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 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 ($r = 0.72$, $p = 6.8 \times 10^{-8}$), and shows clear generalization to both an independent dataset and across brain atlases. The model identifies stronger functional connectivity within sensory networks as a robust predictor of poorer response across both treatments. In contrast, causal analysis implicates frontoparietal and default mode networks as key moderators of differential response, with stronger 5-HT1A- and 5-HT2A-related signalling in the frontoparietal network predicting escitalopram response but psilocybin resistance. Overall, this work advances precision medicine and biomarker discovery in depression.

Keywords: precision medicine, geometric deep learning, transformers, graphormers, 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

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

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 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.

While previous studies using machine learning (ML) have achieved notable successes in predicting antidepressant treatment outcome from brain graph features [25–28], key challenges remain for clinical translation. For instance, prior approaches have largely focused on binary classification of treatment response or remission [25, 27–29], yet binarising continuous depression scores sacrifices valuable information and relies on statistically arbitrary thresholds. Furthermore, 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 [30, 31].

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 undesirable. To ensure safe and reliable decisions, it is crucial to understand why a specific treatment outcome was predicted.

Finally, standard predictive approaches 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 move beyond mere prediction towards *causal* inference.

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. Lastly, we extend our model within a causal inference framework [32], 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 [33] 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).

Our model, called `graphTRIP` (graph-based Treatment Response Interpretability and Prediction), combines a variational graph-transformer 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 fixed-density thresholding of FC, retaining only the global 20% strongest edges. Node features encoded REACT values for 5-HT1A, 5-HT2A, and 5-HTT, denoted as R5-HT1A, R5-HT2A, and R5-HTT to distinguish 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 $\boldsymbol{z}$. This vector, combined with the drug condition (psilocybin or escitalopram) and baseline clinical data (QIDS and BDI depression scores, and a binary indicator of prior SSRI use), serves as input to the MLP.

We applied two analyses to probe model predictions: 1) Regional attribution analysis estimates brain regional contributions to the prediction; 2) GRAIL quantifies learned associations between brain-graph features and treatment outcome (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).

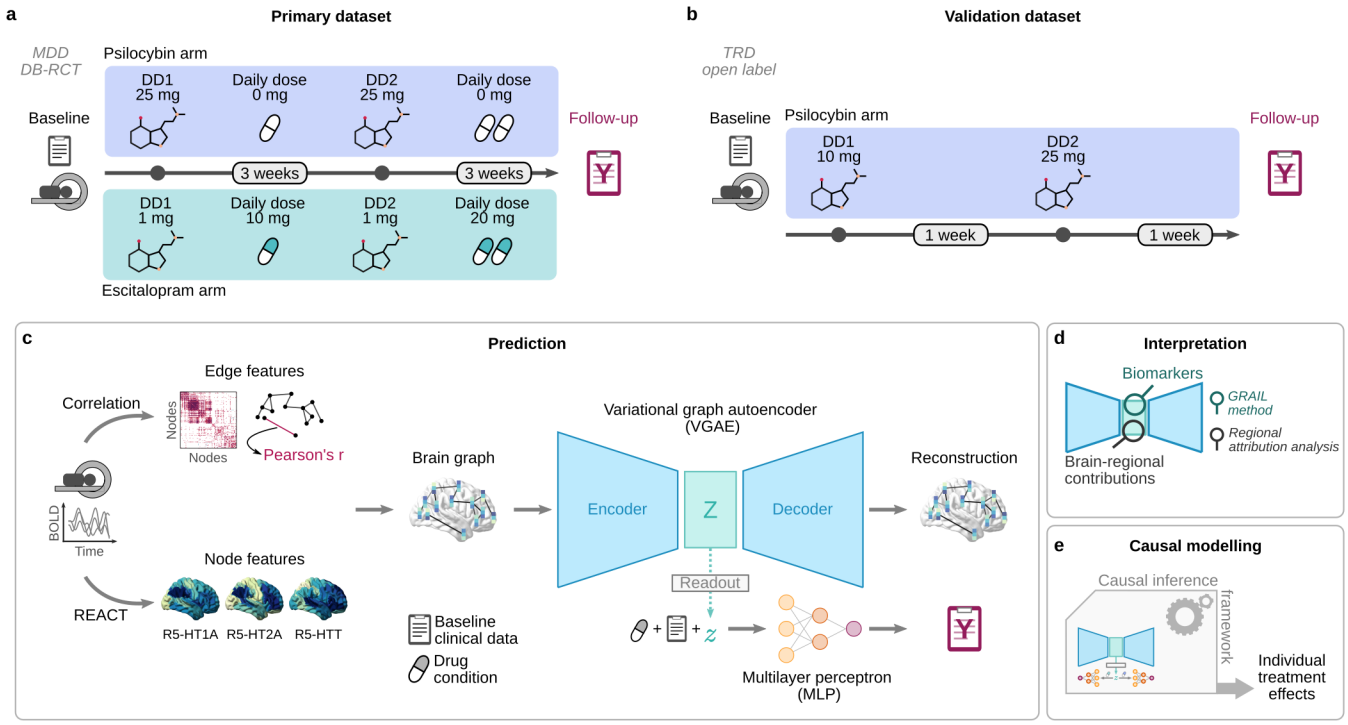

FIG. 1. **Overview of study design and methodological framework.** **a**, Primary dataset: 42 MDD patients from a DB-RCT, comparing escitalopram ($N = 20$) and psilocybin ($N = 22$) treatment. Escitalopram and psilocybin groups, respectively, received inactive and high doses of psilocybin on each dosing day (DD1, DD2). 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. QIDS was measured 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.

Accurate, robust predictions of treatment response

`graphTRIP` achieved a strong correlation between true and predicted post-treatment QIDS scores ($r = 0.7220$, $p = 6.8 \times 10^{-8}$; Fig. 2a) and showed good reconstruction performance (Fig. 2c-d). Unless stated otherwise, all reported predictions correspond to mean test predictions averaged across 10 repeated training runs with different random seeds. In contrast, an ordinary least squares (OLS) linear regression model fitted to the clinical data and drug condition (i.e., without neuroimaging data) performed substantially worse ($r = 0.3127$, $p = 0.0437$; Fig. 2a). Further analysis suggests that the linear model primarily predicted the mean QIDS score for each drug condition, which was higher in the escitalopram group [33]. After controlling for the drug condition, the partial correlation between true and predicted scores remained significant for `graphTRIP` ($r = 0.6833$, $p = 6.2 \times 10^{-7}$), but not for the linear model ($r = 0.0530$, $p = 0.7390$; 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. 7a). Across random seeds, `graphTRIP` consistently achieved highly significant performance and significantly outperformed all other benchmark models (Supp. Fig. 7b).

To assess the contribution of latent brain-graph features in the `graphTRIP` model, we performed permutation 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. The latent brain-graph feature vector $\boldsymbol{z}$ was the most important predictor, followed by baseline QIDS and BDI scores (Fig. 2e).

We additionally trained a `graphTRIP` model to predict post-treatment BDI scores. We achieved significant predictions ($r = 0.5254$, $p = 3.5 \times 10^{-4}$; Supp. Fig. 8), 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

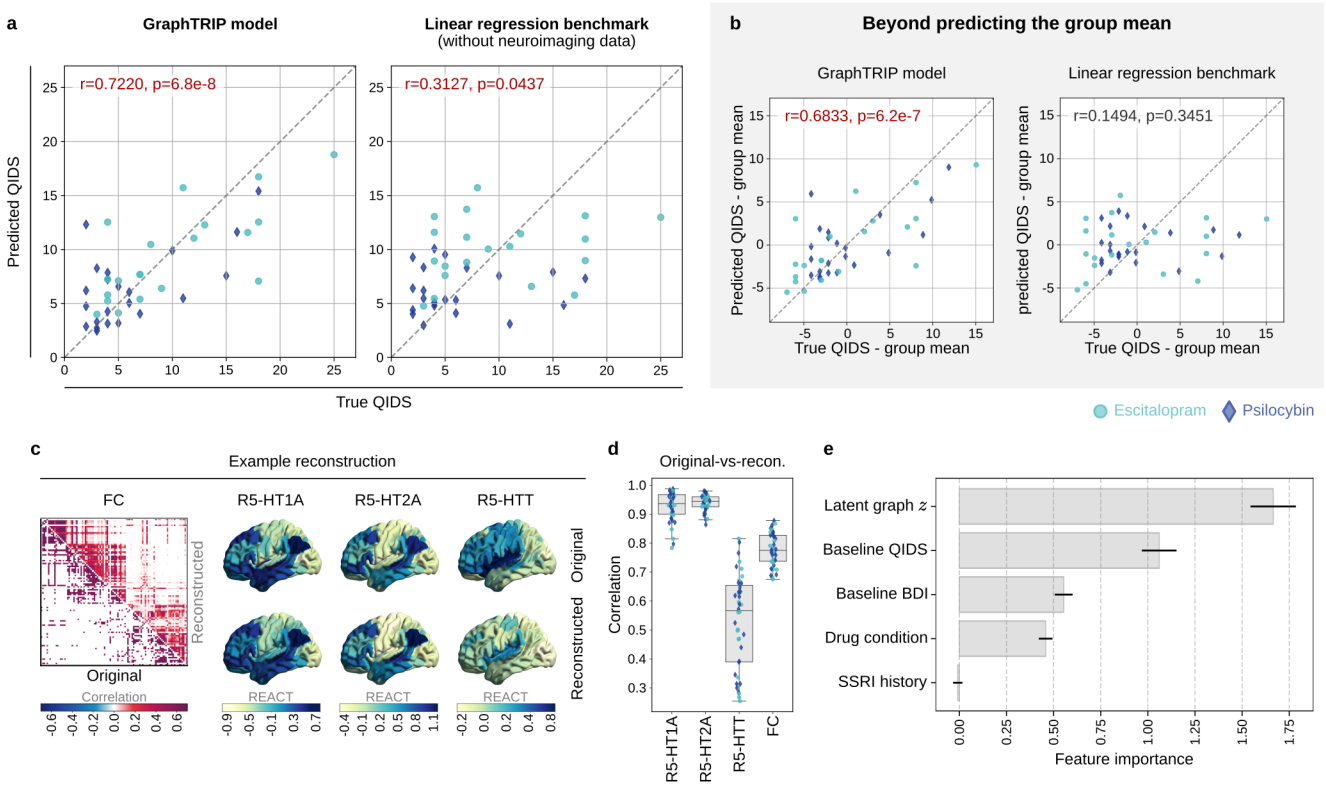

FIG. 2. **Reconstruction and prediction performance of the `graphTRIP` model.** **a**, `graphTRIP` significantly predicts post-treatment QIDS, outperforming an OLS linear regression model trained only on clinical data and drug condition, without latent neuroimaging features. **b**, After controlling for the group difference in post-treatment QIDS, the partial correlation between true and predicted scores remains significant for the `graphTRIP` model, but not for the linear regression model. **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.

the one used for training (Schaefer 100 [34]). On Schaefer 200, which has twice as many parcels, `graphTRIP` maintained highly accurate reconstructions and QIDS prediction performance ($r = 0.7156$, $p = 10^{-7}$; Fig. 3). We further replicated this result on the AAL atlas [35], which differs fundamentally from Schaefer and includes subcortical regions, where `graphTRIP` also generalised well ($r = 0.6049$, $p = 2.2 \times 10^{-5}$; Supp. Fig. 9).

Validation in an independent dataset

We evaluated `graphTRIP` on an independent dataset that differed from the primary dataset in several key aspects (Fig. 1a), including the presence of only one treatment arm (psilocybin), a different psilocybin treatment protocol, patients with treatment-resistant depression (TRD), and a different follow-up time point.

Given these differences, direct application of `graphTRIP` to the validation dataset was not expected to yield strong predictive performance. Nevertheless, the model produced consistently positive true-vs-predicted correlations (Supp. Fig. 10a) and showed only a modest reduction in VGAE reconstruction accuracy (Fig. 4b). This suggests that the model learned transferable structure that could be leveraged through fine-tuning.

Fine-tuning `graphTRIP` on the validation dataset required adapting the MLP prediction head to account for dataset-specific constraints. Specifically, because all validation subjects had TRD and received psilocybin, both SSRI history and drug condition inputs were constant and therefore uninformative for training. We therefore first pretrained new MLP prediction heads on the primary dataset using only latent graph representations – generated by frozen VGAEs from the core `graphTRIP` model – and baseline QIDS and BDI scores as the MLP inputs; that is, excluding drug condition and SSRI history (Fig. 4a). These pretrained models achieved significant prediction performance on the primary dataset ($r = 0.6293$, $p = 8.0 \times 10^{-6}$; Supp. Fig. 10b). We then transferred the pretrained models to the validation dataset, again freezing the VGAEs and fine-tuning only the MLP weights (see Methods).

Following fine-tuning, `graphTRIP` achieves a significant

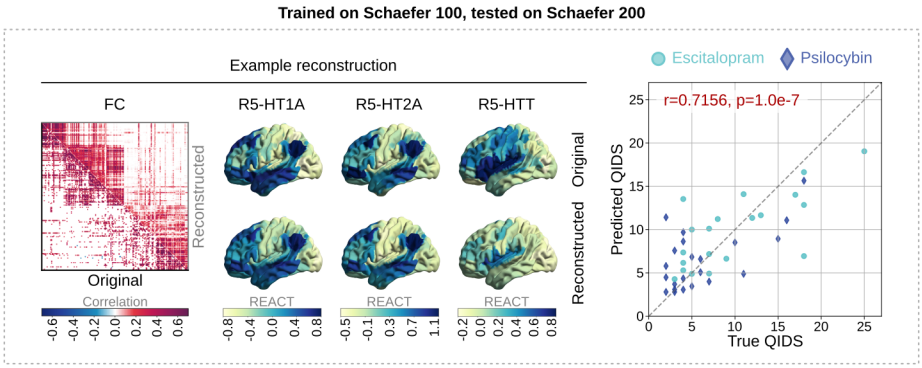

FIG. 3. **graphTRIP generalises across brain atlases without retraining.** The graphTRIP model, trained on the Schaefer 100 brain atlas, generalises to Schaefer 200 while maintaining strong reconstruction and prediction performance. The panel shows FC and node feature reconstructions for an example subject, and the MLP's prediction performance.

correlation between true post-treatment QIDS scores and mean predictions on the validation dataset ($r = 0.5104$, $p = 0.0434$; Fig. 4c). Notably, given the small sample size ($N = 16$), training models from scratch proved infeasible. In contrast, fine-tuning outperformed all newly trained benchmark models across random seeds (Fig. 4d).

Estimating individual treatment effects

Identifying which treatment is most effective for a given patient requires estimating individual treatment effects (ITEs) in a causal analysis. This relies on several assumptions, including strong ignorability (no unmeasured confounding given observed covariates). RCTs are designed to approximate these conditions, and we therefore assume that they largely hold for the present analysis. We note, however, that small-sample effects and post-randomisation factors such as dropouts can introduce subtle imbalances. Indeed, we observed an almost-significant difference in baseline QIDS between treatment groups, with lower baseline severity in the psilocybin arm (two-sample t-test: $t = 1.9928$, $p = 0.0531$, Cohen's $d = 0.6104$).

Under these assumptions, one could in principle estimate ITEs using graphTRIP by generating predictions for each patient under both treatment conditions. However, while graphTRIP provides a robust predictor of post-treatment QIDS, its predictions under psilocybin and escitalopram for the same subject were almost perfectly correlated and differed primarily by a near-constant offset (Supp. Fig. 11). This suggests that the model captured an average treatment effect but did not learn strong brain-drug interaction effects. This behaviour is consistent with known limitations of single-model approaches like graphTRIP, which can underfit treatment-covariate interactions and thus attenuate heterogeneous treatment effects when baseline features are substantially more predictive of outcomes than the treatment itself [36].

To explicitly model potential drug-specific response patterns, we developed Medusa-graphTRIP – an extension of graphTRIP inspired by the Counterfactual Regression framework [32]. Medusa-graphTRIP uses a shared VGAE to learn a common latent brain-graph representation across both treatment arms, followed by two separate MLP heads that predict post-treatment QIDS, one for each treatment condition (Fig. 5a).

Compared to a single-head model, this two-head structure tends to be better at capturing heterogeneous treatment effects when response functions differ across treatments [36]. However, since each head is trained on fewer samples, this approach trades reduced bias for increased variance. Thus, we use graphTRIP as a more stable estimator of general treatment response, whereas Medusa-graphTRIP is introduced to disentangle treatment-specific response patterns.

Medusa-graphTRIP predicted post-treatment QIDS with high accuracy ($r = 0.7315$, $p = 3.8 \times 10^{-8}$; Fig. 5b), and produced faithful brain-graph reconstructions (Supp. Fig. 12a). ITEs were estimated as the difference between predictions from the psilocybin and escitalopram heads, such that negative values indicate predicted better response to psilocybin (Fig. 5a). Consistent with the significant group difference in post-treatment QIDS in the data (two-sample t-test: $t = 2.2174$, $p = 0.0323$, Cohen's $d = 0.6805$), most patients were predicted to respond better to psilocybin (Fig. 5c), with an estimated average treatment effect of ATE $= -4.84 \pm 0.68$ (standard error). Nevertheless, seven patients were estimated to benefit more from escitalopram.

Regional attributions align with the unimodal-transmodal cortical hierarchy

We assessed the contributions of individual brain regions to predicting 1) post-treatment QIDS using graphTRIP, and 2) ITEs using Medusa-graphTRIP. To this end, we performed a regional attribution analysis, which quantifies how small perturbations to each region's

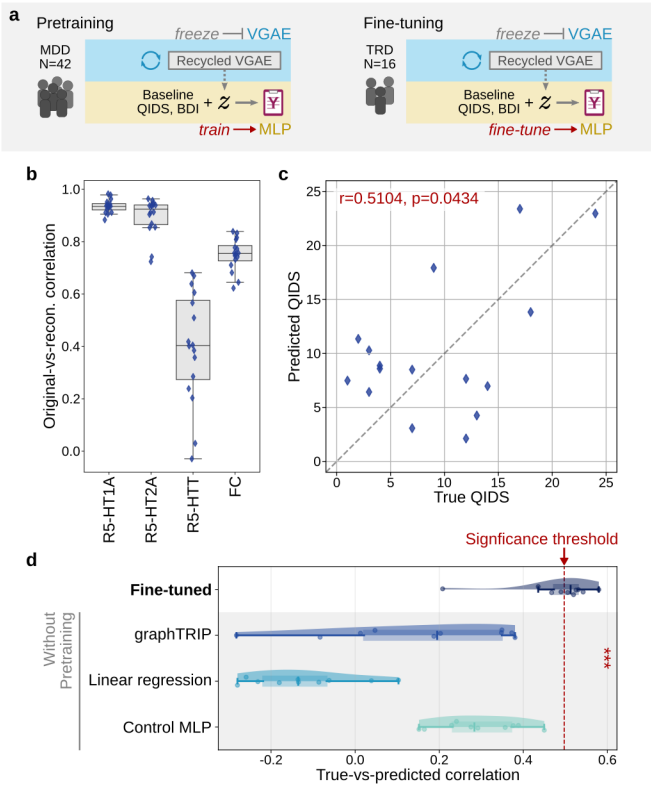

FIG. 4. **Transfer learning enables generalisation to an independent dataset. a**, Overview of the fine-tuning strategy. New MLP prediction heads were first pretrained on the primary dataset using only inputs compatible with the validation data, and with the VGAEs recycled from the core model and kept frozen. These pretrained models were then transferred to the validation dataset and fine-tuned, again updating only the MLP parameters. **b**, Original versus reconstructed brain-graph feature correlations for all patients in the validation dataset. **c**, QIDS prediction performance after fine-tuning. **d**, Training models from scratch on the validation dataset – including randomly initialised `graphTRIP` models, OLS linear regression, and MLPs trained on all available demographic and clinical data – fails due to the small sample size. In contrast, fine-tuning yields significantly better performance across 10 random seeds (Friedman test for global effect: $p = 5.3 \times 10^{-5}$; post-hoc Wilcoxon tests vs. all non-fine-tuned models, $p < 0.01$ after FDR correction).

latent representation influences the predicted outcome for a given patient. Specifically, the gradient of the predicted outcome $\hat{y}$ with respect to the latent space decomposes into node-wise partial gradients $\nabla_{\boldsymbol{z}_i}\hat{y}$, one for each brain region i. Regional attribution scores are defined as the percentage energy (sum of squares) of each node-wise gradient from the total energy of the full gradient vector. This gradient-based approach is conceptually similar to established attribution methods [37].

Averaging attribution profiles across patients revealed that `graphTRIP` assigns the largest contributions to sensory-motor (SMN) and visual (VIS) regions, and the smallest contributions to association cortices (Fig. 6a). In contrast, `Medusa-graphTRIP`'s ITE estimates relied most strongly on regions within the default mode (DMN) and frontoparietal (FPN) networks, with comparatively lower contributions from SMN and VIS regions (Fig. 6b). These mirror-reversed spatial patterns closely align with the unimodal-transmodal cortical gradient, which captures key anatomical, functional, and neurochemical distinctions across the cortex [20, 38–41].

To formally test this relationship, we fitted a linear model predicting regional attribution weights from each region's position along the unimodal-transmodal axis together with ten normative molecular target density maps derived from open-access datasets [20, 41, 42] (see Methods). Model predictions correlated significantly with observed attribution patterns for both `graphTRIP` and `Medusa-graphTRIP` when assessed against 1,000 spatial autocorrelation–preserving null models [43] (`graphTRIP`: $r = 0.4348$, $p = 0.0020$; `Medusa-graphTRIP`: $r = 0.8946$, $p = 0.0010$). Dominance analysis identified the unimodal-transmodal axis as the strongest predictor in both cases, followed by dopamine, noradrenaline, and serotonin transporter densities for `graphTRIP` outcome predictions, and by serotonin receptor densities (5-HT2A, 5-HT4, and 5-HT1A) for `Medusa-graphTRIP` ITE estimates (Fig. 6a,b).

Biomarkers of treatment responsiveness

Beyond regional attributions, we developed GRAIL to systematically link treatment response predictions to brain graph-derived biomarkers (Fig. 6c). 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 higher prediction values (e.g., higher post-treatment QIDS), and negative alignment suggests the opposite. Zero alignment suggests no learned association. Alignment significance was assessed by comparing observed values against null distributions generated using spatial autocorrelation–preserving spin permutations (see Methods).

We applied GRAIL to our models to assess learned associations for a range of candidate biomarkers, 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 antidepres-

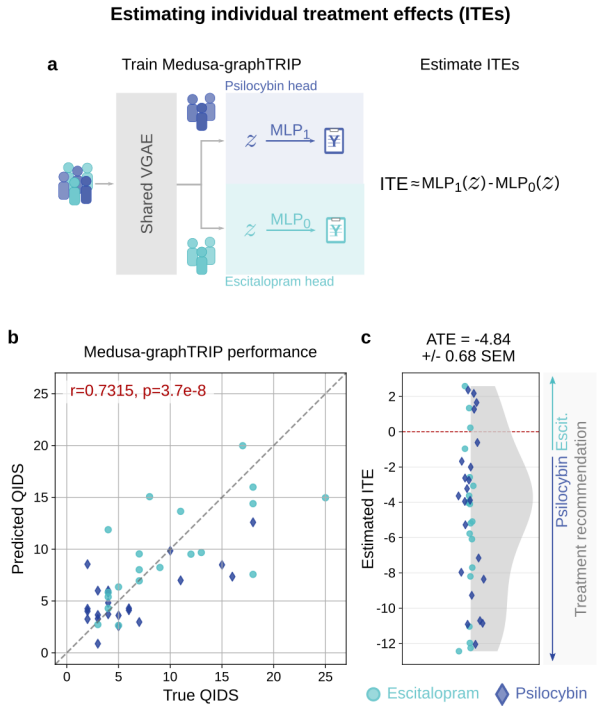

FIG. 5. **Causal inference with `Medusa-graphTRIP` enables estimation of individual treatment effects (ITEs).** **a**, Schematic of the `Medusa-graphTRIP` architecture. A shared VGAE learns a common latent representation of pre-treatment brain graphs, which is passed to two separate MLP outcome heads trained on data from a single treatment condition each (psilocybin, MLP_1; escitalopram, MLP_0). ITEs are estimated as the difference between predictions from the two heads. **b**, `Medusa-graphTRIP` accurately predicts post-treatment QIDS. **c**, Estimated individual treatment effects indicate that most patients are predicted to respond better to psilocybin than escitalopram, with an average treatment effect (ATE) of -4.84 ± 0.68 standard error.

sant treatment [12]: serotonin (5-HT), dopamine (DA), noradrenaline (NA), acetylcholine (ACh). Candidate biomarkers included 1) mean brain-graph features (i.e., FC or REACT maps) within each RSN, and 2) 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.

To distinguish different types of biomarkers, we applied GRAIL in four complementary modes, corresponding to models trained to predict 1) overall response to either treatment (`graphTRIP`); 2) response to escitalopram (`Medusa-graphTRIP`'s escitalopram head); 3) response to psilocybin (`Medusa-graphTRIP`'s psilocybin head); and 4) ITEs (the difference between `Medusa-graphTRIP`'s heads). In the modes (1-3), gradient alignment quantifies whether a given biomarker is associated with better or worse response in general (1), or under the specific treatment (2-3). In mode (4), gradient alignment quantifies whether a biomarker is associated with relatively better response to escitalopram (positive alignment) or psilocybin (negative alignment). Importantly, by comparing the direction of alignment across these modes, biomarkers can be classified according to whether they predict treatment-invariant or drug-specific responsiveness or resistance. For example, a biomarker that showed significant negative alignment across modes (1-3), but no significant alignment with ITE predictions (mode 4), was classified as a shared responsiveness biomarker. Formal definitions of all biomarker categories are provided in the Methods.

We focused on candidate biomarkers that were assigned a significant category in at least 50% of individuals. Using this criterion, GRAIL identified multiple biomarkers associated with shared treatment responsiveness or resistance, as well as two biomarkers of differential response between psilocybin and escitalopram (Fig. 6d). Shared responsiveness biomarkers included higher network modularity within RSNs, and increased FC of regions with high normative 5-HT2A receptor density (FC_corr_5-HT2A). In contrast, higher FC within VIS and SMN regions (FC_mean_VIS, FC_mean_SMN) was associated with overall treatment resistance. Interestingly, mean functional coupling between FPN regions and regions with high normative 5-HT1A and 5-HT2A receptor density (R5-HT1A_mean_FPN, R5-HT2A_mean_FPN) predicted resistance to psilocybin but response to escitalopram.

To validate the biomarker associations identified by GRAIL, we fitted a multivariate OLS regression model including the drug condition and all identified biomarkers as main effects, together with drug–biomarker interaction terms for the biomarkers that were classified as differential (i.e., R5-HT1A_mean_FPN, R5-HT2A_mean_FPN). The overall model was significant ($r = 0.7389$, $p = 0.0168$, $n = 42$), with FC_mean_VIS showing a significant positive main effect ($p = 0.037$) and R5-HT2A_mean_FPN a significant drug interaction effect ($p = 0.017$), consistent with the GRAIL analysis.

While not all biomarkers identified by GRAIL reached significance in the multivariate regression model, univariate biomarker–outcome correlations replicated the sign of GRAIL alignment (Supp. Fig. 13a). Moreover, across all 69 candidate biomarkers, group-averaged alignments were significantly correlated with biomarker–outcome correlations ($r = 0.3896$, $p = 9.4 \times 10^{-4}$, $n = 69$; Supp. Fig. 13b). In contrast, a linear ridge regression model trained to predict post-treatment QIDS directly from the candidate biomarkers did not achieve significant performance ($r = -0.1995$, $p = 0.2052$, $n = 42$; Supp. Fig. 13c). Together, these results show that GRAIL recovers the expected direction of association for biomarkers with linear outcome relationships, while also reflecting additional dependencies learned by the predictive model that are not captured by linear regression.

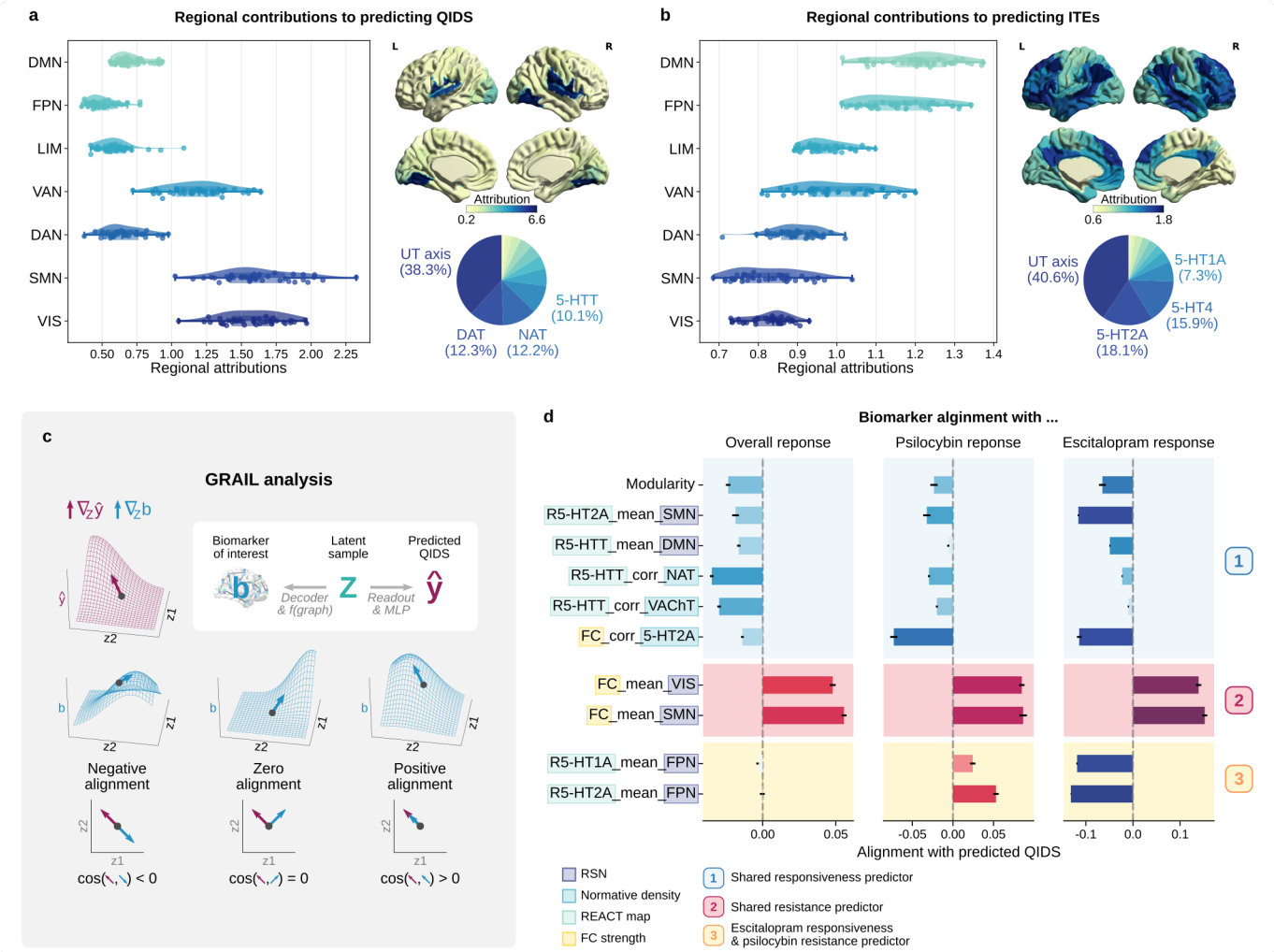

FIG. 6. **Interpretability analyses reveal unimodal regions as general outcome predictors and transmodal cortices as treatment-specific moderators.** **a**, Regional attribution analysis for `graphTRIP` quantifies each brain region's contribution to predicted post-treatment QIDS. Left: mean attributions aggregated within resting-state networks (RSNs). Top right: population-mean regional map. Bottom right: dominance analysis from a linear model explaining regional attributions from unimodal–transmodal (UT) axis scores and normative neuroreceptor density maps; pie segments indicate each regressor's relative contribution to explained variance (R^2). **b**, Regional attribution analysis applied to `Medusa-graphTRIP` individual treatment effect (ITE) predictions. **c**, Overview of GRAIL, which estimates associations between candidate biomarkers and model predictions via gradient alignment in latent space. **d**, Biomarker alignments with the predictions of three models: `graphTRIP` (left), and `Medusa-graphTRIP`'s psilocybin head (centre) and escitalopram head (right). Positive/negative values indicate association with worse/better outcomes under both drugs, psilocybin, and escitalopram, respectively. Bars show mean ± standard error across patients. Numbers indicate biomarker categorizations based on alignment directions across models.

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 `Medusa-graphTRIP`, a causal inference extension of `graphTRIP`, to estimate individual treatment effects and identify treatment-specific moderators.

`Medusa-graphTRIP` estimated an average treatment effect favouring psilocybin over escitalopram at three weeks post-treatment. This finding is consistent with previous analyses of the same dataset [33] 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 several brain-graph biomarkers that were associated with treatment outcome across both treatment arms, suggesting the presence of drug-invariant predictors of treatment responsiveness. Specifically, regional attribution analysis indicated that SMN and VIS regions contributed most strongly to `graphTRIP`'s outcome predictions. Consistent with this, GRAIL linked stronger mean FC in VIS and SMN regions to worse outcomes overall. This finding also aligns with recent work, reporting that baseline connectivity patterns involving VIS (and to a lesser extent SMN) networks are most informative for predicting early response to psilocybin [25].

One possible interpretation is that these biomarkers relate to individual differences in the brain's functional cortical hierarchy. A prominent view posits that cortical organisation follows a unimodal-transmodal gradient, ranging from sensory systems to higher-order association cortices [38], and that psychedelic treatment may act in part by modifying or flattening this hierarchical organisation [45–47]. In this context, stronger baseline coupling within unimodal systems such as VIS and SMN could reflect a cortical organisation that is already relatively shifted towards low-level sensory coupling, leaving less leverage for therapeutically beneficial reconfiguration. Conversely, higher modularity and stronger coupling of regions associated with high normative 5-HT2A density – which tend to be located higher along the cortical hierarchy – may reflect a baseline organisation that supports better response to serotonergic interventions.

Together, these results suggest that baseline cortical hierarchy may contribute to treatment responsiveness in a manner that is partly shared across psilocybin and escitalopram. However, note that a drug-invariant baseline predictor does not imply that both treatments modify cortical hierarchy in the same way; indeed, recent evidence points to distinct hierarchical reconfigurations under psilocybin and escitalopram [27].

Treatment-specific mechanisms of treatment responsiveness

Our analyses further suggested that treatment response heterogeneity is most strongly expressed in higher-order association systems. In particular, regional attribution analysis indicated that DMN and FPN regions contributed most to `Medusa-graphTRIP`'s ITE predictions. Additionally, GRAIL identified mean FPN REACT values for 5-HT1A and 5-HT2A as differential biomarkers, predicting better response to escitalopram but poorer response to psilocybin.

This pattern aligns with extensive literature implicating the FPN and DMN in depression and antidepressant response [16, 47–49], and is also consistent with the view that psilocybin and SSRIs modulate serotonergic signalling through distinct mechanisms: Chronic SSRI use is thought to elevate cortical serotonin levels, which may bias signalling towards post-synaptic 5-HT1A-mediated inhibition and thereby reduce layer-V pyramidal cell excitability, particularly in higher-order association cortices where 5-HT1A is densely expressed [4, 50]. Such 5-HT1A-linked effects have been proposed to contribute to core therapeutic aspects of SSRI treatment, including improved stress regulation and emotional blunting [50]. In contrast, psilocybin acts as a direct agonist of excitatory 5-HT2A receptors, which are often co-expressed on the same layer-V pyramidal neurons that also express 5-HT1A [50]. A prominent hypothesis is that acute 5-HT2A activation transiently increases pyramidal cell excitability and initiates downstream plasticity-related processes that support longer-lasting therapeutic effects [4].

Taken together, while the antidepressant mechanisms of both drugs remain incompletely understood, converging evidence suggests that they may modulate pyramidal cell excitability in association cortices via opposing 5-HT1A- and 5-HT2A-mediated effects. This aligns with our finding that FPN serotonergic coupling differentially predicts response to escitalopram versus psilocybin.

Limitations and future work

One potential source of confusion lies in the use of two separate models to predict treatment outcome (`graphTRIP`) and treatment effects (`Medusa-graphTRIP`). 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. This can occur when a treatment-invariant predictive signal dominates the data, such that the model can achieve strong performance without learning additional treatment-specific features. In such settings, meta-learning theory suggests that models with separate predictors for each treatment arm can be better suited to capture heterogeneous treatment effects, albeit at the cost of higher variance when per-arm sample sizes are small [36]. `Medusa-graphTRIP` was therefore introduced to explicitly model treatment-specific outcome functions.

Importantly, `graphTRIP` remains valuable in our setting because its single prediction head is trained on outcomes from both arms, providing a more stable estimator of overall treatment response and enabling the identification of drug-invariant biomarkers. Analysing interpretability results across both models allowed us to distinguish shared biomarkers from drug-specific modera-

tors in a conservative manner – that is, supported by convergent findings across architectures with different inductive biases. With larger samples, this two-model strategy may no longer be necessary, and either `graphTRIP` or `Medusa-graphTRIP` alone may be sufficient.

Estimating causal treatment effects from observed data is inherently challenging because it requires reasoning about counterfactual outcomes which are never directly observed. Consequently, causal conclusions are only as reliable as the assumptions under which they are derived. One key assumption is ignorability, i.e., that treatment assignment is independent of potential outcomes conditional on baseline covariates [32]. While RCTs are designed to approximate this condition, small-sample effects can still introduce subtle imbalances. In our data, treatment assignment was modestly predictable from pretreatment variables, primarily due to slightly higher baseline QIDS in the psilocybin group (Supp. Fig. 12b,c). However, `Medusa-graphTRIP`'s ITE estimates were only weakly sensitive to baseline QIDS (Supp. Fig. 12d), suggesting that this imbalance is unlikely to be a major driver of our counterfactual predictions. Nevertheless, confirming treatment-specific biomarkers and improving the reliability of ITE estimation will require larger RCTs, ideally including comparisons across multiple antidepressant interventions. This would be a major step towards truly advancing personalised treatment of depression.

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. 14b). Furthermore, 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 [51–56]. 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, 57, 58].

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, enables patient-specific biomarker discovery, and – via its causal extension – can estimate which treatment is likely to work best for each patient. Our analysis highlights the role of unimodal brain regions in early treatment response to both escitalopram and psilocybin, and reveals serotonergic signalling in association networks as a treatment-specific moderator. 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 [47].

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 = 41.9 $\pm$ 11.0 SD years; 8 female) and 20 in the escitalopram arm (mean age = 38.7 $\pm$ 11.0 SD 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 $\pm$ 2.6 SD 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 $\pm$ 10.15 SD 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. [34] 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 [35]. 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 [47].

Functional connectivity

BOLD time series were parcellated and z-scored using the Schaefer atlases with 100 or 200 parcels [34], and the AAL atlas [35]. Functional connectivity (FC) was then computed as the Pearson correlation coefficient between the z-scored time series of each pair of brain regions, using Nilearn's `ConnectivityMeasure` function (`kind='correlation'`). This resulted in an $N \times N$ FC matrix for each participant and parcellation.

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 maps were parcellated to obtain one value for each brain region, participant, and atlas.

Statistical analysis

Statistical analyses were performed in Python (version 3.12). Pearson correlations were used to evaluate prediction accuracy (true vs. predicted values). Cohen's d was calculated as the difference in group means divided by the pooled standard deviation for two-sample comparisons, and as the mean divided by the standard deviation for one-sample comparisons. All p-values are two-sided. Where indicated, p-values were corrected for multiple comparisons using false discovery rate (FDR) correction.

Model and training configurations

Full hyperparameter configurations of all models, including architectural details and training parameters are provided in Supplementary Tables II-III. Additionally, 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 QIDS scores based on these latent representations, the drug condition, 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 3-dimensional feature vector, encoding the REACT values for 5-HT1A, 5-HT2A, and 5-HTT. Additionally, three conditional node features encode the 3D spatial coordinates of each brain region in MNI space. These conditional features provide the model with anatomical information about regional identity, while preserving permutation invariance of the graph representation. Since these features are normative and do not contain patient-specific information, they are used only by the encoder and are not reconstructed by the decoder. Edges exist between regions if the absolute FC value is among the global top 20% strongest connections to obtain a fixed connection density of $\rho = 0.2$ for each brain graph, and edge attributes are the original (non-thresholded) FC values. This thresholding approach was shown to yield robust and reliable network topology [31]. Self-connections were set to 1.

VGAE architecture. The VGAE consists of a node-level encoder and decoder and a graph-level readout layer. We will describe each component in turn.

Encoder. The encoder is based on a Graphormer-style architecture [59], which extends the transformer architecture to graph-structured data. Each node's feature vector x_i is first concatenated with its conditional node features c_i (the MNI coordinates) and projected into a shared embedding space as

$$h_i^{(0)} = W_{\text{in}}[x_i \,\|\, c_i] \ , \tag{1}$$

where W_{in} is a learnable linear projection.

The model then applies three Graphormer layers. At each layer ℓ, node embeddings are updated using multi-head self-attention with two learned attention biases, derived from shortest-path distances between nodes and edge attributes (the FC weights). Specifically, the shortest-path bias b_{ij}^{spd} is obtained by embedding the discretised shortest-path distance between nodes i and j, while the edge bias b_{ij}^{edge} is computed by passing the corresponding FC value through a two-layer MLP with ReLU hidden-layer activation. Both biases are added to the attention scores e_{ij} between nodes i and j in a head-specific manner as follows.

$$e_{ij}^{(\ell,h)} = \frac{(Q_i^{(\ell,h)})^\top K_j^{(\ell,h)}}{\sqrt{d_h}} + b_{ij,h}^{\text{spd}} + b_{ij,h}^{\text{edge}} \tag{2}$$

Here, h indexes attention heads, $Q_i^{(\ell,h)}$ and $K_j^{(\ell,h)}$ are linear projections of the node embeddings, and d_h is the head dimension.

Attention weights are obtained via softmax normalisation over all nodes within the same graph,

$$\alpha_{ij}^{(\ell,h)} = \frac{\exp(e_{ij}^{(\ell,h)})}{\sum_k \exp(e_{ik}^{(\ell,h)})} \ . \tag{3}$$

After applying a dropout layer, the normalised scores are used to compute the attention output as

$$\hat{h}_i^{(\ell)} = \sum_h \sum_j \alpha_{ij}^{(\ell,h)} V_j^{(\ell,h)} \ . \tag{4}$$

where $V_j^{(\ell,h)}$ is another linear projection of the node embeddings.

The outputs from all attention heads are then linearly projected to allow mixing of information across heads,

$$\tilde{h}_i^{(\ell)} = W_{\text{attn}} \, \hat{h}_i^{(\ell)} \ , \tag{5}$$

where W_{attn} is a learnable projection matrix.

Each Graphormer layer further includes a position-wise feedforward network with residual connections, layer normalisation, and dropout. Concretely, the layer update takes the form

$$u_i^{(\ell)} = \text{LN}\Big(h_i^{(\ell)} + \text{Dropout}\Big(\tilde{h}_i^{(\ell)}\Big)\Big) \ , \tag{6}$$

$$h_i^{(\ell+1)} = \text{LN}\Big(u_i^{(\ell)} + \text{Dropout}\Big(\Phi_{\text{ff}}(u_i^{(\ell)})\Big)\Big) \ , \tag{7}$$

where $\text{LN}(\cdot)$ denotes layer normalisation and $\Phi_{\text{ff}}(\cdot)$ is a two-layer MLP with ReLU activation in the hidden layer.

After the final Graphormer layer, node embeddings are projected to a lower-dimensional space,

$$h_i = W_{\text{out}}\, h_i^{(L)} \ . \tag{8}$$

A final linear layer maps each node embedding h_i to a Gaussian distribution parameterised by a mean μ_i and log-variance $\log \sigma_i^2$,

$$(\mu_i, \log \sigma_i^2) = W_{\text{enc}} h_i \ . \tag{9}$$

Latent node representations z_i are then sampled using the reparameterisation trick,

$$z_i = \mu_i + \sigma_i \odot \epsilon \ , \tag{10}$$

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

Decoder. The VGAE decoder consists of three independent sub-modules:

- **Node decoder:** A 3-layer MLP with LeakyReLU activation and dropout, and a linear output layer to reconstruct node features from z_i.
- **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 readout layer aggregates the latent node vectors $\{z_i\}$ into graph-level representations. Before aggregating, it first applies a single multi-head self-attention block to the latent node vectors to allow each node representation to integrate information from other regions. This block works analogously to the encoder attention mechanism described above, but without the structural attention biases, and – again, just like in the encoder – it is followed by a position-wise feedforward network (a two-layer MLP) with residual connections, layer normalisation, and dropout.

The resulting contextualised node embeddings are then aggregated using mean pooling to obtain a single graph-level representation,

$$z = \frac{1}{N} \sum_{i=1}^{N} \tilde{z}_i \ , \tag{11}$$

where $\tilde{z}_i$ denotes the contextualised latent embedding of node i and N is the number of nodes in the graph.

MLP prediction head. The input to the MLP prediction head consists of a concatenation of the graph representation vector z and clinical data: baseline BDI and QIDS 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 three layers with LeakyReLU activation and dropout in the hidden layers (see Tab. II), and a linear output layer to predict post-treatment QIDS scores.

`Medusa-graphTRIP` model

To estimate individual treatment effects (ITEs), we developed `Medusa-graphTRIP`, a causal extension of `graphTRIP` with separate prediction heads for each treatment condition. The model is inspired by the counterfactual regression framework introduced by Shalit et al. [32].

`Medusa-graphTRIP` employs a single VGAE to learn a shared latent brain-graph representation across patients from both treatment groups. The VGAE architecture is identical to that used in `graphTRIP`, with the exception of the readout layer (Tab. III). Specifically, `Medusa-graphTRIP` uses a simple mean-standard deviation pooling, where the per-graph representation is formed by concatenating the mean and standard deviation of latent node vectors. We chose this less expressive pooling module to ensure that the two treatment-specific prediction heads receive statistically comparable latent representations.

Downstream of the shared VGAE, `Medusa-graphTRIP` includes two MLP prediction heads, each trained exclusively on patients from a single treatment condition. As in `graphTRIP`, the MLP heads receive the latent brain-graph representation together with baseline clinical data (baseline QIDS, baseline BDI, and SSRI history) and are trained to predict post-treatment QIDS scores. Treatment condition is not provided as an explicit input to the MLPs, but is used solely to route each patient's latent representation to the appropriate treatment-specific head during training.

Following Shalit et al. [32], individual treatment effects for each patient i were estimated as the difference between the predictions of the two heads,

$$\tilde{D}(x_i) = \mu_1(x_i) - \mu_0(x_i) \ , \tag{12}$$

where $\tilde{D}(x_i)$ denotes the estimated ITE, x_i represents the patient's covariates, and μ_0 and μ_1 are the prediction

heads trained on escitalopram and psilocybin patients, respectively. Thus, μ_0 estimates response under escitalopram, while μ_1 estimates response under psilocybin.

Model training

Unless stated otherwise, `graphTRIP` and `Medusa-graphTRIP` were trained using identical training protocols. All models were implemented in PyTorch and trained using the Adam optimiser with a learning rate of 0.001.

Training objective. The VGAE and downstream MLP(s) were trained jointly to optimise both graph reconstruction and outcome prediction. For `graphTRIP`, the total training objective was defined as a weighted combination of the VGAE loss and the MLP loss,

$$\mathcal{L}_{\text{graphTRIP}} = \alpha \mathcal{L}_{\text{VGAE}} + (1-\alpha)\mathcal{L}_{\text{MLP}} \ , \tag{13}$$

with $\alpha = 0.5$.

For `Medusa-graphTRIP`, an additional maximum mean discrepancy (MMD) penalty was used to encourage similar latent representations across treatment groups, as recommended by Shalit et al. [32]:

$$\mathcal{L}_{\text{Medusa}} = \alpha \mathcal{L}_{\text{VGAE}} + (1-\alpha)\mathcal{L}_{\text{MLP}} + \beta \mathcal{L}_{\text{MMD}} \ , \tag{14}$$

where $\beta = 1$ controls the strength of the MMD penalty. The MMD was computed between latent representations of patients receiving psilocybin and escitalopram using a Gaussian radial basis function kernel with bandwidth estimated from the data.

The VGAE loss comprises a reconstruction term, a Kullback-Leibler (KL) divergence term, and an L2 regularisation term on the VGAE parameters,

$$\mathcal{L}_{\text{VGAE}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{reg}} \tag{15}$$

$$= \sum_{i=1}^{N} \|x_i - \hat{x}_i\|_2^2 + \sum_{(i,j)\in\mathcal{E}} \|e_{ij} - \hat{e}_{ij}\|_2^2 \tag{16}$$

$$+ \mathbb{E}_{q(z|X)}[\text{KL}(q(z|X) \,\|\, p(z))] + \lambda \sum_k W_{\text{VGAE},k}^2 \ , \tag{17}$$

where x_i and $\hat{x}_i$ denote the original and reconstructed node features, e_{ij} and $\hat{e}_{ij}$ the original and reconstructed edge attributes (FC values), N is the number of nodes, and $\mathcal{E}$ denotes the set of unique (upper-triangular) edges. The final term corresponds to standard weight-decay regularisation over all VGAE parameters W_{VGAE} with $\lambda = 0.01$.

The MLP loss consists of a mean squared error prediction term and an L2 weight-decay regularisation term,

$$\mathcal{L}_{\text{MLP}} = \sum_{i=1}^{n} \|y_i - \hat{y}_i\|_2^2 + \lambda \sum_k W_{\text{MLP},k}^2 \ , \tag{18}$$

where y_i and $\hat{y}_i$ are the true and predicted post-treatment QIDS scores for the i-th patient in a batch of size n, $\lambda = 0.01$ is the regularisation strength, and W_{MLP} denotes the MLP parameters.

Training protocol. All models were trained for 300 epochs using k-fold cross-validation (CV). The number of folds and batch size were chosen based on dataset size, with the aim of obtaining approximately balanced test folds and batches (7 folds; batch size 9 for the primary dataset and 7 for the validation dataset). For the primary dataset, CV splits were stratified by treatment condition to maintain balanced proportions of psilocybin and escitalopram patients in each fold.

Latent-sample augmentation was used only for `Medusa-graphTRIP`. Specifically, during training, the MLP prediction heads received three independently sampled latent representations per patient. The rationale behind this was to slightly augment the effective training data for the treatment-specific MLP heads, which are trained on fewer samples compared to `graphTRIP`'s single MLP head due to treatment stratification.

Predictions and reconstructions for each patient were always generated using the CV fold in which that patient was assigned to the test set. Each full training procedure was repeated ten times with different random seeds. Unless stated otherwise, all reported predictions correspond to mean test-set predictions averaged across these ten runs.

Fine-tuning

To evaluate transfer of `graphTRIP` to an independent validation dataset ($n = 16$), we first pretrained new MLP prediction heads on the primary dataset using only inputs compatible with the validation data, and subsequently fine-tuned these models on the validation dataset.

Pretraining transferable MLP heads. For each random seed (0-9), we loaded the set of seven VGAEs obtained from the corresponding 7-fold CV run and froze their weights. For each frozen VGAE, we initialised a new MLP prediction head with the same architecture and training configuration as in the core model. These MLPs were trained on the primary dataset using the VGAE latent graph representations and baseline QIDS and BDI scores as inputs, but excluding drug condition and SSRI history. Each MLP was trained using the same loss function and optimisation protocol as during core-model training. This procedure yielded a set of seven pretrained VGAE–MLP models per random seed, whose inputs were fully compatible with the validation dataset.

Fine-tuning on the validation dataset. The pretrained VGAE–MLP models were then transferred to the validation dataset for fine-tuning. The VGAE weights were kept frozen and only the MLP parameters were updated. Fine-tuning was performed using 7-fold CV, with a batch size of 7 to ensure balanced batches. The MLPs were fine-tuned for 300 epochs using the same loss

function, learning rate (lr = 0.001), L_2 regularisation strength ($\lambda = 0.01$), and dropout rate (0.25) as during core-model training.

To reduce sensitivity to initialisation and fold assignment, each CV fold was fine-tuned seven times, each time initialised from a different pretrained VGAE–MLP CV-fold model. This resulted in seven test predictions per patient in the validation dataset for each random seed, which were averaged to obtain a mean test prediction for that seed. As in the main analyses, final validation predictions were obtained by averaging these mean test predictions across random seeds.

Gradient Alignment for Interpreting Latent-variable Models (GRAIL)

The GRAIL method was developed to estimate learned associations between candidate brain-graph-derived biomarkers and model predictions (e.g. post-treatment QIDS or ITEs). Positive alignment indicates that higher values of a biomarker are associated with a higher prediction value, while negative alignment indicates the opposite. An alignment of zero indicates no association between the biomarker and predicted treatment response.

GRAIL leverages the two-part architecture of the `graphTRIP` and `Medusa-graphTRIP` models, consisting of a VGAE and an MLP prediction head. The VGAE encoder maps input brain graphs to latent representations z, which serve as inputs to both the VGAE decoder (reconstructing brain graphs) and the MLP (predicting post-treatment QIDS or ITEs). Biomarkers of interest (e.g., mean FC within an RSN) are computed from reconstructed graphs. The only requirement for a biomarker is that it can be expressed as a differentiable function of the reconstructed brain graph.

GRAIL computes the latent gradients of the MLP prediction $\hat{y}$ and the candidate biomarker b, i.e., $\nabla_z \hat{y}$ and the $\nabla_z b$. These gradients describe directions in latent space that locally increase the predicted outcome and the biomarker value, respectively. Alignment between prediction and biomarker is quantified as the cosine similarity between the gradients,

$$\text{Alignment}(b) = \frac{\nabla_z \hat{y} \cdot \nabla_z b}{\|\nabla_z \hat{y}\| \|\nabla_z b\|} \ , \tag{19}$$

which ranges from -1 (maximal negative alignment) to 1 (maximal positive alignment).

A first-stage screening ensured that only biomarkers with stable local associations around the patient's latent representation were considered: For each patient and trained model, we generated 25 samples from a Gaussian distribution centred at the patient-specific latent mean. For each latent sample, we computed the predicted outcome (or ITE), all candidate biomarkers, their respective gradients, and the corresponding alignment values. For each biomarker, we then performed a one-sample t-test across the 25 latent samples to test whether alignment values were significantly different from zero. Biomarkers with a Cohen's $d > 0.5$ and FDR-corrected $p < 0.05$ were retained for further analysis. We stored the mean alignment across latent samples for all biomarkers.

For biomarkers passing the first-stage screening, statistical significance was assessed using spatially informed null models. All candidate biomarkers were derived from either regional assignments to RSNs, or normative molecular target density maps. Accordingly, surrogate biomarkers were generated, for each latent sample, using 1000 spatial-autocorrelation preserving spin permutations of RSN assignments or normative density maps. Gradient alignments were computed for each surrogate biomarker in the same manner as for the observed biomarker and then averaged across latent samples. This yielded a null distribution of mean alignment values for each candidate biomarker. Observed mean alignment values were compared against these null distributions, and biomarkers were deemed significant if their alignment exceeded the null distribution after FDR correction at $p < 0.05$.

The above procedure was applied independently to all models trained with different random seeds and CV folds. For each patient and biomarker, this resulted in a binary indicator of whether the biomarker was found to be significantly aligned in a given model. To identify biomarkers that were robustly associated with predictions across model initializations, we tested whether the number of significant detections exceeded chance expectation. Specifically, for each patient and biomarker, we computed the probability of observing k or more significant detections out of n trained models under a binomial model,

$$P(X \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}, \tag{20}$$

where p denotes the nominal false-positive rate. Biomarkers with low tail probabilities were interpreted as robustly aligned across random initializations for a particular patient. For reporting alignment values, we computed performance-weighted mean alignments across models, with weights proportional to test-fold prediction accuracy (true-vs-predicted Spearman correlation).

Biomarker categories

To distinguish shared versus drug-specific biomarkers of treatment responsiveness, we applied GRAIL in four complementary modes: 1) `graphTRIP` (shared outcome prediction across both treatment arms); 2) the psilocybin head of `Medusa-graphTRIP`; 3) the escitalopram head of `Medusa-graphTRIP`; and 4) `Medusa-graphTRIP`'s ITE predictions. For outcome-prediction models (1-3), positive alignment indicates biomarkers associated with higher predicted post-treatment QIDS (treatment resis-

tance), whereas negative alignment indicates biomarkers associated with lower predicted post-treatment QIDS (treatment responsiveness). For the ITE mode (4), positive alignment indicates biomarkers associated with relatively better predicted response to escitalopram, whereas negative alignment indicates relatively better predicted response to psilocybin.

For each patient, we assigned each biomarker to a category based on the combination of significant alignment directions across the four modes (Table I). Biomarkers whose alignment patterns did not match any valid combination were considered invalid and remained unassigned. For downstream analyses, we retained biomarkers that were assigned a valid category in at least 50% of patients.

Regional attribution analysis

To quantify the contribution of individual brain regions to model predictions, we performed a regional attribution analysis based on gradients in latent space. For each patient, we computed the gradient of the model prediction with respect to the latent node representations of the brain graph. This analysis was applied both to predicted post-treatment QIDS scores (using `graphTRIP`) and to ITE estimates (using `Medusa-graphTRIP`). The resulting region-specific gradient vectors quantify the local sensitivity of the prediction to perturbations in each region's latent representation.

Regional attribution scores were computed by first normalising the full gradient vector (obtained by concatenating all node-wise gradients) to unit Euclidean norm, and then computing, for each region, the sum of squares of its partial gradient components. This yields non-negative regional attribution scores that sum to one (or to 100 after rescaling) and can be interpreted as the percentage contribution of each region to the total gradient energy. Under small Gaussian perturbations of the latent variables, this quantity is proportional to the expected contribution of each region to the variance of the predicted outcome, consistent with established gradient-based attribution methods [37].

To obtain robust patient-level estimates, we accounted for both latent sampling variability and model stochasticity. For each patient, we sampled 100 latent vectors from a Gaussian distribution centred at the patient-specific latent mean, computed regional attribution vectors for each sample, and averaged these to obtain a stable attribution profile. We then aggregated attributions across all CV-fold models and random seeds using a performance-weighted average, with weights proportional to test-fold prediction accuracy (true–vs-predicted Spearman correlation). As a sanity check, attribution profiles were highly consistent across folds and seeds for the same patient (Supp. Fig. 15).

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.

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. Patient-level clinical data are subject to patient confidentiality and data sharing restrictions. A minimal dataset sufficient to reproduce core figures and results is available in the supplementary materials. The normative receptor density maps from in vivo PET used in this study are publicly available: serotonin receptor and transporter maps can be accessed at https://nru.dk/index.php/allcategories/category/90-nru-serotonin-atlas-and-clustering; all other receptor and transporter maps 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

COMPETING INTERESTS

R.L.C.-H. is a scientific advisor to TRYP Therapeutics, MindState Design Labs, and Red Light Holland. D.N. has received lecture fees from Takeda, Lundbeck, Otsuka, and Janssen, and consulting fees from Algernon, Beckley Psytech, Leith Pharma, and Amitis Partners. He is a director of GABA Labs and Chief Research Officer at Solvonis (formerly Awaknlifesciences), and holds shares/options in PsychedWellness and Neurotherapeutics. D.N. is also Head of the Imperial College Centre for Psychedelic Research, which has received in-kind support from COMPASS Pathways and USONA (psilocybin), and Beckley Psytech (5-MeO-DMT). All other authors declare no competing interests.

TABLE I. Valid biomarker category assignments based on GRAIL alignment direction across four modes. '–' indicates significant negative alignment, '+' indicates significant positive alignment, and '0' indicates non-significant alignment.

Shared	Psilocybin	Escitalopram	ITE	Biomarker category
–	–	–	0	Shared responsiveness
–	–	–	+	Shared responsiveness
–	–	–	–	Shared responsiveness
+	+	+	0	Shared resistance
+	+	+	+	Shared resistance
+	+	+	–	Shared resistance
–	0	–	+	Escitalopram responsiveness
0	0	–	+	Escitalopram responsiveness
–	–	0	–	Psilocybin responsiveness
0	–	0	–	Psilocybin responsiveness
+	0	+	–	Escitalopram resistance
0	0	+	–	Escitalopram resistance
+	+	0	+	Psilocybin resistance
0	+	0	+	Psilocybin resistance
0	+	–	+	Escitalopram responsiveness & psilocybin resistance
0	–	+	–	Psilocybin responsiveness & escitalopram resistance

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). L.R. is supported by the Huxley Foundation. 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, Cross-national epidemiology of dsm-iv major depressive episode, BMC Medicine **9**, 90 (2011).

[2] S. Shorey, E. D. Ng, and C. H. J. Wong, Global prevalence of depression and elevated depressive symptoms among adolescents: A systematic review and meta-analysis., 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, Major depressive disorder: hypothesis, mechanism, prevention and treatment, Signal Transduction and Targeted Therapy **9** (2024).

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

[5] A. Rush, Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A star*d report, 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, Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis, 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, Psilocybin-assisted therapy for depression: A systematic review and meta-analysis, 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, Psilocybin for treatment-resistant depression: fmri-measured brain mechanisms, 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, Single-dose psilocybin for a treatment-resistant episode of major depression., The New England journal of medicine **387 18**, 1637 (2022).

[10] T. Watford and N. Masood, Psilocybin, an effective treatment for major depressive disorder in adults - a systematic review, Clinical Psychopharmacology and Neuroscience **22**, 2–12 (2023).

[11] R. L. Carhart-Harris and G. M. Goodwin, The therapeutic potential of psychedelic drugs: Past, present, and future, 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, Major depressive disorder, 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, Toward mapping neurobehavioral heterogeneity of psychedelic neurobiology in humans, Biological Psychiatry **93**, 1061 (2023).

[14] E. T. Bullmore and O. Sporns, Complex brain networks: graph theoretical analysis of structural and functional systems, Nature Reviews Neuroscience **10**, 186 (2009).

[15] M. P. van den Heuvel and O. Sporns, A cross-disorder connectome landscape of brain dysconnectivity, 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, Altered resting-state functional connectivity of multiple networks and disrupted correlation with executive function in major depressive disorder, Frontiers in Neurology **11** (2020).

[17] R. H. Kaiser, J. R. Andrews-Hanna, T. D. Wager, and D. A. Pizzagalli, Large-scale network dysfunction in major depressive disorder: A meta-analysis of resting-state functional connectivity., JAMA psychiatry **72 6**, 603 (2015).

[18] O. Dipasquale, P. Selvaggi, M. Veronese, A. S. Gabay, F. E. Turkheimer, and M. A. Mehta, Receptor-enriched analysis of functional connectivity by targets (react): A novel, multimodal analytical approach informed by pet to study the pharmacodynamic response of the brain under mdma, Neuroimage **195**, 252 (2019).

[19] T. Lawn, M. A. Howard, F. E. Turkheimer, B. Mišić, G. Deco, D. F. D. Martins, and O. Dipasquale, From neurotransmitters to networks: Transcending organisational hierarchies with molecular-informed functional imaging, 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, Mapping neurotransmitter systems to the structural and functional organization of the human neocortex, 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.*, In vivo mapping of pharmacologically induced functional reorganization onto the human brain's neurotransmitter landscape, Science advances **9**, eadf8332 (2023).

[22] T. Lawn, A. Giacomel, D. Martins, M. Veronese, M. Howard, F. E. Turkheimer, and O. Dipasquale, Normative modelling of molecular-based functional circuits captures clinical heterogeneity transdiagnostically in psychiatric patients, Communications Biology **7** (2024).

[23] D. Martins, M. Veronese, F. E. Turkheimer, M. A. Howard, S. C. Williams, and O. Dipasquale, A candidate neuroimaging biomarker for detection of neurotransmission-related functional alterations and prediction of pharmacological analgesic response in chronic pain, Brain Communications **4**, 10.1093/braincomms/fcab302 (2021).

[24] T. Lawn, O. Dipasquale, A. Vamvakas, I. Tsougos, M. A. Mehta, and M. A. Howard, Differential contributions of serotonergic and dopaminergic functional connectivity to the phenomenology of lsd, Psychopharmacology **239**, 1797 (2022).

[25] D. Copa, D. Erritzoe, B. Giribaldi, D. Nutt, R. Carhart-Harris, and E. Tagliazucchi, Predicting the outcome of psilocybin treatment for depression from baseline fmri functional connectivity, 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, Whole-brain turbulent dynamics predict responsiveness to pharmacological treatment in major depressive disorder, Molecular Psychiatry **30**, 1069–1079 (2024).

[27] G. Deco, Y. S. Perl, S. Johnson, N. Bourke, R. L. Carhart-Harris, and M. L. Kringelbach, Different hierarchical reconfigurations in the brain by psilocybin and escitalopram for depression, Nature Mental Health (2024).

[28] R. Liu, X. Hou, S. Liu, Y. Zhou, J. Zhou, K. Qiao, H. Qi, R. Li, Z. Yang, L. Zhang, J. Cui, C. Jin, A. Yu, and G. Wang, Predicting antidepressant response via local-global graph neural network and neuroimaging biomarkers, NPJ Digital Medicine **8** (2025).

[29] 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, Cross-trial prediction of treatment outcome in depression: a machine learning approach., The lancet. Psychiatry **3 3**, 243 (2016).

[30] S. B. Eickhoff, B. T. T. Yeo, and S. Genon, Imaging-based parcellations of the human brain, Nature Reviews Neuroscience **19**, 672 (2018).

[31] 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.*, Systematic evaluation of fmri data-processing pipelines for consistent functional connectomics, Nature Communications **15**, 4745 (2024).

[32] U. Shalit, F. D. Johansson, and D. A. Sontag, Estimating individual treatment effect: generalization bounds and algorithms, ArXiv **abs/1606.03976** (2017).

[33] 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, Trial of psilocybin versus escitalopram for depression., The New England journal of medicine **384 15**, 1402 (2021).

[34] A. Schaefer, R. Kong, E. M. Gordon, T. O. Laumann, X.-N. Zuo, A. J. Holmes, S. B. Eickhoff, and B. T. T. Yeo, Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri, Cerebral Cortex (2018).

[35] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain, NeuroImage **15**, 273 (2002).

[36] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, Metalearners for estimating heterogeneous treatment effects using machine learning, Proceedings of the National Academy of Sciences of the United States of America **116**, 4156 (2019).

[37] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, Smoothgrad: removing noise by adding noise, ArXiv **abs/1706.03825** (2017).

[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, Situating the default-mode network along a principal gradient of macroscale cortical organization, 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, The natural axis of transmitter receptor distribution in the human cerebral cortex, 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, A synergistic core for human brain evolution and cognition, 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.*, Neurodevelopment of the association cortices: Patterns, mechanisms, and implications for psychopathology, 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, A high-resolution in vivo atlas of the human brain's serotonin system, The Journal of Neuroscience **37**, 120 (2017).

[43] F. Váša and B. Mišić, Null models in network neuroscience, 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, The organization of the human cerebral cortex estimated by intrinsic functional connectivity., Journal of neurophysiology **106 3**, 1125 (2011).

[45] R. L. Carhart-Harris and K. J. Friston, Rebus and the anarchic brain: Toward a unified model of the brain action of psychedelics, Pharmacological Reviews **71**, 316 (2019).

[46] M. Girn, L. Roseman, B. C. Bernhardt, J. Smallwood, R. L. Carhart-Harris, and N. Spreng, Serotonergic psychedelic drugs lsd and psilocybin reduce the hierarchical differentiation of unimodal and transmodal cortex, NeuroImage **256** (2022).

[47] 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, Increased global integration in the brain after psilocybin therapy for depression, Nature Medicine **28**, 844 (2022).

[48] C. J. Lynch, I. G. Elbau, T. H. Ng, A. Ayaz, S. Zhu, D. L. Wolk, N. Manfredi, M. Johnson, M. Chang, J. Chou, I. Summerville, C. Ho, M. Lueckel, H. Bukhari, D. M. Buchanan, L. W. Victoria, N. Solomonov, E. Goldwaser, S. Moia, C. Caballero-Gaudes, J. Downar, F. Vila-Rodriguez, Z. J. Daskalakis, D. M. Blumberger, K. Kay, A. Aloysi, E. M. Gordon, M. T. Bhati, N. R. Williams, J. D. Power, B. Zebley, L. Grosenick, F. M. Gunning, and C. Liston, Frontostriatal salience network expansion in individuals in depression, Nature **633**, 624 (2024).

[49] A. Scalabrini, B. Vai, S. Poletti, S. Damiani, C. Mucci, C. Colombo, R. Zanardi, F. Benedetti, and G. Northoff, All roads lead to the default-mode network—global source of dmn abnormalities in major depressive disorder, Neuropsychopharmacology **45**, 2058 (2020).

[50] R. Carhart-Harris and D. J. Nutt, Serotonin and brain function: a tale of two receptors, Journal of Psychopharmacology (Oxford, England) **31**, 1091 (2017).

[51] R. Cofré, R. Herzog, P. A. M. Mediano, J. I. Piccinini, F. E. Rosas, Y. S. Perl, and E. Tagliazucchi, Whole-brain models to explore altered states of consciousness from the bottom up, Brain Sciences **10** (2020).

[52] G. Deco, J. Cruzat, J. Cabral, G. M. Knudsen, R. L. Carhart-Harris, P. C. Whybrow, N. K. Logothetis, and M. L. Kringelbach, Whole-brain multimodal neuroimaging model using serotonin receptor maps explains non-linear functional effects of lsd, Current biology **28**, 3065 (2018).

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

[54] J. B. Burt, K. H. Preller, M. Demirtas, J. L. Ji, J. H. Krystal, F. X. Vollenweider, A. Anticevic, and J. D. Murray, Transcriptomics-informed large-scale cortical model captures topography of pharmacological neuroimaging effects of lsd, Elife **10**, e69320 (2021).

[55] A. I. Luppi, S. P. Singleton, J. Y. Hansen, K. W. Jamison, D. Bzdok, A. Kuceyeski, R. F. Betzel, and B. Misic, Contributions of network structure, chemoarchitecture and diagnostic categories to transitions between cognitive topographies, Nature Biomedical Engineering **8**, 1142 (2024).

[56] 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.*, An engine for systematic discovery of cause-effect relationships between brain structure and function, bioRxiv , 2025 (2025).

[57] R. van den Bosch and R. Cools, Validating molecular target-enriched fmri for disentangling drug effects on dopamine, bioRxiv (2025).

[58] T. Lawn and M. A. Mehta, Towards molecular specificity in pharmacoimaging, PsyArXiv (2025), commentary on van den Bosch and Cools (2025).
[59] C. Ying, T. Cai, S. Luo, S. Zheng, G. Ke, D. He, Y. Shen, and T.-Y. Liu, Do transformers really perform bad for graph representation?, ArXiv **abs/2106.05234** (2021).

SUPPLEMENTARY

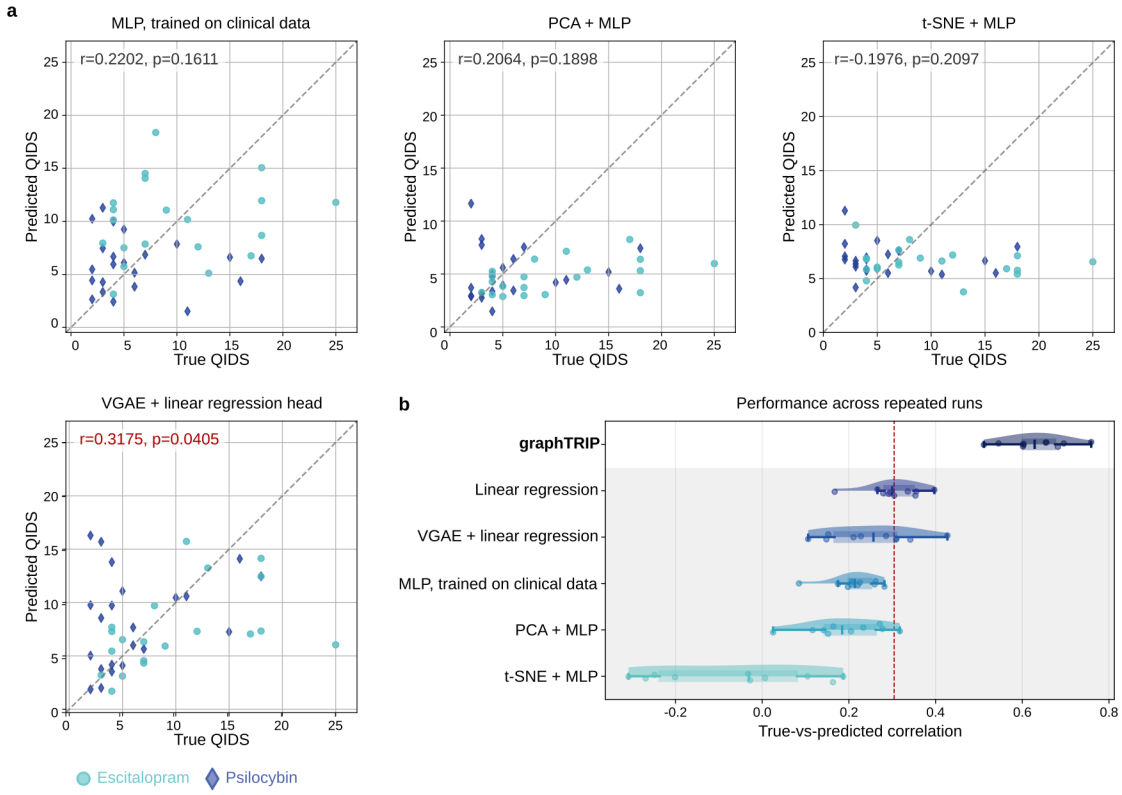

FIG. 7. **Benchmark models for predicting post-treatment QIDS. a**, True versus predicted post-treatment QIDS scores for four benchmark models evaluated with the same 7-fold cross-validation protocol as `graphTRIP`. (1) MLP, trained on clinical data: an MLP with the same architecture as the `graphTRIP` prediction head, trained only on baseline clinical data (baseline QIDS, baseline BDI, SSRI history) and treatment condition (baseline scores standardised). (2-3) PCA+t-SNE MLPs: brain graphs were constructed as in `graphTRIP`, but FC matrices were not thresholded to ensure homogeneous input across patients; node features and the upper triangle of the unthresholded FC matrix were flattened, concatenated, standardised across patients, and reduced using PCA (32 components) or t-SNE (perplexity = 30) before training the MLP. The MLP had the same architecture as in `graphTRIP`. (4) VGAE+linear regression head: identical to `graphTRIP` except that the MLP prediction head was replaced by a single linear output layer. **b**, Distribution of test-fold correlations between true and predicted outcomes across 10 repeated training runs with different random seeds for each benchmark model, alongside `graphTRIP`. (Panel **a** shows mean test predictions averaged across seeds.) The vertical dashed line indicates the correlation threshold for statistical significance. `graphTRIP` achieved consistently significant performance across all seeds, whereas benchmark models did not. Overall performance differed significantly across models (Friedman test: statistic= 38.97, $p = 2.4 \times 10^{-7}$), and `graphTRIP` outperformed all benchmarks in pairwise comparisons (Wilcoxon signed-rank tests; FDR-corrected $p < 0.01$).

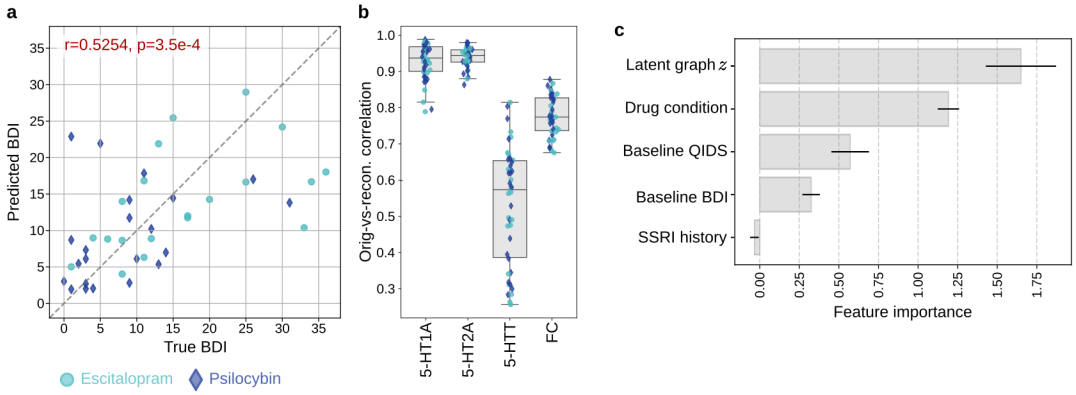

FIG. 8. **Predicting post-treatment BDI scores with `graphTRIP`.** A separate `graphTRIP` model was trained to predict post-treatment BDI scores, using the same configuration and training parameters as the main model predicting post-treatment QIDS. This analysis demonstrates the flexibility of our pipeline 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.

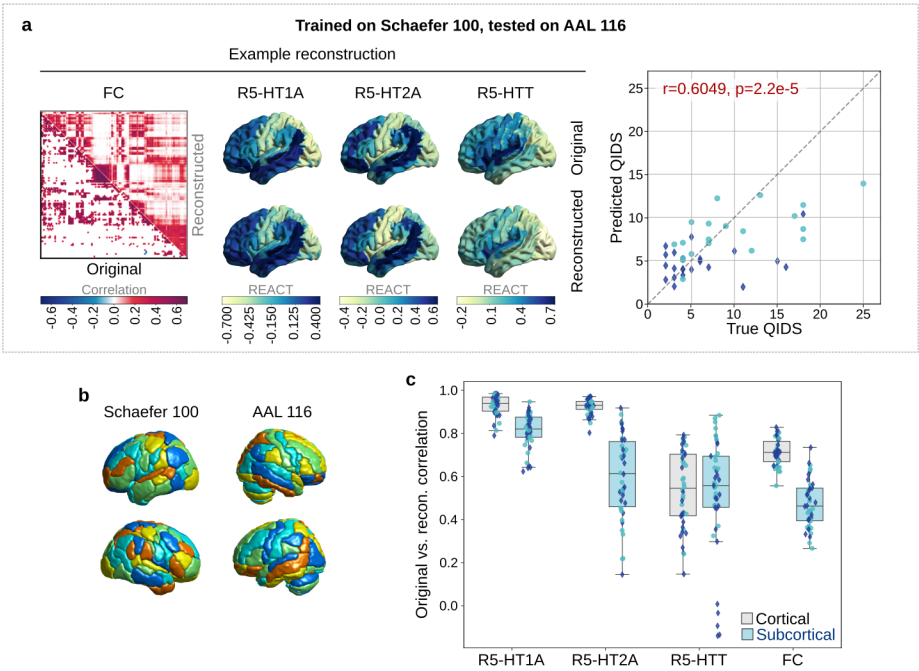

FIG. 9. `graphTRIP` **generalises to brain graphs defined using the AAL atlas.** **a**, `graphTRIP`, trained on brain graphs constructed with the Schaefer 100 atlas, generalises to graphs using the AAL atlas, maintaining strong reconstruction and prediction performance. The panel shows example reconstructions 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.

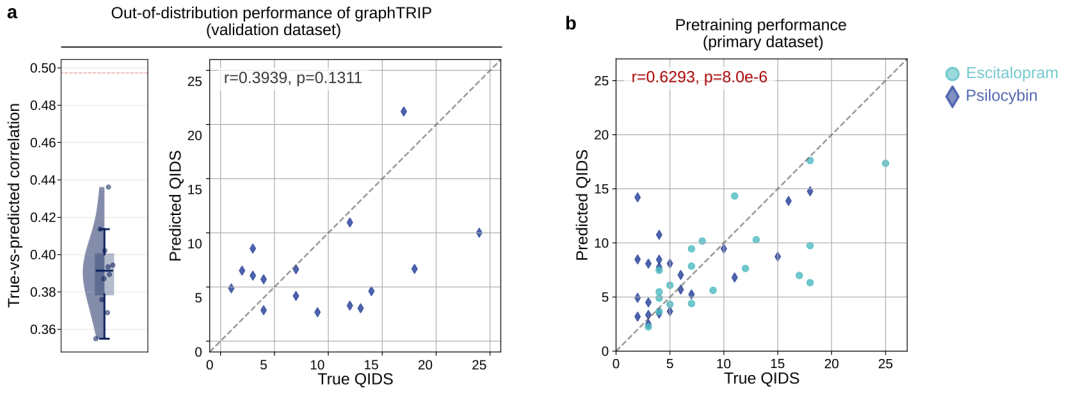

FIG. 10. **Out-of-distribution performance and MLP pretraining indicate that `graphTRIP` learned transferable representations.** **a**, Evaluating the non-fine-tuned core `graphTRIP` models, trained with 10 random seeds, on the validation dataset yields consistently positive true-vs-predicted correlations across seeds (left). The right panel shows true post-treatment QIDS scores plotted against mean test predictions averaged across seeds and CV folds. **b**, Pretraining newly initialised MLP prediction heads on the primary dataset using only inputs compatible with the validation dataset, while keeping the VGAEs frozen, yields significant prediction performance.

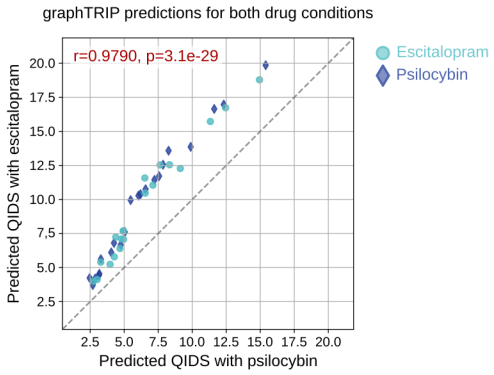

FIG. 11. **`graphTRIP` shows limited sensitivity to treatment condition.** Scatter plot of `graphTRIP` predictions for each patient when the treatment indicator is fixed to escitalopram (y-axis) versus psilocybin (x-axis). Predictions are nearly perfectly correlated and differ primarily by a near-constant offset, indicating that `graphTRIP` captures strong treatment-invariant prognostic signal while only weakly modelling treatment-specific interactions.

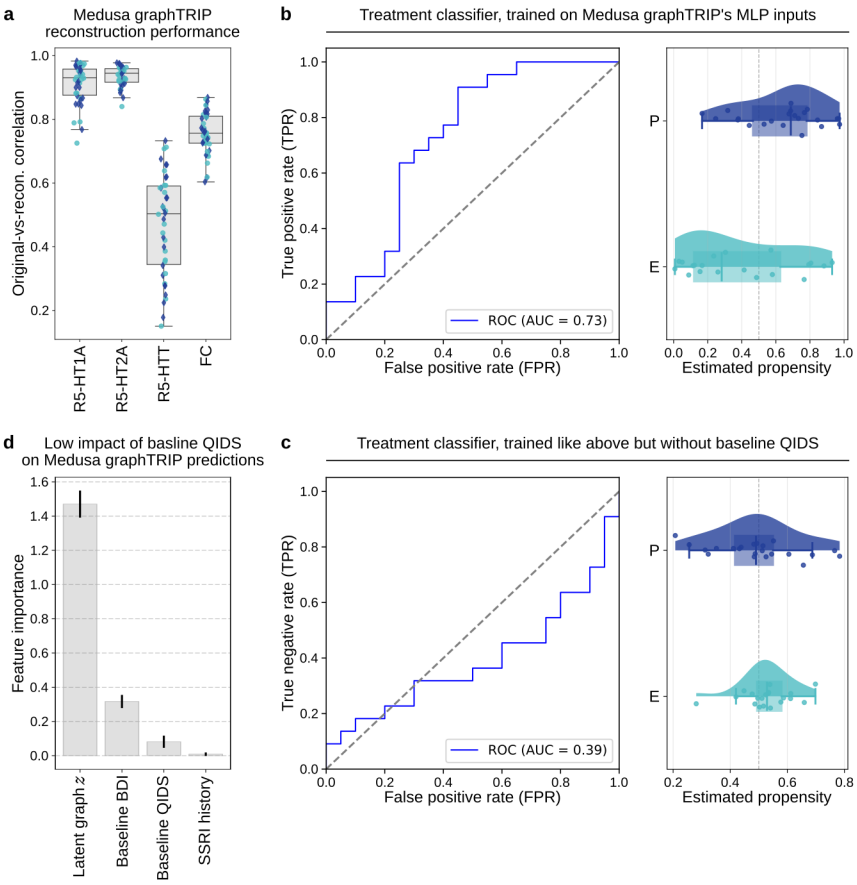

FIG. 12. **Model validation and overlap assessment for `Medusa-graphTRIP`. a**, Correlations between original and reconstructed brain-graph features indicate good reconstruction performance of `Medusa-graphTRIP`'s VGAE. **b**, Treatment assignment was modestly predictable from pre-treatment covariates (AUC= 0.73). A logistic-regression model trained on the inputs to the `Medusa-graphTRIP` outcome heads (frozen VGAE latent representations and clinical covariates) achieved above-chance classification of treatment groups. The ROC curve (left) and the distribution of predicted treatment probabilities (right) indicate that, despite this predictability, estimated propensities remain bounded away from 0 and 1 for most patients, suggesting sufficient overlap between treatment groups. **c**, When baseline QIDS was excluded from the classifier inputs, treatment prediction performance collapsed (AUC= 0.39), with predicted probabilities concentrated near 0.5 for both treatment conditions, indicating that treatment predictability was driven primarily by baseline severity. **d**, Permutation importance analysis for `Medusa-graphTRIP` outcome prediction shows that latent brain-graph representations contribute most strongly to model predictions, whereas baseline QIDS exhibits only weak influence.

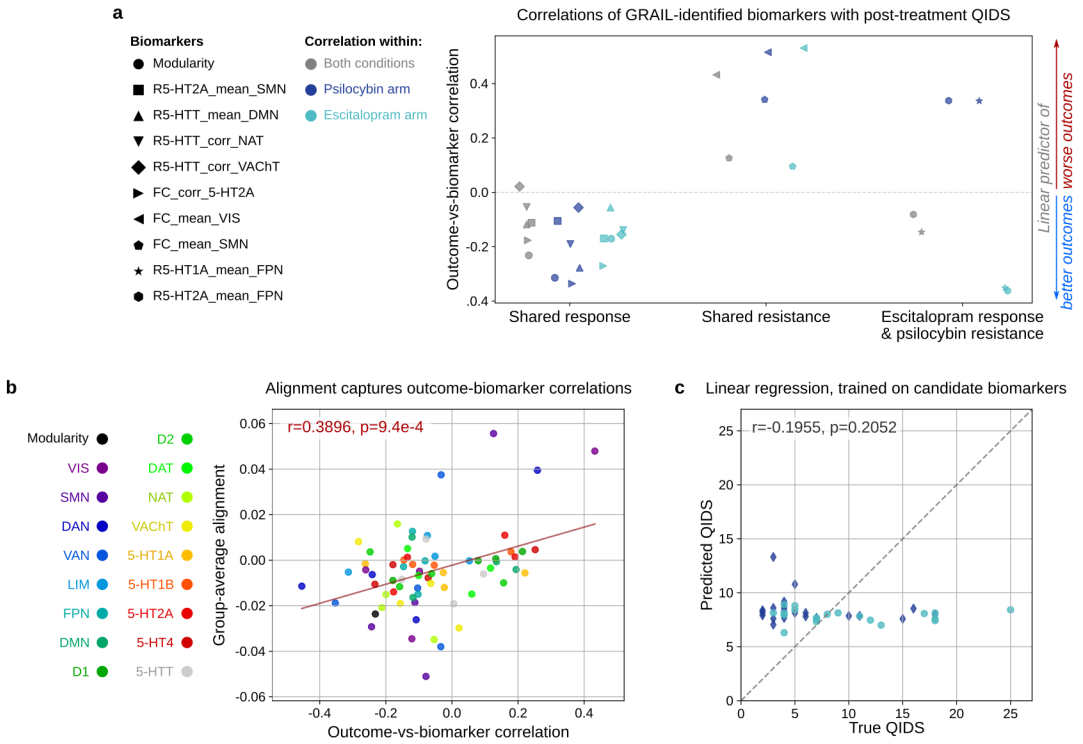

FIG. 13. **Validation of GRAIL biomarker relevance. a**, Univariate correlations between each GRAIL-identified biomarker and post-treatment QIDS. Grey markers show correlations computed across all patients, while blue and green markers show within-condition correlations for the psilocybin and escitalopram groups, respectively. Biomarkers are grouped by the category assigned by our analysis, illustrating that the direction of univariate associations matches the direction of GRAIL alignment: shared responsiveness predictors show negative correlations with outcome, shared resistance predictors show positive correlations, and biomarkers classified as escitalopram responsiveness but psilocybin resistance predictors show opposite correlation signs across treatment arms. **b**, Correspondence between GRAIL-derived alignment and direct biomarker-outcome associations. For each biomarker, we computed its Pearson correlation with outcome and compared these signed correlations across biomarkers to the group-averaged signed alignment values. The resulting significant association ($r = 0.3896$, $p = 9.4 \times 10^{-4}$) indicates that GRAIL captures the direction of biomarker relevance while reflecting model-specific dependencies beyond simple univariate effects. Notably, gradient alignments are patient-specific, whereas feature-outcome correlations are computed at the group level, so perfect correspondence is not expected. **c**, A ridge regression model trained to predict post-treatment QIDS directly from all candidate biomarker values (computed from the input brain graphs) does not achieve significant predictive performance.

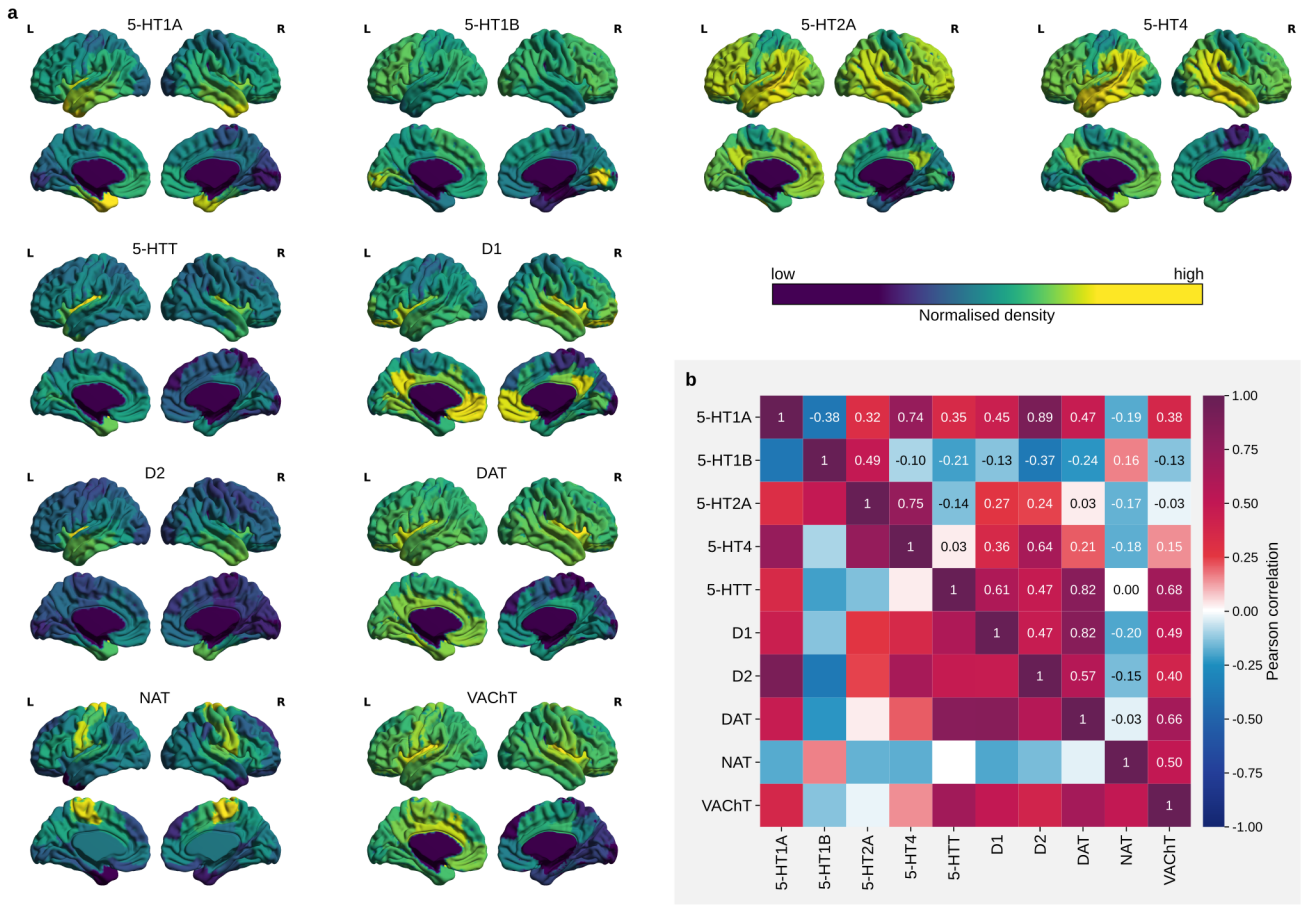

FIG. 14. **Normative molecular target distributions and cross-correlations. a**, Normative molecular target maps used in this study. Each map is individually normalised; the colour scale therefore indicates relative density within a given target map and should not be compared across targets. **b**, Cross-correlation 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.

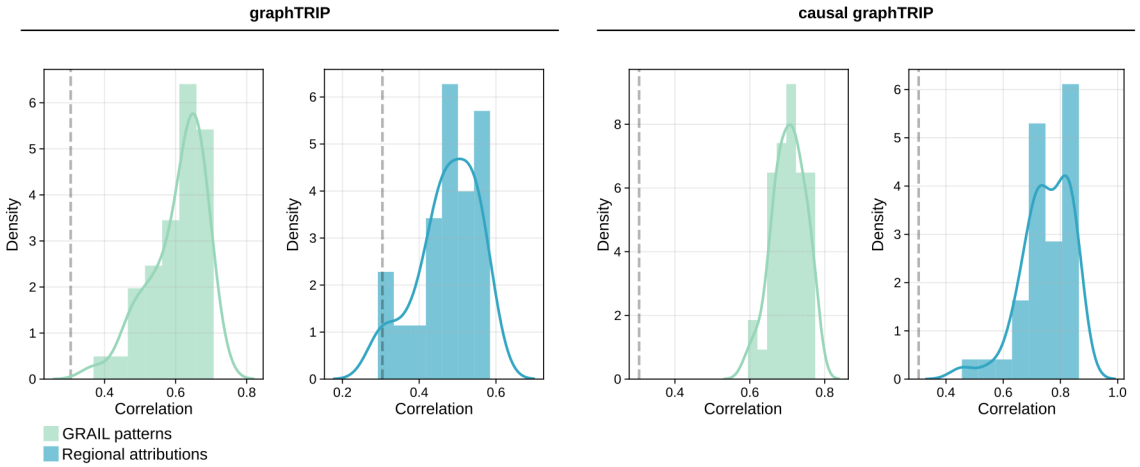

FIG. 15. **Cross-validation model agreement in regional attributions and gradient alignment.** `graphTRIP` and `Medusa-graphTRIP` were trained using 7-fold cross-validation and 10 random seeds, resulting in 70 independently trained models per architecture. This enables estimating regional attribution and gradient alignment profiles for each patient across 70 independent model instances. For each patient, we computed the pairwise correlation between every unique pair of models and summarised agreement as the mean pairwise correlation. Histograms show these mean correlations across all patients for regional attributions and gradient alignment, separately for `graphTRIP` (left) and `Medusa-graphTRIP` (right). Dashed lines indicate the correlation threshold for statistical significance. Overall, both attribution and alignment outputs were highly consistent across models, with mean correlations typically well above the significance threshold.

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

Configuration	**Setting**
Training configuration	
Learning rate	0.001
Number of epochs	300
Number of latent samples	1
Alpha	0.5
Batch size	9
Number of cross-validation folds	7
MLP configuration	
Regularisation strength	0.01
Dropout probability	0.25
MLP architecture	Input: 36; Layers: [64 → 64 → 1] with LeakyReLU (slope = 0.01) and Dropout(0.25) between layers
VGAE configuration	
Latent dimension	32
VGAE regularisation strength	0.01
Encoder	Graphormer-based node encoder with input projection (6 → 16), edge and shortest-path distance encodings used as additive attention biases, followed by 3 Graphormer layers (hidden dim = 16, FFN dim = 32, dropout = 0.25), and an output projection (16 → 32); variational encoding via Linear(32 → 64) to parameterise $(\mu, \log \sigma^2)$
Pooling layer	GraphTransformer pooling that first applies multi-head self-attention and a feedforward network to contextualise node embeddings (32 → 32 → 32), followed by LayerNorm, Dropout(0.25), and mean pooling
Edge decoder	Linear(64 → 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(32 → 32) → LeakyReLU → Dropout(0.25) → Linear(32 → 32) → LeakyReLU → Dropout(0.25) → Linear(32 → 3)

TABLE III. **Model and training configuration for `Medusa-graphTRIP`**. The causal model shares the same VGAE architecture as the core `graphTRIP` model, but replaces the pooling layer and prediction head to enable treatment-specific outcome estimation.

Configuration	Setting
Training configuration	
Learning rate	0.001
Number of epochs	300
Number of latent samples	3
Alpha	0.5
Batch size	9
Number of cross-validation folds	7
Beta (MMD loss parameter)	1.0
CFR prediction head configuration	
Regularisation strength	0.01
Dropout probability	0.20
CFRHead architecture	Two treatment-specific RegressionMLP heads (escitalopram, psilocybin), each with Input: 67; Layers: [32 → 1] using LeakyReLU (slope = 0.01) and Dropout(0.20)
Treatment selection	Predictions selected via treatment indicator t
VGAE configuration	
Latent dimension	32
VGAE regularisation strength	0.01
Encoder	Graphormer-based node encoder with input projection (6 → 16), edge and shortest-path distance encodings used as additive attention biases, followed by 3 Graphormer layers (hidden dim = 16, FFN dim = 32, dropout = 0.25), and an output projection (16 → 32); variational encoding via Linear(32 → 64) to parameterise $(\mu, \log \sigma^2)$
Pooling layer	Mean–Std pooling: concatenation of per-graph mean and standard deviation of latent node embeddings (output dimension = 2×32)
Edge decoder	Linear(64 → 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(32 → 32) → LeakyReLU → Dropout(0.25) → Linear(32 → 32) → LeakyReLU → Dropout(0.25) → Linear(32 → 3)