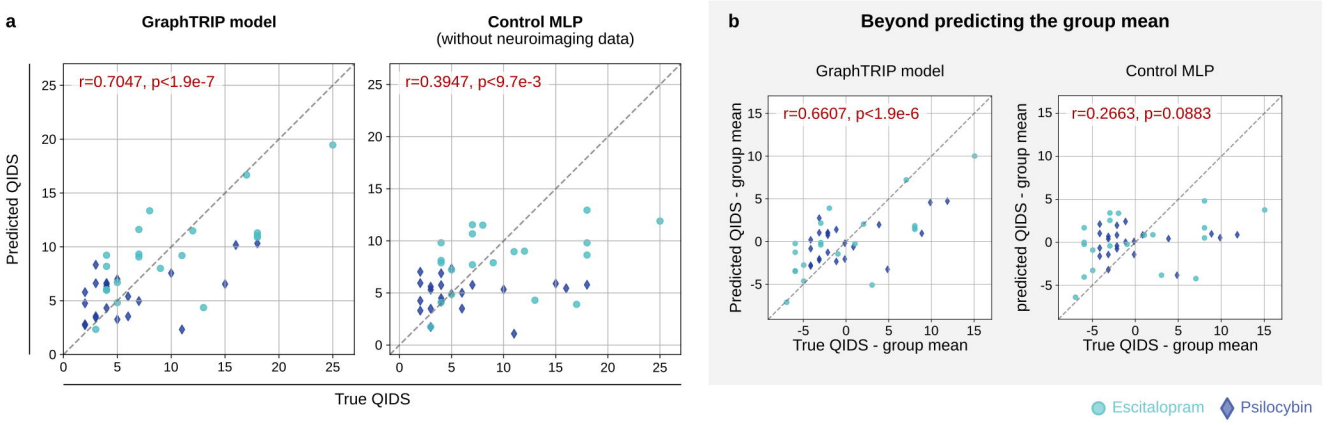
Identifying treatment-specific moderators

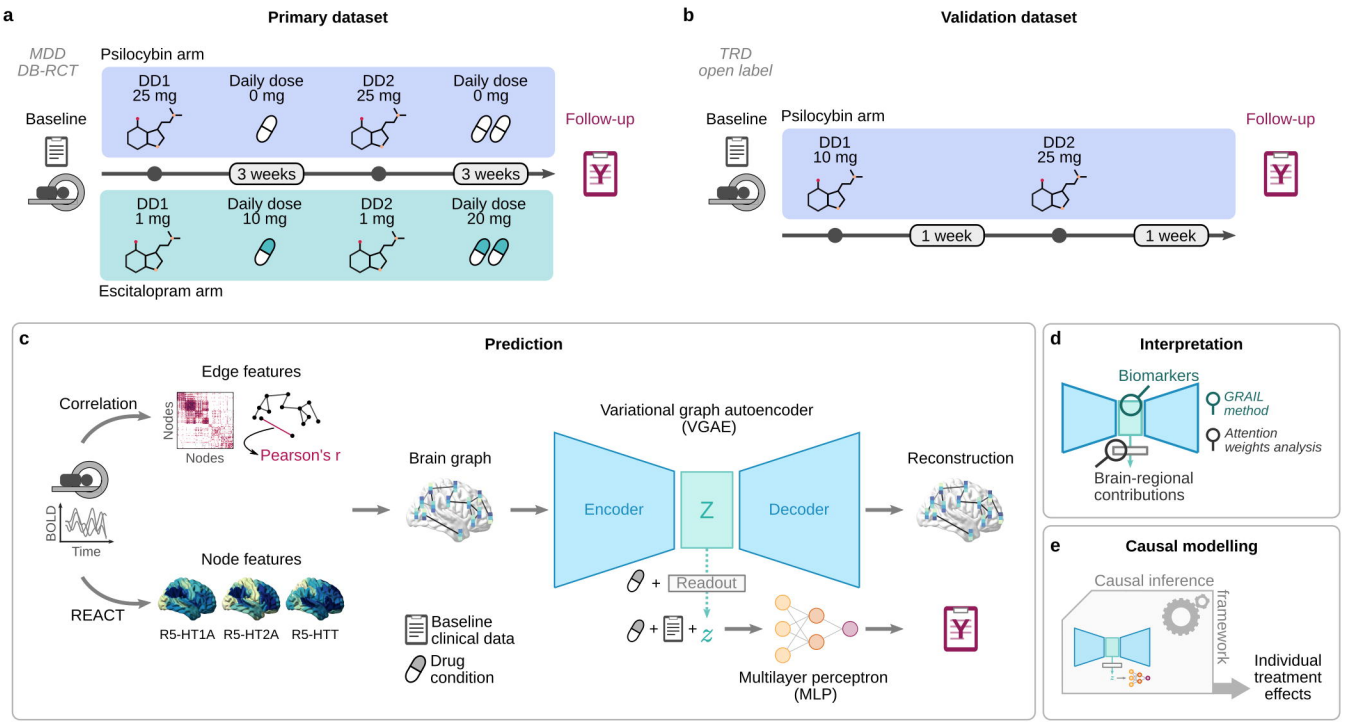



a
Attention Weights Analysis
Latent node vectors
$z_1\ z_2 \cdots z_N$
+ Readout
Attention weights
$w_1 w_2 \cdots w_N$
Latent brain graph
$z = \sum_i w_i z_i$
Attention
0.00
0.03
b
5-HT2A 8.2%
5-HT1A 12.1%
5-HT4 12.2%
Unimodal-transmodal axis 40.4%
0
20
40
60
80
100
Relative importance (%)
c
DMN
LIM
FPN
VAN
DAN
SMN
VIS
0.00
0.01
0.02
0.03
0.04
0.05
Mean attention
d
Attention shift
Escitalopram
Psilocybin
-0.002
-0.001
0.000
0.001
0.002
Mean attention difference
e
DMN
LIM
FPN
VAN
DAN
SMN
VIS
f
GRAIL Method
Biomarker of interest
Latent sample
Predicted QIDS
Decoder f(graph)
Readout MLP
Negative alignment
Zero alignment
Positive alignment
cos(↖,↘) < 0
cos(↖,↗) = 0
cos(↖,↖) > 0
Gradient of ŷ w.r.t Z
Gradient of b w.r.t Z
g
Biomarkers
Patients
Subtype 1
Subtype 2
Larger biomarker value predicts
worse outcomes
better outcomes
Alignment
0.04
0.02
0.00
-0.02
-0.04
FC_mean_VIS
R5-HT1A_mean_VIS
R5-HT2A_mean_VIS
R5-HTT_mean_VIS
R5-HT1A_mean_SMN
R5-HT2A_mean_SMN
R5-HTT_mean_SMN
R5-HT1A_mean_DAN
R5-HT2A_mean_DAN
R5-HT1A_corr_D1
R5-HT2A_corr_D1
R5-HT2A_corr_D2
R5-HT1A_corr_NAT
R5-HT1A_corr_VAChT
R5-HT2A_corr_VAChT
R5-HT2A_corr_5-HT4
R5-HTT_corr_5-HT4
FC_corr_5-HT1A
R5-HT1A_corr_R5-HT2A
R5-HT1A_corr_R5-HTT
R5-HT2A_corr_R5-HTT
VIS
SMN
DAN
DA
NA
ACh
5-HT
RSN
Normative density
REACT map
FC strength
mean: Mean
corr: Correlation


a
Pretraining
train → VGAE
MDD
N=42
+ SSRI history + Readout
Baseline QIDS, BDI + z
train → MLP
Fine-tuning
freeze ⊣ VGAE
TRD
N=16
1 + 1 + Readout
Baseline QIDS, BDI + z
train → MLP
b
Original-vs-recon. correlation
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
R5-HT1A
R5-HT2A
R5-HTT
FC
c
r=0.5764, p<1.9e-2
Predicted QIDS
True QIDS
0
5
10
15
20
25
d
Signficance threshold
Without pretraining
GraphTRIP
MLP
Pretrained graphTRIP
Before fine-tuning
After fine-tuning
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
True-vs-predicted correlation


a
Trained on Schaefer 100, tested on Schaefer 200
Example reconstruction
Schaefer 100
Schaefer 200
FC
Reconstructed
Original
Correlation
R5-HT1A
R5-HT2A
R5-HTT
Original
Reconstructed
REACT
r=0.5633, p<1.0e-4
Predicted QIDS
True QIDS
Boosting performance by combining graphTRIP with atlas-bound latent features
b
Drug condition
Baseline clinical data
GraphTRIP z
Atlas-bound z
c
r=0.7507, p<1.0e-8
Predicted QIDS
True QIDS
d
Baseline QIDS
GraphTRIP z
Drug condition
Baseline BDI
Atlas-bound z
SSRI history
Feature importance
Escitalopram
Psilocybin


a
GraphTRIP model
Control MLP
(without neuroimaging data)
r=0.7047, p<1.9e-7
r=0.3947, p<9.7e-3
Predicted QIDS
True QIDS
b
Beyond predicting the group mean
GraphTRIP model
Control MLP
r=0.6607, p<1.9e-6
r=0.2663, p=0.0883
Predicted QIDS - group mean
True QIDS - group mean
Escitalopram
Psilocybin
c
Example reconstruction
FC
R5-HT1A
R5-HT2A
R5-HTT
Reconstructed
Original
Correlation
REACT
d
Original-vs-recon.
Correlation
R5-HT1A
R5-HT2A
R5-HTT
FC
e
Baseline QIDS
Latent graph z
Baseline BDI
SSRI history
Drug condition
Feature importance


a
Primary dataset
MDD
DB-RCT
Psilocybin arm
DD1
25 mg
Daily dose
0 mg
DD2
25 mg
Daily dose
0 mg
Baseline
Follow-up
3 weeks
3 weeks
DD1
1 mg
Daily dose
10 mg
DD2
1 mg
Daily dose
20 mg
Escitalopram arm
b
Validation dataset
TRD
open label
Baseline
Psilocybin arm
Follow-up
DD1
10 mg
DD2
25 mg
1 week
1 week
c
Prediction
Edge features
Correlation
Nodes
Nodes
Pearson's r
BOLD
Time
Node features
REACT
R5-HT1A
R5-HT2A
R5-HTT
Brain graph
Variational graph autoencoder
(VGAE)
Encoder
Z
Decoder
Reconstruction
Readout
Baseline
clinical data
Drug
condition
Multilayer perceptron
(MLP)
d
Interpretation
Biomarkers
GRAIL
method
Attention
weights analysis
Brain-regional
contributions
e
Causal modelling
Causal inference
framework
Individual
treatment
effects