npj | digital medicine

Article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-025-01912-8

Predicting antidepressant response via local-global graph neural network and neuroimaging biomarkers

Check for updates

Rui Liu[1,2,8], Ximan Hou[1,2,8], Shuyu Liu[3,8], Yuan Zhou[1,2,4,5], Jingjing Zhou[1,2], Kaini Qiao[1,2], Han Qi[1,2], Ruinan Li[1,2], Zhi Yang[1,2], Ling Zhang[1,2], Jian Cui[6,7], Cheng Jin[1,2,3] ✉, Aihong Yu[1,2] ✉ & Gang Wang[1,2] ✉

Depressed mood and anhedonia, the core symptoms of major depressive disorder (MDD), are linked to dysfunction in the brain's reward and emotion regulation circuits. To develop a predictive model for treatment remission in MDD based on pre-treatment neurocircuitry and clinical features. A total of 279 untreated MDD patients were analyzed, treated with selective serotonin reuptake inhibitors for 8–12 weeks, and assigned to training, internal validation, and external validation datasets. A hierarchical local-global imaging and clinical feature fusion graph neural network model was constructed. The model achieved 76.21% accuracy (AUC = 0.78) in predicting remission. Validation on the internal and external independent datasets yielded similar performance (accuracy = 72.73%, AUC = 0.74; accuracy = 71.43%, AUC = 0.72). Key contributing brain regions included the right globus pallidus, bilateral putamen, left hippocampus, bilateral thalamus, and bilateral anterior cingulate gyrus. These findings highlight the role of specific circuits in guiding antidepressant treatment.

Major depressive disorder (MDD) is a significant global mental health issue, characterized by a high incidence and disability rate, imposing a substantial burden on patients and society. Antidepressants, particularly selective serotonin reuptake inhibitors (SSRIs), are widely used as the first-line treatment for depression due to their efficacy in improving depressive symptoms. However, MDD presents with high clinical heterogeneity, leading to individual variability in treatment response. In current clinical practice, antidepressant prescribing often relies on empirical, trial-and-error strategies, resulting in significant variability among clinicians[1,2]. Studies have reported remission rates for first-time antidepressant treatments ranging from 36% to 48%[3,4]. Insufficient remission rates can prolong illness and lead to chronic or treatment-resistant depression[5]. Therefore, identifying objective and effective methods to predict antidepressant response is crucial for achieving personalized and precise therapy.

Previous studies primarily focused on clinical or neuroimaging features independently for remission prediction. Clinical and demographic information, such as age, sex, education, and illness duration, offers insight into patients' subjective experiences and disease backgrounds[6]. In contrast, neuroimaging features serve as objective biomarkers, reflecting changes in brain structure and function. Integrating these two modalities is essential for constructing predictive models of antidepressant efficacy.

Depressed mood and anhedonia, core symptoms of MDD, are often associated with poor treatment outcomes. SSRIs alleviate depressive mood by inhibiting 5-hydroxytryptamine (5-HT) reuptake, with low mood identified as a potential predictor of SSRIs' efficacy[7]. Anhedonia has also been closely linked to treatment outcomes, with studies highlighting its role as a predictor of remission time in SSRIs-resistant MDD adolescents[8]. Systematic reviews have shown that monoaminergic agents, glutamatergic drugs, psychedelics, and stimulants are associated with varying degrees of improvement in anhedonia among adults with MDD[9]. Consequently, brain circuits related to depressed mood and anhedonia hold significant potential as predictors of antidepressant efficacy. This study focuses on neuroimaging circuits associated with these symptoms for feature selection.

Recent advancements in functional magnetic resonance imaging (fMRI) have enabled the study of MDD pathophysiology and the identification of biomarkers for predicting antidepressant efficacy. Studies suggest

[1]Beijing Key Laboratory of Mental Disorders, National Clinical Research Center for Mental Disorders & National Center for Mental Disorders, Beijing Anding Hospital, Capital Medical University, Beijing, China. [2]Advanced Innovation Center for Human Brain Protection, Capital Medical University, Beijing, China. [3]Medical Robot Research Institute, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China. [4]State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China. [5]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China. [6]Department of Psychiatry, Shandong Daizhuang Hospital, Jining, China. [7]Precision Medicine Laboratory, Shandong Daizhuang Hospital, Jining, China. [8]These authors contributed equally: Rui Liu, Ximan Hou, Shuyu Liu. ✉e-mail: chengjin520@sjtu.edu.cn; yuaihong@mail.ccmu.edu.cn; gangwangdoc@ccmu.edu.cn

that dysfunction in the reward and emotion regulation circuits is closely associated with anhedonia and depressed mood, key mechanisms underlying depression[10–12]. The 5-HT-mediated emotion regulation circuit includes the prefrontal cortex, hippocampus, amygdala, orbitofrontal cortex (OFC), and anterior cingulate cortex (ACC)[12]. Changes in the structure and function of these regions in MDD patients are linked to symptom trajectories and serve as significant predictors of antidepressant efficacy[13,14]. Similarly, the reward circuit, which includes the ventral striatum, ventral pallidum, dorsolateral prefrontal cortex (DLPFC), OFC, ACC, and thalamus, has been implicated in anhedonia[15,16]. A review indicated that the reward circuit comprises several key brain areas such as the ventral striatum, ventral pallidum, DLPFC, OFC, ACC, and thalamus[17]. Research indicates that 5-HT levels in the insula and ventral striatum are associated with anhedonia and that SSRIs may alleviate this symptom by modulating these levels[18]. Structural and functional changes in these circuits following SSRIs treatment are closely associated with symptom trajectories and antidepressant efficacy[19–23].

Baseline clinical characteristics and demographic information have been associated with antidepressant outcomes[24]. Hamilton Depression Rating Scale (HAMD) score have been proposed as predictors of early remission[25]. Moreover, psychosocial functioning, a stable predictor of long-term prognosis, has been suggested as an indicator of long-term treatment efficacy[26–28]. This study aims to integrate age, sex, education level, illness duration, HAMD score, Quality of Life Enjoyment and Satisfaction (QLES) Questionnaire score, and neuroimaging features to develop a predictive model for antidepressant treatment outcomes.

Machine learning (ML) based on fMRI has been increasingly applied to predict treatment responses in MDD. Previous studies primarily utilized traditional ML models such as support vector machines, random forests, or logistic regression for predicting antidepressant efficacy[29–31]. However, these models often fail to fully explore the complex topological structures of data and the intricate dependencies between features[32,33], potentially limiting insights into the mechanisms underlying brain changes induced by antidepressant treatment. Advanced ML methods are therefore essential for improving the prediction of antidepressant efficacy.

Traditional network analysis methods, including metric indices, path analysis, and network models, have been used to evaluate macroscopic characteristics and structural properties of networks. Nonetheless, these methods struggle to capture complex network topologies and higher-order dependencies. Graph neural networks (GNNs), by contrast, update node representations by aggregating information from neighboring nodes without relying on a fixed number or order of neighbors, effectively capturing complex topological structures within graphs. Unlike traditional network analysis, which is limited to direct interactions, GNNs accommodate versatile topological structures, with nodes and edges corresponding to the brain's regions of interest (ROIs) and their structural and functional connections. This approach has been shown to enhance prediction performance.

Recent advances in the application of GNNs to neuro-images such as fMRI have laid the groundwork for learning informative brain representations and supporting downstream tasks such as brain age prediction, sex classification, and disease diagnosis. Notably, BrainGNN[34] introduced ROI-aware graph convolutional layers and ROI-selection pooling, enabling adaptive learning of region-specific features and interpretable biomarker identification. Extending this progress, LGGNet[35] successfully adopted a local-global paradigm for brain-computer interfaces using electroencephalogram (EEG) data, learning both intra- and inter-functional brain activities. Similarly, PLI-GCNN[36] leveraged hybrid feature representations by combining electrode-level characteristics and global topological patterns to automate the detection of alcoholism. Building on this foundation, BNT[37] proposed an orthonormal clustering readout for self-supervised soft clustering, further enhancing the discriminability of node embeddings across functional brain modules. BrainRGIN[38] integrated clustering-based embeddings and graph isomorphism networks to better capture modular brain sub-network organization and enhance graph-level representations via attention-based readout functions. CI-GNN[39] proposed a Granger causality-inspired GNN to identify the most influential subgraph that is causally related to the decision.

Despite these advances, many of these models primarily focused on static, subject-specific graphs, consequently limiting their generalizability across clinically heterogeneous populations. Moreover, they often overlooked the integration of population-level information, which is crucial for understanding the diversity of brain function and pathology across individuals. In response, recent studies have explored hierarchical GNN architectures to simultaneously model intra-subject brain features and inter-subject similarities. For instance, SFC-GNN[40] combined regional graph perception with a structure feature pooling strategy, constructing population-level graphs via similarity kernels and enabling node classification through community-aware embeddings. However, these models did not support end-to-end joint optimization of local and global graphs, which can hinder the alignment between individual- and population-level learning objectives. LG-GNN[41] further introduced a two-level framework in which a self-attention-based local ROI-GNN captures regional biomarkers, while a global subject-level GNN integrates both imaging and non-imaging data to enhance classification performance. This model achieved end-to-end optimization of the local-global network structure, highlighting the importance of integrating non-imaging data and subject relationships. However, it still lacked the capability to dynamically update graph structures based on task-specific signals derived from time-resolved neural features.

Alongside these developments, there has been growing attention to the temporal dimension of fMRI data, particularly the dynamic fluctuations in BOLD signals. Models such as BrainGNN and BrainRGIN treated fMRI as static snapshots, ignoring the temporal progression of neural activity. Addressing this limitation, Graph Clustered Transformer[42] conducted a comparative analysis of resting-state functional magnetic resonance imaging (rs-fMRI) time points and resting-state functional connectivity (rs-FC) for Alzheimer's prediction, utilizing temporal convolutional networks (TCNs) for temporal feature embedding. Deep-Spatiotemporal[43] integrated GNNs and temporal networks, including TCNs and LSTM, to model spatiotemporal dynamics in rs-fMRI data, and confirmed findings that RNNs and CNNs can provide similar performance. DSAM[44] further proposed a dynamic spatiotemporal attention framework that employed TCNs to extract multi-scale temporal features and leveraged self-attention to learn task-specific FC directly from the time series. These works validate the feasibility of integrating temporal features into GNN architectures and further support the construction of task-specific brain connectivity matrices.

In addition to these modeling challenges, the integration of multimodal data remains underdeveloped in the context of neuroimaging-based prediction. While Deep-Spatiotemporal[43] proposed an architecture combining TCNs and GNNs for spatiotemporal learning and claimed support for multimodal integration, its fusion strategy was confined to combining fMRI and structural connectivity from diffusion-weighted imaging (DWI). MS^2-GNN[45] proposed a GNN-based multimodal fusion strategy, which successfully investigated the heterogeneity/homogeneity among audio and EEG modalities for the subsequent MDD detection task. Following this trend, LGMF-GNN[46] proposed the local-global multimodal fusion GNN, which jointly modeled local ROI connectivity and global population-level patterns, integrating functional and structural MRI with clinical data to enhance MDD diagnostic accuracy. These advances collectively underscore the critical importance of multimodal fusion in enhancing the predictive power and clinical applicability of GNNs.

In this context, as illustrated in Fig.1, we propose a hierarchical local-global imaging and clinical feature fusion GNN (LGCIF-GNN) to predict antidepressant efficacy in the acute phase of SSRIs treatment. First, the model performs dynamic graph structure optimization by adaptively updating adjacency matrices based on pairwise similarities of ROI-level temporal embeddings extracted via a bidirectional GRU (bi-GRU) encoder. This learnable, task-driven graph construction captures richer temporal dependencies than static correlation methods and aligns graph topology with treatment prediction.

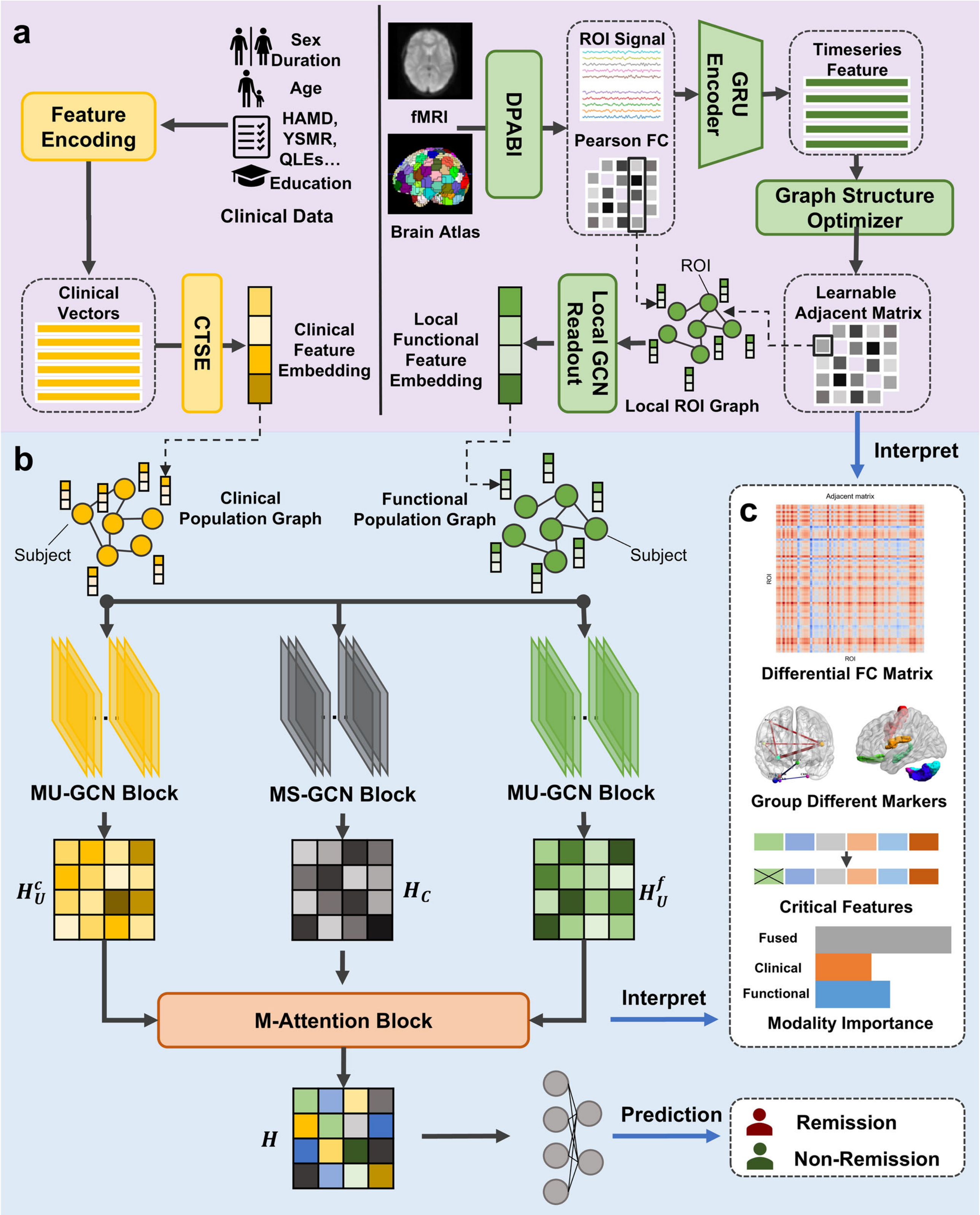

Fig. 1 | Schematic overview of the LGCIF-GNN framework. a Local graph construction and encoding. At the local level, subject-specific functional graphs are constructed from ROI-wise BOLD time series and processed via a GRU encoder and GCN-based readout to extract individual brain embeddings. In parallel, clinical variables are embedded into feature vectors. **b** Global graph modeling and multimodal fusion. The local embeddings are used to construct population-level functional and clinical graphs, where nodes represent subjects and edges encode modality-specific similarity between subjects. Global graph modules extract shared and unique representations across modalities, which are fused via attention and passed to an MLP for individualized prediction. **c** Model interpretability and marker mining. The model supports interpretation by identifying discriminative functional connections, relevant clinical features, and modality contributions to treatment outcome prediction.

Table 1 | Demographic and clinical information of train data

	Remission (74)	Non-remission (90)	*p* value (two-tailed)
Age (median, range)	28 (18–56)	27.0 (18–59)	0.485
Sex (male/ female)	24/50	23/67	0.332
Education level[a]	22/41/11	14/59/17	0.091
HAMD-17	19.53 ± 3.75	21.88 ± 4.03	<0.001
YSMR	0.50 ± 1.15	0.70 ± 1.86	0.488
QLES	36.26 ± 7.43	32.87 ± 8.99	0.005
Duration (day)	995.53 ± 1196.53	1643.99 ± 1934.82	0.012
Episode (first-episode/ recurrence)	40/34	44/46	0.510

[a] The education level of adult groups was divided into three categories: primary school/junior, senior high school/bachelor's degree, and above. *HAMD-17* 17-item Hamilton Depression Rating Scale, *YMRS* Young Mania Rating Scale, *QLES* Quality of Life Enjoyment and Satisfaction Questionnaire.

Second, we introduce a local-global architecture that jointly models intra-subject ROI-level dynamics and inter-subject population-level similarities. The local network focuses on modeling fine-grained, ROI-level functional dynamics within each subject's brain activity, encoding rich temporal dependencies that reflect neural processes underlying MDD. In parallel, the global network operates over population graphs based on functional and clinical similarity among subjects, capturing inter-individual relationships and common patterns. By integrating these two levels, the end-to-end network effectively fuses personalized features with population-wide trends. This design provides complementary, multi-angle supervision signals that enhance the model's ability to identify both individual-specific and group-consistent features predictive of SSRIs treatment response.

Third, to bridge the longstanding gap in multimodal fusion, we integrated neuroimaging and clinical data within a unified graph-based architecture. Specifically, we constructed phenotype-informed population graphs using a clinical trait similarity encoder (CTSE) and employed specialized modules—modality-unique graph convolutional networks (MU-GCNs), modality-shared GCNs (MS-GCNs), and a modality-attention fusion block—to extract and integrate complementary information from each modality. These design choices facilitate not only improved performance in predicting SSRIs treatment response in MDD but also provide clinically interpretable insights, representing a step toward precision psychiatry.

Focused on reward and emotion regulation circuits—including the nucleus accumbens, striatum, thalamus, amygdala, hippocampus, DLPFC, OFC, and ACC—this study employed ablation analysis to evaluate the impact of radiographic features on model performance. An independent external validation set was used to confirm the model's predictive accuracy and generalizability. It was hypothesized that clinical and neuroimaging features from the reward and emotion regulation circuits would effectively predict remission in MDD patients after 8–12 weeks of antidepressant treatment.

Results

Demographic and clinical characteristics

In the training set, 19 patients were removed due to incomplete scale data or duration information. For the internal validation dataset, 2 patients were excluded due to head movement. For the external validation dataset, 1 patient was excluded due to head movement, and an additional 2 patients were removed due to incomplete scale data. Ultimately, 279 patients were included in the final analysis.

The demographic and clinical characteristics of the training set are summarized in Table 1. The remission rate for MDD patients after 8 or 12 weeks of SSRIs treatment was 45.12% (74 out of 164 patients achieved remission). No significant differences were found between groups regarding sex, age, education level, or episode frequency (all $p \geq 0.05$). Compared to the remission group, the non-remission group exhibited higher HAMD-17 scores and longer illness duration ($p < 0.001$ and $p = 0.012$, respectively), and lower QLES scores ($p = 0.005$). Details of demographic and medication information for the 66 MDD patients in the internal validation dataset and the 49 MDD patients in the external validation dataset are provided in the Supplementary Materials (Supplementary Tables 1 and 2).

Prediction model and performance validation

The area under the curve (AUC) of the prediction model was 0.78, with a sensitivity of 75.20%, a specificity of 77.48%, and an accuracy of 76.21% (Fig. 2a). Validation on the internal independent validation set yielded comparable results, with an accuracy of 72.73%, sensitivity of 73.53%, specificity of 71.88%, and AUC of 0.74 (Fig. 2b).

To further evaluate the generalizability of our approach, we tested the model on an external cohort from an independent clinical site. The LGCIF-GNN achieved an AUC of 0.72, with an accuracy of 71.43%, sensitivity of 70.00%, and specificity of 72.41% (Fig. 2c). Despite differences in imaging protocols and demographic composition, the model maintained robust predictive performance, indicating its ability to generalize across sites and capture clinically relevant patterns.

To better understand the contribution of each modality, an ablation study of imaging features was conducted. Model performance declined significantly when using only clinical features without incorporating neuroimaging data into the fusion model, resulting in an accuracy of 69.41%, sensitivity of 66.01%, specificity of 73.45%, and AUC of 0.71 (Fig. 2d). This comparison clearly demonstrates the added value provided by integrating imaging modality.

Model interpretation

The attention score for functional imaging data was 0.3312, indicating a stronger influence compared to clinical data, which had a score of 0.3284. The multimodal (MC) embedding, which integrates both modalities, achieved the highest attention score of 0.3404, underscoring its critical contribution to the model's predictive performance.

The differential rs-FC matrix identified the top five enhanced and top five diminished rs-FC characteristics within the reward and emotion regulation circuits (Fig. 3). The primary brain regions involved included the right globus pallidus, bilateral putamen, left hippocampus, bilateral thalamus, and bilateral ACC. Among these regions, the left hippocampus was the most frequently selected.

Feature masking analysis revealed that imaging features contributed more significantly to the model's predictive performance than clinical features (Fig. 4). The top ten node features included brain regions such as the bilateral ACC, right globus pallidus, bilateral dorsolateral prefrontal cortex (dPFC), OFC, right hippocampus, and right thalamus. Among clinical features, HAMD item 5 (Sleep poorly) and item 8 (Retardation) showed slightly higher contributions compared to other clinical features (Fig. 4).

The neural substrates showing significant alterations in rs-FC between remission and non-remission groups included the left ACC, right thalamus, and right globus pallidus. These brain regions were identified as pivotal contributors, ranking among the top ten nodes with the most substantial impact on network prediction performance, as determined by the feature masking strategy.

Discussion

The primary objective of this study was to develop a GNN model to predict the efficacy of SSRIs during the acute treatment period based on rs-fMRI features of the reward and emotion regulation circuits, along with clinical characteristics. The model demonstrated satisfactory predictive performance with an accuracy of 76.21%, and its robustness and generalizability were confirmed using an independent validation set (accuracy = 72.73%). Ablation studies revealed that neuroimaging features made a significant

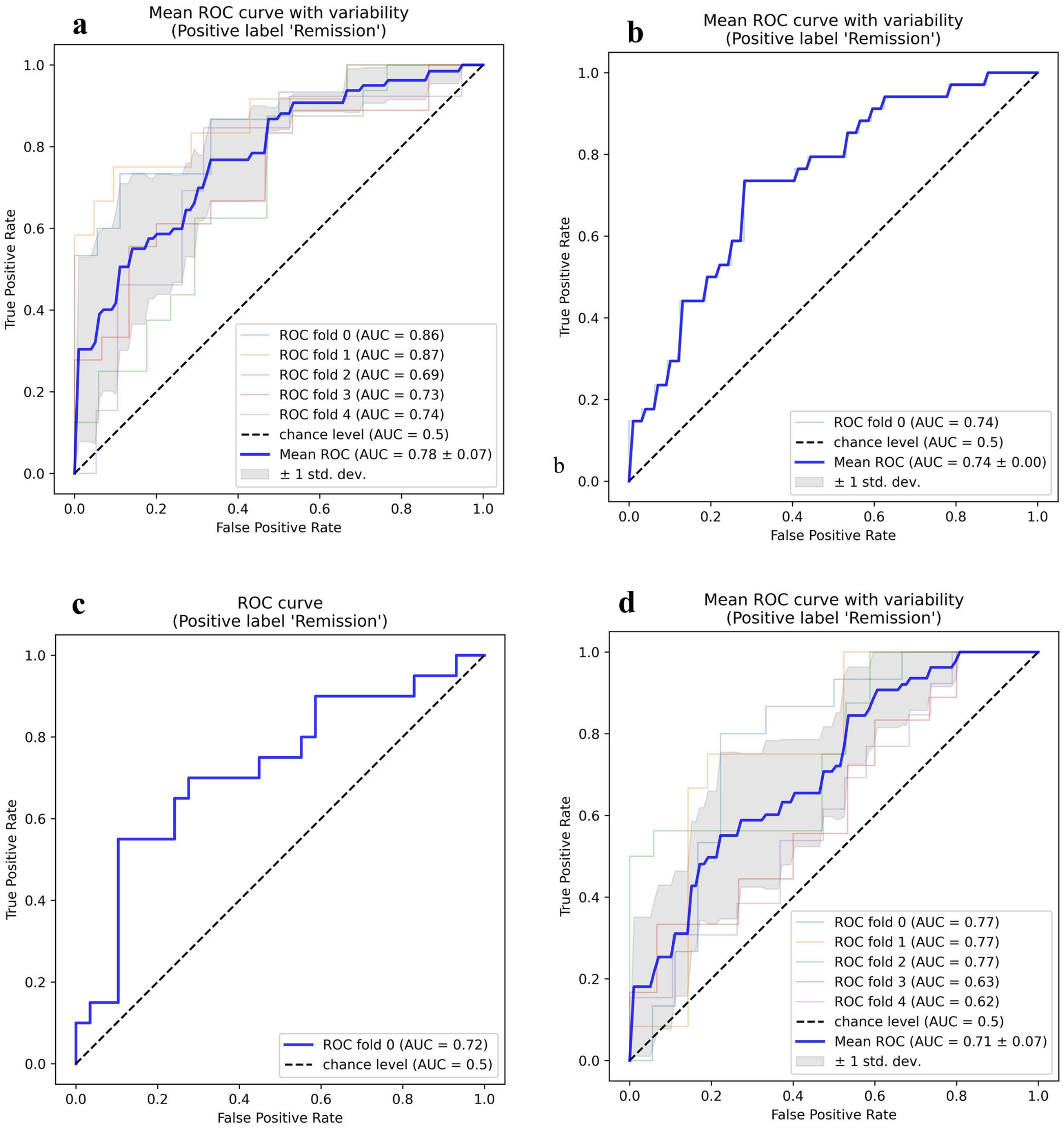

Fig. 2 | Performance evaluation and generalizability of the LGCIF-GNN. a The ROC of the 5-fold cross-validation. The AUC is 0.78, accuracy is 76.21%, sensitivity is 75.20%, and specificity is 77.48%. **b** The ROC on the internal independent validation set. The AUC is 0.74, accuracy is 72.73%, sensitivity is 73.53%, and specificity is 71.88%. **c** The ROC on the external independent validation set from a separate center. The AUC is 0.72, accuracy is 71.43%, sensitivity is 70.00%, and specificity is 72.41%. **d** The ROC of the ablation study using clinical features alone. The AUC is 0.71, accuracy is 69.41%, sensitivity is 66.01%, and specificity is 73.45%.

contribution, with the most impactful brain regions including the right globus pallidus, bilateral putamen, left hippocampus, bilateral thalamus, and bilateral ACC. These findings underscore the potential of a model combining clinical and neuroimaging features of the reward and emotion regulation circuits for predicting SSRIs response in MDD patients.

Recent advances in spatiotemporal GNNs have substantially shaped the landscape of neuroimaging-based predictive modeling. Models such as BrainGNN[34], BrainRGIN[38], and BNT[37] have demonstrated the utility of ROI-aware architectures or cluster representation for brain network modeling, integrating specific brain regions and subnetworks. Concurrently, spatiotemporal frameworks such as DSAM[44], Graph Clustered Transformer[42], and Deep-Spatiotemporal[43] have pioneered hybrid temporal-graph approaches for capturing dynamic brain states. LG-GNN[41], SFC-GNN[40], and MS^2-GNN[45] have made strides toward integrating intra-subject features and population-level information or fusing features from different modalities. While building upon these foundations, our Local-Global Imaging and Clinical Feature Fusion GNN (LGCIF-GNN) offers a meaningful advancement by combining dynamic, task-adapted graph

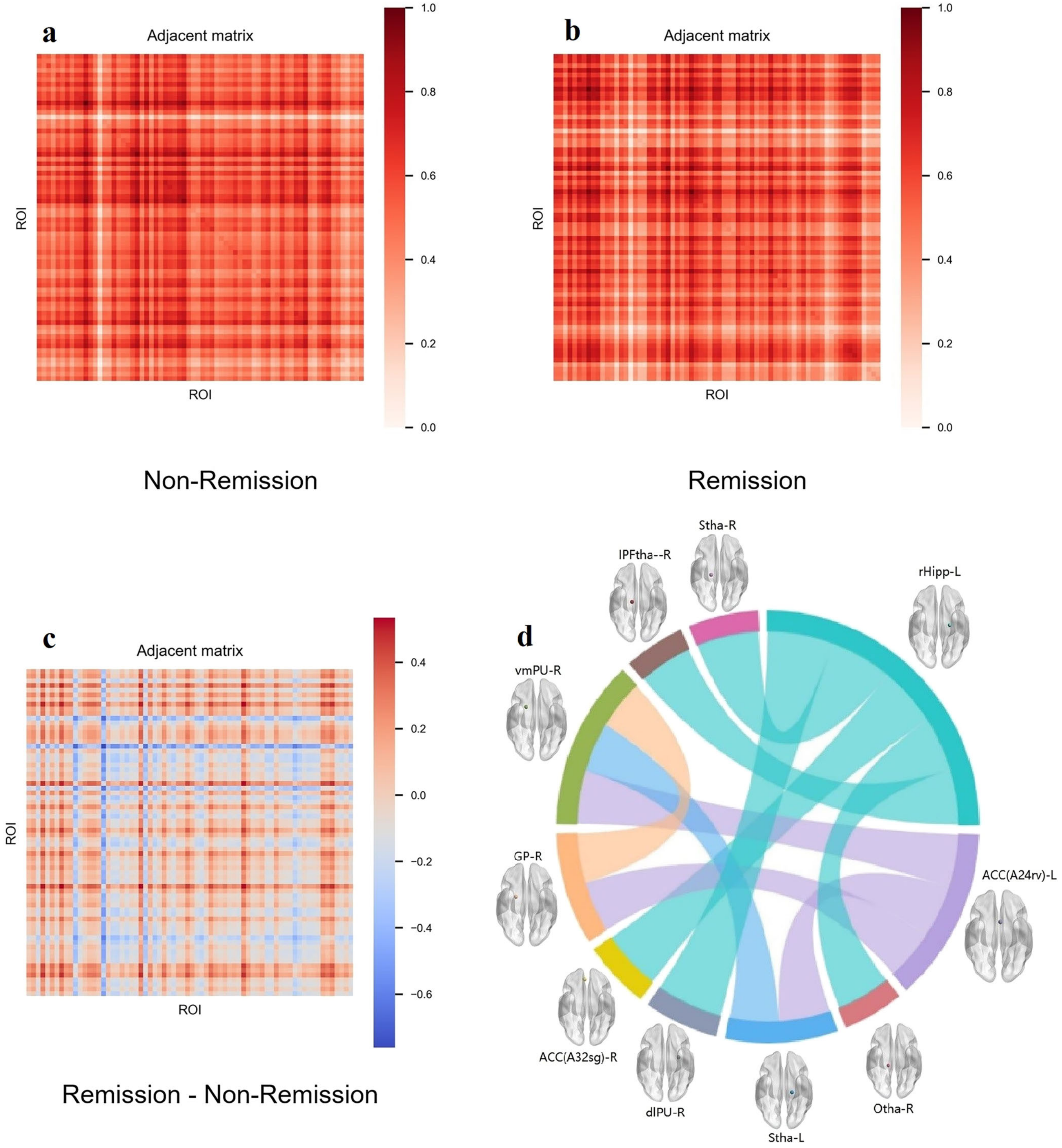

Fig. 3 | Differential FC matrix and altered FC in remission and non-remission Groups. a–c The differential FC matrix between the remission and non-remission groups. The color bar indicates the strength of functional connectivity. **d** The top five enhanced and the top five diminished FC were shown in a chord picture.

construction, local-global modeling that bridges subject-specific neurobiological precision with population-level similarity, and interpretable multimodal disentanglement and fusion.

First, unlike prior edge-enhanced GNNs such as BrainGNN[34] and BrainRGIN[38], which rely on static FC matrices or precomputed similarity graphs, LGCIF-GNN constructs subject-specific functional graphs during training by leveraging the dynamic similarity of ROI-level temporal embeddings. This approach allows the model to directly align graph topology with the treatment prediction task, improving representational specificity and model interpretability. Additionally, this design avoids hard clustering or top-k pooling strategies in previous works, such as Graph Clustered Transformer[42] and BNT[37], which are sensitive to hyperparameters and risk discarding relevant features. Specifically, our attention-based mechanism enables soft graph readout, preserving critical regional information and improving robustness.

Second, previous studies have mainly defined the “spatiotemporal” or “local-global” concept within the scope of individual subjects, focusing on fine-grained temporal dynamics within single ROIs and spatial or global relationships across brain regions. For example, models like DSAM[44] and Deep-Spatiotemporal[43] integrate TCNs with GNNs to capture intra-subject

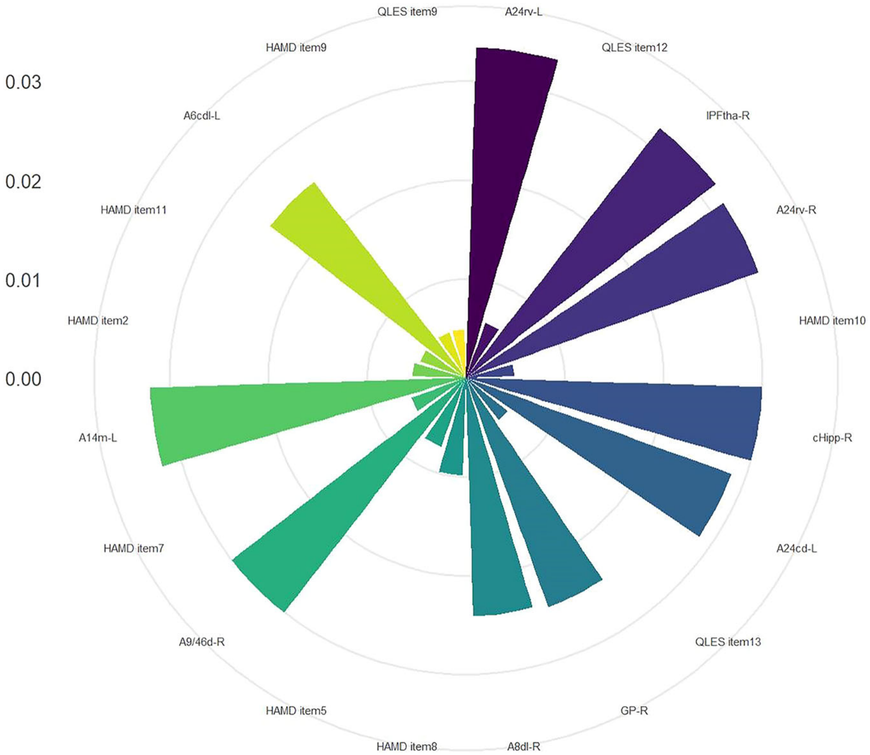

Fig. 4 | Weight distribution of the top ten contributing clinical and neuroimaging features. The radius or the area covered by each sector in the wind direction rose map represented the magnitude of the weight values for each feature in the classification model. The results revealed that neuroimaging features have a greater contribution to the model than clinical features.

spatiotemporal patterns using unidirectional and multi-scale temporal modeling. In contrast, LGCIF-GNN replaces the commonly used TCN modules with bi-GRU encoders, enabling the extraction of long-range, bidirectional temporal dependencies—crucial for capturing feedback and recurrent processes in emotion and reward-related circuits. Further, our framework advances this paradigm through a population-tiered hierarchy. This architecture dynamically links individual-level ROI dynamics (“local”) to population graphs constructed from functional and clinical similarity (“global”), enabling simultaneous learning from personalized neurodynamics and group-level pathophysiological patterns. This dual-scale supervision captures both individual heterogeneity and population-consistent biomarkers of treatment response in an end-to-end manner—a capability underexplored in prior local-global GNNs.

Third, while recent multimodal GNNs have started incorporating non-imaging data, they often do so in a limited manner. For instance, Deep-Spatiotemporal[43] replaces the adjacency matrix with DWI-derived structural connectivity, without true feature-level integration; LG-GNN[41] includes only age, sex, and site as phenotypic inputs. MS^2-GNN[45] recently demonstrated the promise of integrating EEG and audio data for MDD detection, but the application of GNNs to fuse fMRI and clinical features in psychiatry—especially for outcome prediction—remains underexplored. In contrast, LGCIF-GNN presents a more comprehensive and clinically grounded multimodal framework. It incorporates a diverse range of phenotype- and symptom-relevant clinical features, including age, sex, education, disease duration, and multiple clinical scale scores (HAMD, QLES, YMSR), which reflect both underlying biological vulnerability and external clinical status. Through modality-specific and shared feature extraction (MU-GCN, MS-GCN) and adaptive fusion, it achieved task-specific, interpretable multimodal integration tailored to SSRIs response prediction. This integration notably enhanced predictive performance, with the area under the receiver operating characteristic (ROC) curve (AUC) increasing from 0.71 to 0.78, a 10% improvement, after incorporating imaging features. These results highlight the added value of including neuroimaging data to enable more precise and individualized predictions of treatment response.

Moreover, previous studies on antidepressant efficacy prediction often utilized small sample sizes, unbalanced datasets, or uncontrolled polypharmacy approaches[7,30,31], leading to potential confounding in model assessments. In contrast, this study employed a larger sample size, balanced training and testing datasets, and a rigorous single-drug therapy protocol, enhancing the stability and generalizability of the predictive model. Additionally, the use of an independent validation set strengthens the model’s robustness. Unlike earlier studies relying solely on internal cross-validation, this study employed a validation dataset acquired with different scanning parameters and acquisition batches, facilitating a more comprehensive evaluation of generalization performance.

The proposed model achieved an AUC of 0.74 and an accuracy of 72.73% on the internal independent validation set, and an AUC of 0.72 and an accuracy of 71.43% on the external independent validation set, demonstrating its robustness and reliability. External validation is critical for assessing a prediction model’s adaptability to real-world scenarios beyond its development data and population[47]. For comparison, a recent study by Poirot et al. developed an XGBoost model to predict sertraline response, achieving an internal cross-validation accuracy of 68% and an external validation accuracy of 65%[48], which is notably lower than the accuracy achieved by our model.

Consistent with the latest meta-study[49], neuroimaging features contributed more significantly to the model than clinical features. The ablation experiments and feature importance analysis demonstrated that resting-state brain imaging features played a more substantial role in predicting treatment outcomes compared to clinical features. Combining the results of the differential FC matrix analysis and feature masking analysis of node degrees revealed that the ACC, globus pallidus, hippocampus, and thalamus were common key features in predicting SSRIs treatment outcomes.

The right globus pallidus, a key component of the reward circuit[50,51], showed FCs with the left rostroventral area 24 and the right ventromedial putamen, forming an integrative network responsible for modulating emotional and motivational states. The globus pallidus plays a crucial role in regulating anxiety- and depression-like behaviors and can integrate and transmit signals related to motivation, reward, stress, and depression in the brain[52]. A recent study has shown that the plasticity of the cholinergic neuronal circuit in the ventral globus pallidus regulates pain-like and depression-like behaviors in mice[53]. Another study has shown that there are structural and functional abnormalities in the putamen among patients with depression and those at genetic risk in their families, suggesting that the putamen is a potential biomarker for depression[54,55]. The left hippocampus, a core component of the emotion regulation circuit[11], and the right thalamus, which is involved in both the reward and emotion regulation circuits[7,12,33], were also identified as critical nodes. FCs between the left rostral hippocampus and subregions of the right thalamus, the right subgenual area 32, and the right dorsolateral putamen form an integrative network mediating interactions among multiple brain systems involved in emotional and reward processing[56,57]. The hippocampus plays a crucial role in emotional regulation and is closely related to the pathophysiological mechanism of depression[58]. Research has shown that SSRIs promote neurogenesis in the dentate gyrus of the hippocampus and selectively act on the serotonergic pathway[59]. Recently, many studies have emphasized that the pathways involving the thalamus may be the targets for the treatment of depression. Zhang et al. found a circuit from the visual cortex to the lateral posterior thalamic nucleus regulates depression-like behaviors in male mice[60]. Zhang et al. found that the pathway from the thalamic reticular nucleus to the lateral habenula regulates depression-like behaviors in chronic stress and chronic pain[61]. In this study, our findings highlight the central role of functional alterations in the hippocampus, thalamus, and globus pallidus in predicting antidepressant efficacy. Dysfunctions in these connections may underlie core depressive and anhedonia symptoms. Antidepressant interventions likely exert their therapeutic effects by restoring functional synchrony, offering promising targets for MDD treatment strategies.

Regarding clinical features, HAMD item 5 (“Sleep poorly”) and item 8 (“Retardation”) contributed slightly more than other clinical features. Previous studies have suggested that the initial presentation of retardation and sleep disturbances may influence antidepressant efficacy[62–64], aligning with our findings. While neuroimaging features performed well independently, adding clinical features further enhanced model accuracy.

However, this study has several limitations. Firstly, the FC structure was initially estimated using Pearson correlation, a widely adopted approach in neuroimaging research due to its practicality and interpretability. While our model incorporates a graph structure optimization mechanism to adapt and refine the FC using temporally informed representations, future investigations into alternative FC metrics may offer additional benefits in expressiveness and biological fidelity. Second, for poorly covered masks, NaN values were replaced with the mean of non-NaN values from other MDD patient signals. Future research could explore personalized mask configurations. Lastly, this study focused solely on establishing a predictive model for SSRIs efficacy. Considering the variety of real-world medications, future studies should include additional antidepressant types to build more comprehensive efficacy prediction models.

In summary, the Local-Global Imaging and Clinical Feature Fusion Graph Neural Network (LGICF-GNN) was successfully applied to predict the acute-phase efficacy of SSRIs treatment in MDD patients, and it demonstrated consistently strong performance on the independent internal and external validation sets. The integration of clinical and functional imaging features achieved optimal predictive performance, with imaging features contributing significantly more than clinical features. These findings highlight the potential of neuroimaging features from the reward and emotion regulation circuits as predictors of antidepressant response. The current results represent an important step toward biomarkers of antidepressant response.

Methods

Participants

The MDD cohort in this study was derived from three cohort studies (ChiCTR-OOC-17012566, MR-11-23-003930, and ChiCTR2200059053) conducted at Beijing Anding Hospital, Capital Medical University, from September 2018 to October 2023. The study included 183 MDD patients in Cohorts 1 and 2, and 68 MDD patients in Cohort 3. Cohorts 1 and 2 were used as the training data, with all patients receiving SSRIs treatment for either 8 or 12 weeks. Additionally, Cohort 3 was designated as the internal validation set to assess model stability and generalization, with all patients receiving SSRIs treatment for 8 weeks. To further evaluate the model’s generalizability across different clinical sites, we included an external validation dataset comprising 52 MDD patients recruited from Shandong Daizhuang Hospital between March 2023 and February 2025, all of whom received SSRIs treatment for 8 weeks. Treatment outcomes were assessed using the same clinical criteria as in the primary cohort to ensure consistency in endpoint definition. The inclusion and exclusion criteria were consistent with those applied in the discovery cohort. This external dataset was acquired under different scanning protocols and demographic conditions, providing a realistic testbed for evaluating the model’s robustness across variations in scanner hardware, acquisition parameters, and population characteristics.

Inclusion criteria were: (1) adults aged 18–65 years; (2) Han ethnicity and right-handedness; (3) diagnosis of MDD based on the Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) for Cohorts 1 and 2, or the DSM-V for Cohort 3; (4) no prior antidepressant use or use for no more than seven days within the preceding 14 days; (5) willingness to undergo SSRIs treatment. Exclusion criteria included: (1) significant non-depression DSM-IV or DSM-V diagnosis ; (2) previous intolerance or lack of response to SSRIs; (3) MRI contraindications; (4) presence of psychotic symptoms.

The study was approved by the Human Research and Ethics Committee of Beijing Anding Hospital, Capital Medical University, and all participants provided informed consent (Approval No. 2017-24, 2020-106, 2022-14-202221FS-2). This study follows the STROBE statement checklist.

Treatment and clinical assessment

The HAMD-17 was used to assess depressive symptoms. Patients achieving a HAMD score ≤7 after 8 or 12 weeks of SSRIs treatment were classified as the remission group, while those with a score >7 were classified as the non-remission group. The Young Mania Rating Scale (YMRS) was used to assess manic symptoms, consisting of 11 items with total scores ranging from 0 to 44, where higher scores indicate more severe symptoms. The quality of life was measured using the 16-item QLES Questionnaire[65]. The total score was calculated by summing the first 14 items, each rated on a 5-point Likert scale (1 = very poor to 5 = very good), resulting in a total score range of 14–90, with higher scores indicating better quality of life.

MRI image acquisition

Baseline neuroimaging data were collected using 3.0 T Siemens superconducting MRI scanners equipped with 64-channel head coils at two sites. Both sagittal T1-weighted magnetization-prepared rapid gradient-echo (MPRAGE) and gradient-recall echo-planar imaging (EPI) sequences were acquired. Participants were scanned in a supine position with earplugs to reduce noise, and were instructed to relax and minimize head movement. The internal and external validation cohorts followed the same scanning protocol, while the discovery cohort utilized a different protocol.

Scanning protocol of the training cohort

T1-MPRAGE sequence: repetition time (TR) = 2530 ms, echo time (TE) = 1.85 ms, flip angle (FA) = 9°, matrix = 256 × 256, slice thickness = 1 mm, gap = 0 mm, number of slices = 192, field of view (FOV) = 256 mm × 256 mm.

EPI sequence: TR = 2000 ms, TE = 30 ms, FA = 90°, matrix = 64 × 64, slice thickness = 3.5 mm, gap = 0.7 mm, number of slices = 33, FOV = 200 mm × 200 mm, 200 time points.

Scanning protocol of internal and external validation cohorts

T1-MPRAGE sequence: repetition time (TR) = 2530 ms, echo time (TE) = 4.21 ms, flip angle (FA) = 7°, matrix = 256 × 256, slice thickness = 1 mm, gap = 0 mm, number of slices = 192, field of view (FOV) = 256 mm × 256 mm.

EPI sequence: TR = 2000 ms, TE = 30 ms, FA = 90°, matrix = 64 × 64, slice thickness = 3.5 mm, number of slices = 33, FOV = 224 mm × 224 mm, 240 time points.

MRI image preprocessing

The PhiPipe tool was used for preprocessing rs-fMRI data[66]. The steps included head motion adjustment using AFNI's 3dvolreg, slice acquisition correction with AFNI's 3dTshift, boundary-based registration with FreeSurfer's bbregister, and masking for brain regions based on T1 processing and BOLD-T1 registration. Motion outliers were interpolated using neighboring volumes. Nuisance signals, including mean white matter and ventricle signals, and Friston's 24-parameter head motion model, were regressed out. Bandpass filtering (0.01–0.1 Hz) was applied, and BOLD images were transformed into MNI152 standard space using combined T1-MNI152 and BOLD-T1 registration.

Functional connectivity analysis

FC analysis was performed using DPARSF (http://rfmri.org/DPARSF). Following preprocessing, the FC matrix of the reward and emotion regulation circuits was constructed by calculating Pearson's correlation between the time courses of each ROI. Brain region coordinates were obtained from the Brainnetome Atlas template and relevant literature, including the nucleus accumbens, striatum, amygdala, globus pallidus, DLPFC, thalamus, OFC, ACC, and parahippocampus, totaling 70 ROIs (Supplementary Table 3).

ROIs were defined as 5 mm radius spheres around the peak coordinates of each cluster. The time series of voxels within each ROI was extracted and averaged. Rs-FC between ROI pairs was calculated using Pearson correlation, followed by Fisher R-to-Z transformation. NaN values for ROIs with poor coverage were replaced with the mean of non-NaN signals from other MDD patients.

Overall model design and computational pipeline

To predict SSRIs treatment outcomes in patients with MDD, this study introduces a hierarchical local-global imaging and clinical feature fusion graph neural network (LGCIF-GNN). The architecture integrates fine-grained brain functional activity with population-level patterns to enhance predictive accuracy and interpretability. An overview of the model architecture and computation pipeline is illustrated in Fig. 1.

LGCIF-GNN takes rs-fMRI signals and clinical variables as input, processing them through a hierarchical local-global graph framework. At the local level, subject-specific brain ROI graphs are constructed to capture individual functional dynamics across ROIs using temporal encoding and connectivity optimization. In parallel, clinical and demographic data are structured and embedded. As a result, for each subject, the model derives both functional and clinical embeddings, which serve as node features in two modality-specific population graphs. In these graphs, each node corresponds to a subject, and edges represent pairwise similarities in either FC patterns or clinical profiles. The two global population graphs are then jointly processed at the global level through modality-specific and shared GCN[67] branches, followed by an attention-based fusion mechanism. The fused representation is then passed to a final multi-layer perceptron (MLP) for individualized prediction of SSRIs treatment response (remission vs. non-remission). Importantly, the model supports interpretability by highlighting discriminative functional connections, clinically relevant traits, and relative modality contributions, thereby enabling both accurate prediction and biologically meaningful insights. In the following, we describe the key components of the framework and elaborate on the underlying computational workflow.

Initially, rs-fMRI signals are preprocessed using the DPABI toolbox to extract ROI-wise BOLD time series for each subject, using a standardized brain atlas. These ROI-specific time sequences are then fed into a bi-GRU[68] encoder, which captures the intrinsic temporal dynamics of each brain region and generates region-level embeddings. These embeddings serve as input to the graph structure optimizer, a module that learns subject-specific FC matrices in a task-driven manner. This approach yields individualized adjacency matrices that capture functionally meaningful neural interactions tailored to the treatment prediction objective, thereby addressing the limitations of conventional static, correlation-based connectivity measures. The resulting adjacency matrix defines the structure of each subject's local ROI-level brain graph, wherein each node corresponds to an ROI defined by a standard brain atlas, and the node features are derived from the corresponding row of the Pearson correlation-based FC matrix. Subsequently, the Local GCN Readout module takes both the node features and the optimized graph structure as input, applies graph convolutional operations coupled with attention mechanisms[69] to update and aggregate regional features, producing a compact graph-level embedding that captures individual functional patterns within the reward and emotion regulation circuits. This embedding is then propagated to the global level, where each subject is represented as a node in the functional population graph.

In parallel, clinical and demographic variables—such as age, sex, clinical assessment scale scores, disease duration, and education level—are numerically encoded and concatenated into structured vectors in the Feature Encoding module. These vectors are used by the CTSE to project clinical features to a shared latent space and compute a pairwise similarity matrix across subjects, forming a clinical population graph that reflects clinical phenotypic proximity between subjects.

Both the subject-level functional embeddings and clinical features are input into the global graph modules as two population graphs. To disentangle and integrate modality-specific and modality-shared information across these population graphs, the global model employs a three-branch design: the Modality-Unique (MU) GCN block extracts distinct representations from each modality using independent multi-hop residual GCNs (MHR-GCNs); the Modality-Shared (MS) GCN block captures common cross-modal patterns via weight sharing; and the Multimodal Attention (M-Attention) block adaptively fuses the outputs of all branches into a unified representation, with attention weights reflecting their relative contributions to final treatment outcome prediction. Finally, an MLP receives the fused embedding and outputs the individualized prediction of the SSRIs treatment response. Additional details on model construction and implementation are provided in the Supplementary Tables 4 and 5.

Through this design, LGCIF-GNN achieves end-to-end optimization of both graph structures and multi-modality feature fusion, leveraging both local fine-grained functional patterns and global population-level similarities to improve the robustness and interpretability of multimodal predictive modeling.

Local graph construction and encoding

In local graphs, different functional ROIs of the reward and emotion regulation circuit, according to neurological knowledge, were defined as nodes, and the FC between these ROIs is defined as edges. Specifically, 70 ROIs were used as nodes, and the FC strength between nodes was established as edge weights to construct the local graph, with each graph corresponding to one subject. The FC derived from the correlation computations was utilized to initialize the graph structure of the local graph.

As depicted in Fig. 1a, the local ROI-based GNN consists of three main components: (1) a GRU regional time series encoder, (2) an FC matrix adjustment optimizer and graph generator, and (3) a readout module for generating graph-level embeddings and local predictions.

The temporal resolution of fMRI-derived time series signals is typically low. To extract the temporal features of the fMRI BOLD signals for each ROI while avoiding overfitting, this paper employs a lightweight bi-GRU[68] as the temporal encoder. Specifically, for an input BOLD time series $X \in R^{n \times t}$ of a subject, where n represents the number of ROIs and t is the length of the time series, the GRU encoder produces a regional embedding for each ROI, $h_e = Encoder(x)$, where $h_e \in \mathbb{R}^{n \times d}$ and d is the output dimension of the GRU encoder.

To enable the graph structure defined by the FC matrix to be continuously optimized and adjusted in the network training stage, rather than being determined solely by coarse and inflexible correlation computation methods, a graph structure optimizer has been designed. This module constructs the adjacency matrix A based on the node feature vectors. The cosine similarity between the feature vectors of the ith and jth nodes is used as the weight at the position A_{ij} of the adjacency matrix: $A_{ij} = h_e \cdot h_e^T$.

After the first two modules, the local graph structure has been learned and optimized. The local readout module updates the node features based on the GCN[67] and employs an attention mechanism to perform a weighted aggregation of the features across all nodes in the entire local graph, thereby generating a graph-level embedding. This graph-level embedding maps the functional characteristics of various brain regions and the inter-regional FC patterns within the reward and emotion regulation circuit of an individual into a hidden space. Specifically, the node feature h_i of node i was initialized with the ith row of the Pearson correlation FC matrix A_p .

$$\mathrm{H}_l^0 = \left[h_{l1}^0, h_{l2}^0, \ldots, h_{\mathrm{ln}}^0\right] \tag{1}$$

$$\mathrm{h}_\mathrm{i}^0 = A_p[i, :] \tag{2}$$

A 3-layer GCN was used to update the node feature, and the kth GCN layer is defined referring to the GCN proposed by Kipf and Welling[67] as:

$$H_l^k = \sigma\left(GCN\left(H_l^{k-1}, A\right)\right) = \sigma\left(D^{-\frac{1}{2}}AD^{-\frac{1}{2}}H_l^{k-1}W^k\right) \tag{3}$$

Where D is the diagonal matrix, A is the adjacent matrix derived from the graph structure optimizer and W^t is a trainable weight matrix of the kth layer, which is a two-layer MLP in our implementation. The final embedding of the whole local graph is the concatenation of node embedding weighted by the attention score.

$$a_{ROI} = Attention(A) = softmax\left(\sum_{j=0}^{n} A_{ij}\right) \cdot n \tag{4}$$

$$H_{lG} = concat\left(a_{ROI}H_l^k\right) \tag{5}$$

The local classification is determined by a 3-layer MLP classification head:

$$\hat{y}_l = MLP\left(H_{lG}\right) \tag{6}$$

Global graph construction and cross-modal fusion

Each subject was used as a node to construct two global graphs, including a functional graph and a demographic characteristics graph. The global feature vector (readout) of the subject's fMRI signals learned from the local graph was used as the node feature of the functional graph. The demographic characteristics graph was constructed by using the subject's age, sex, duration, education level, YMSR, HAMD, and QLE scale scores as node features. Subsequently, an attention mechanism was used to perform multimodal feature fusion of the two graphs. The node-level classification was used as the supervisory signal by the local graph.

As illustrated in Fig. 1b, the global GNN updates node features on both the functional population graph $G_f = (V_f, E_f, W_f)$ and the clinical population graph $G_c = (V_c, E_c, W_c)$, facilitating the fusion of clinical and imaging modalities, and performs node-level classification to obtain the final prediction of treatment outcome. In these population graphs, each node represents a patient, and the weights of the edges between nodes indicate the similarity between two patients in the corresponding modality. In the functional population graph, the node features are the functional embeddings generated by the local GNN, and the edge weights are the cosine similarity between the features of two nodes. In the clinical population graph, the node features remain as the functional embeddings produced by the local GNN, while the edge weights are determined by the CTSE based on clinical information such as age, sex, education level, disease duration, YSMR, QLES, and HAMD scale scores of the two patients. In the global GNN, the modality-unique GNN module is designed to extract unique features from each of the two modalities. The modality-shared GNN module is utilized to capture the common features across both modalities. Meanwhile, the modality-attention fusion module is employed to facilitate the fusion of multimodal features.

The CTSE starts by accepting the clinical feature vector h_c, a concatenation of age, sex, education level, disease duration, and the sub-item total scores of clinical scales, as input and maps each input to a common latent space $h_{ci} \in \mathbb{R}^{D_c}$. In the $D_c = 128$ dimensional space, the cosine similarity can be better applied. The projection network is a 1-layer MLP to avoid overfitting. Thereby, the CTSE calculates the similarity between node i and j as:

$$W_c(i, j) = \frac{\cos\left(MLP\left(h_{ci}\right), MLP\left(h_{cj}\right)\right) + 1}{2} \tag{7}$$

where cos denotes the cosine similarity between two input vectors.

In this study, we propose the MHR-GCN to address the over-smoothing issue and enhance the aggregation of multi-scale information in GCNs, inspired by the snowball GCN block. Our network architecture incorporates residual connections to facilitate the training of deeper networks, mitigating the vanishing gradient problem and enabling the training of GCNs beyond four layers. The MHR-GCN block concatenates the output of each layer before the final GCN layer to enrich feature representation, where the output of each hidden layer is a description of the center node with its different-hop neighbors. Residual connections are added after each GCN layer to learn the residual information, and the final output layer aggregates features from all hidden layers to produce node embeddings. This design allows the MHR-GCN to effectively aggregate information from various receptive fields, providing comprehensive node representations in the Subject-Graph, thus mitigating over-smoothing and enhancing the network's ability to learn complex graph representations. The structure of snowball MHR-GCN is as follows:

$$H_g^0 = H_{lG} \tag{8}$$

$$H_g^{l+1} = H_g^l + Tanh\left(H_g^l W_l\right),\ l = 0, 1, 2, \ldots, N-1 \tag{9}$$

$$C = Tanh\left(\left[H_g^0, H_g^1, \ldots, H_g^n\right] W_n\right) \tag{10}$$

$$H_{gG} = normalize\left(L^p C W_c\right) \tag{11}$$

Where N is the number of MHR-GCN layers, W_l, W_n, W_c is the trainable matrix, $H_g^0, H_g^1, \ldots, H_g^l$ are extracted features, $p \in \{0, 1\}$, H_{gG} is the global Subject-Graph embedding of one specific modality. When $p = 0,\ L^p = I$ and when $p = 1,\ L^p = L = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$, which means that we project C back onto the Fourier basis, which is necessary when the graph structure encodes much information.

The MU-GCN is applied to extract modality-unique embeddings, which are defined as follows:

$$H_u^f = MHR - GCN\left(X^f, A^f\right) = MHR - GCN\left(H_{lG}, A^f\right) \tag{12}$$

$$H_u^c = MHR - GCN(X^c, A^c) = MHR - GCN\left(H_{lG}, A^c\right) \tag{13}$$

where X^fandX^c are the node features for rs-fMRI and clinical modalities, respectively. And H_u^fandH_u^c are the modality-unique representations. The weights of the three SnowballGCN networks are independent of each other, making it possible to extract unique features more effectively.

Although the data structures and semantic information of various modalities possess distinct characteristics, it is far from trivial to completely disentangle these diverse data types. When performing the same task, data from different modalities often contain overlapping information. Extracting this shared information not only aids in distilling high-quality features for problem-solving but also reduces redundancy during the integration of multimodal information. To achieve this objective, we introduce the MS-GCN module, which facilitates the sharing of weight matrices across different modalities during the execution of the MHR-GCN operation. The specific formulations are as follows:

$$H_s^f = MHR - GCN\left(X^f, A^f\right) = normalize\left(L^{fp} C^f W_{cs}\right) \tag{14}$$

$$H_s^c = MHR - GCN(X^c, A^c) = normalize\left(L^{cp} C^c W_{cs}\right) \tag{15}$$

where H_s^f and H_s^c are the modality-common representations for rs-fMRI and clinical modalities, respectively. And W_{cs} is the shared trainable matrix. By sharing weights in this way, modality-common features can be filtered out. The final common embedding is obtained by the weighted sum of the two $H_s = \alpha H_s^f + \beta H_s^c$, where α and β are hyperparameters measuring the importance of each modality's common embedding. In the implementation, we set $\alpha = \beta = 0.5$ to pay equal attention to all modalities.

The varying impact of various types of information on the ultimate treatment outcome prediction depends on the specific illness being targeted. In order to focus more on the informative methods and relegate the less critical ones to a supporting function, we employ an M-Attention Block on two unique embeddings H_u^f and H_u^c and one shared embedding H_s.

$$a_f, a_c, a_s = Attention\left(H_u^f, H_u^c, H_s\right) \tag{16}$$

$$e_i = \mathrm{Tanh}\left(W_{ai} H_u^i + b_i\right) \tag{17}$$

$$a_i = \frac{\exp\left(e_i\right)}{\exp\left(e_f\right) + \exp\left(e_c\right) + \exp\left(e_s\right)}, \; i \in \{f, c, s\} \tag{18}$$

Once the attention score has been derived, the final embedding can be computed by integrating the representation with the weight as follows:

$$H = a_f H_u^f + a_c H_u^c + a_s H_s \tag{19}$$

Finally, an MLP layer is employed for class prediction.

$$\hat{y}_g = MLP(H) \tag{20}$$

Training and implementation details

The optimization objective of the proposed LGCIF-GNN model integrates three complementary loss components: a classification loss, a modality-unique decorrelation loss, and a modality-shared consistency loss:

$$L = L_{cls} + L_{unique} + L_{shared} \tag{21}$$

First, to supervise the predictive task, we adopt a cross-entropy loss framework. Since our architecture supports prediction at both local and global levels, the classification loss is formulated as a weighted combination of both contributions:

$$L_{cls} = L_{cls_{global}} + \lambda L_{cls_{local}} \tag{22}$$

Here, λ is a hyperparameter regulating the influence of the local loss term. In our implementation, we set λ = 0.2 to prioritize the global classification signal during training, while still retaining local-level supervision to enhance representation learning.

Second, to disentangle modality-specific from modality-invariant information, we introduce a statistical independence constraint between the learned modality-unique embeddings H_u and shared embeddings H_s. This is achieved by minimizing the Hilbert-Schmidt Independence Criterion (HSIC)[70], which quantifies dependence between distributions in a reproducing kernel Hilbert space (RKHS):

$$HSIC\left(H_u, H_s\right) = (m-1)^{-2} tr\left(K_s R K_u R\right) \tag{23}$$

where $K\left(H^i, H^j\right) = \; <\phi\left(H^i\right), \phi\left(H^j\right)>$ is the kernel function mapping the input embeddings into an RKHS, $R = I - \frac{1}{m} ee^T$ is the centering matrix, I is the identity matrix, and e is an all-ones vector. The total decorrelation loss aggregates HSIC values across multiple modality pairs:

$$L_{unique} = HSIC\left(H_u^f, H_s\right) + HSIC\left(H_u^c, H_s\right) \tag{24}$$

Additionally, to encourage alignment among modality-invariant embeddings, we enforce similarity across shared representations derived from distinct modalities. After L_2-normalizing each embedding matrix, we compute pairwise differences between their similarity matrices:

$$N_s = H_s \cdot H_s^T \tag{25}$$

$$L_{shared} = ||N_s^f - N_s^c||^2 \tag{26}$$

The proposed model was implemented in Python 3.7 and PyTorch 1.12.1, utilizing an NVIDIA RTX 4090 GPU. For optimization, the Adam optimizer was employed with an initial learning rate of 5e-3, which was halved every 100 epochs. Across all experiments, the model was trained for a maximum of 200 epochs. The weight hyperparameter for the local classification loss was fixed at 0.2, giving precedence to the global classification loss during optimization. Detailed architectural and implementation specifications are provided in Supplementary Table 5.

Performance validation

To ensure accuracy and robustness in feature selection, the model's performance was evaluated using 5-fold cross-validation, along with metrics including the AUC of the ROC, accuracy, sensitivity, and specificity. The training dataset was randomly divided into five equal parts. In each cycle, four parts were used as sub-training sets to build the model, and the remaining part was used to test the model's performance. This process was repeated five times so that each subset served as a test set once. The optimal hyperparameter settings determined during cross-validation were applied to construct the final model using the entire training dataset. The final model was then tested on the internal and external validation sets for further verification. All ML analyses were conducted using Python.

To assess the model's generalization capability, we conducted validation on two held-out datasets. The internal validation set ($n = 66$) provided an in-site assessment of predictive stability on temporally distinct data. In contrast, the external validation set ($n = 49$) served to evaluate the model's cross-center generalizability. Specifically, it tested the robustness of the learned representations under differences in MRI scanner types, acquisition protocols, and patient population characteristics—factors known to challenge the reproducibility of neuroimaging-based biomarkers.

Additionally, an ablation study was performed to confirm the significance of radiographic features in improving model performance. During the ablation study, only the clinical features of 164 MDD patients were used in 5-fold cross-validation to assess their independent contribution.

Interpretation of functional imaging

As described in the section "Local graph construction and encoding", the Graph Structure Optimizer in this study was designed to extract temporal sequence features from brain regions within the reward and emotion regulation circuits during training of the efficacy prediction network. This process constructs and optimizes an FC matrix specifically for predicting treatment outcomes. Through joint optimization of the local and global networks, the learned adjacency matrix encapsulates deeper temporal information while accounting for the constraints of functional brain imaging and clinical feature similarities among patients. This matrix reflects FC patterns relevant to treatment outcome prediction.

To understand the functional characteristics of the reward and emotion regulation circuits in non-remission and remission groups, FC matrices generated by the trained model for 164 subjects were analyzed. Subjects were grouped by treatment outcomes (non-remission or remission), and their FC matrices were averaged separately to obtain mean FC matrices for each group. These average matrices represent the general FC patterns within the two populations. A differential FC matrix, created by subtracting the mean non-remission FC matrix from the mean remission FC matrix, highlights differences in connectivity patterns between the groups.

The differential FC matrix was analyzed to identify brain regions and connections with the most significant differences between remission and non-remission groups. The specific rows and columns corresponding to the top five highest and lowest values in the matrix represent the FCs most enhanced or diminished in remission patients compared to non-remission patients. To determine the most impactful brain regions, the absolute values of the differential FC matrix were summed across rows to calculate node degrees. Higher node degrees indicated greater discrepancies in functional connections between the groups. Using this method, the 10 brain regions with the highest node degrees were identified as the most indicative for predicting treatment outcomes.

Assessment of modality importance

To evaluate the predictive contributions of different data modalities, the Modality-Attention (M-Attention) module within the global GNN model was utilized. This module fuses functional imaging and clinical data modalities, generating three distinct feature embeddings: unique imaging features, unique clinical features, and shared features between the two. The M-Attention module performs weighted fusion of these embeddings based on attention weights, allowing an assessment of the relative contributions of unique and shared features to the prediction process.

To compare the impact of specific clinical information and functional signals from different brain regions on treatment outcome predictions, a feature masking strategy was employed. During the inference phase on the independent test set, elements corresponding to clinical features were set to zero, or time series signals of specific ROIs were nullified. The model's performance was evaluated using the masked features as input, and its efficacy was assessed by the AUROC.

Performance loss caused by masking individual features was visualized as a heatmap. Features with greater predictive importance caused larger reductions in AUROC, indicated by deeper blue tones in the heatmap. This analysis identified the clinical and functional features most critical to predicting antidepressant treatment efficacy.

Statistical analysis

Demographic data were analysed separately for the training and independent validation sets using SPSS Statistics 26.0. Age and clinical characteristics were compared using the Mann–Whitney *U* test or two-sample *t*-test. Chi-square tests were conducted to evaluate differences in sex and education level between the remission and non-remission groups.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code availability

The full implementation of the proposed LGCIF-GNN model is publicly available at https://github.com/ATP-BME/LGCIF-GNN.

Received: 8 April 2025; Accepted: 24 July 2025;
Published online: 12 August 2025

References

1. Maslej, M. M. et al. Individual differences in response to antidepressants: a meta-analysis of placebo-controlled randomized clinical trials. *JAMA Psychiatry* **78**, 490–497 (2021).
2. Rathnam, S. et al. Heterogeneity in antidepressant treatment and major depressive disorder outcomes among clinicians. *JAMA Psychiatry* **81**, 1003–1009 (2024).
3. Park, L. T. & Zarate, C. A. Jr. Depression in the primary care setting. *N. Engl. J. Med.* **380**, 559–568 (2019).
4. Saveanu, R. et al. The International Study to Predict Optimized Treatment in Depression (iSPOT-D): outcomes from the acute phase of antidepressant treatment. *J. Psychiatr. Res.* **61**, 1–12 (2015).
5. Whiteford, H. A. et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *Lancet* **382**, 1575–1586 (2013).
6. Grzenda, A. et al. Machine learning prediction of treatment outcome in late-life depression. *Front. Psychiatry* **12**, 738494 (2021).
7. Wu, H. et al. Prediction of remission among patients with a major depressive disorder based on the resting-state functional connectivity of emotion regulation networks. *Transl. Psychiatry* **12**, 391 (2022).
8. McMakin, D. L. et al. Anhedonia predicts poorer recovery among youth with selective serotonin reuptake inhibitor treatment-resistant depression. *J. Am. Acad. Child Adolesc. Psychiatry* **51**, 404–411 (2012).
9. Cao, B. et al. Pharmacological interventions targeting anhedonia in patients with major depressive disorder: a systematic review. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* **92**, 109–117 (2019).
10. Ma, Y. et al. Altered neural activity in the reward-related circuit associated with anhedonia in mild to moderate major depressive disorder. *J. Affect Disord.* **345**, 216–225 (2024).
11. Lai, C. H. Fronto-limbic neuroimaging biomarkers for diagnosis and prediction of treatment responses in major depressive disorder. *Prog. Neuro-Psychopharmacol. Biol. Psychiatry* **107**, 110234 (2021).
12. Phillips, M. L. et al. Identifying predictors, moderators, and mediators of antidepressant response in major depressive disorder: neuroimaging approaches. *Am. J. Psychiatry* **172**, 124–138 (2015).
13. Nichols, E. S. et al. Emotion regulation in emerging adults with major depressive disorder and frequent cannabis use. *Neuroimage Clin.* **30**, 102575 (2021).
14. Langenecker, S. A. et al. Frontal and limbic activation during inhibitory control predicts treatment response in major depressive disorder. *Biol. Psychiatry* **62**, 1272–1280 (2007).
15. Gong, L. et al. Disrupted reward and cognitive control networks contribute to anhedonia in depression. *J. Psychiatr. Res.* **103**, 61–68 (2018).
16. Liu, R. et al. Anhedonia correlates with functional connectivity of the nucleus accumbens subregions in patients with major depressive disorder. *Neuroimage Clin.* **30**, 102599 (2021).
17. Haber, S. N. & Knutson, B. The reward circuit: linking primate anatomy and human imaging. *Neuropsychopharmacology* **35**, 4–26 (2010).
18. Hahn, A. et al. Dynamics of human serotonin synthesis differentially link to reward anticipation and feedback. *Mol. Psychiatry* https://doi.org/10.1038/s41380-024-02696-1 (2024).
19. Hanuka, S. et al. Reduced anhedonia following internet-based cognitive-behavioral therapy for depression is mediated by enhanced reward circuit activation. *Psychol. Med.* **53**, 4345–4354 (2023).
20. Sheline, Y. I. et al. Treatment course with antidepressant therapy in late-life depression. *Am. J. Psychiatry* **169**, 1185–1193 (2012).
21. Gerlach, A. R. et al. MRI predictors of pharmacotherapy response in major depressive disorder. *Neuroimage Clin.* **36**, 103157 (2022).
22. Vai, B. et al. Fronto-limbic effective connectivity as possible predictor of antidepressant response to SSRI administration. *Eur. Neuropsychopharmacol.* **26**, 2000–2010 (2016).

23. Godlewska, B. R., Browning, M., Norbury, R., Cowen, P. J. & Harmer, C. J. Early changes in emotional processing as a marker of clinical response to SSRI treatment in depression. *Transl. Psychiatry* **6**, e957 (2016).
24. Rost, N., Binder, E. B. & Brückl, T. M. Predicting treatment outcome in depression: an introduction into current concepts and challenges. *Eur. Arch. Psychiatry Clin. Neurosci.* **273**, 113–127 (2023).
25. Helmreich, I. et al. Hamilton depression rating subscales to predict antidepressant treatment outcome in the early course of treatment. *J. Affect Disord.* **175**, 199–208 (2015).
26. Iancu, S. C., Wong, Y. M., Rhebergen, D., van Balkom, A. & Batelaan, N. M. Long-term disability in major depressive disorder: a 6-year follow-up study. *Psychol. Med.* **50**, 1644–1652 (2020).
27. Sheehan, D. V., Nakagome, K., Asami, Y., Pappadopulos, E. A. & Boucher, M. Restoring function in major depressive disorder: a systematic review. *J. Affect Disord.* **215**, 299–313 (2017).
28. Hohls, J. K., König, H. H., Quirke, E. & Hajek, A. Anxiety, depression and quality of life—a systematic review of evidence from longitudinal observational studies. *Int. J. Environ. Res. Public Health* **18**, https://doi.org/10.3390/ijerph182212022 (2021).
29. Tian, S. et al. Predicting escitalopram monotherapy response in depression: The role of anterior cingulate cortex. *Hum. Brain Mapp.* **41**, 1249–1260 (2020).
30. Patel, M. J. et al. Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. *Int. J. Geriatr. Psychiatry* **30**, 1056–1067 (2015).
31. Nguyen, K. P. et al. Patterns of pretreatment reward task brain activation predict individual antidepressant response: key results from the EMBARC randomized clinical trial. *Biol. Psychiatry* **91**, 550–560 (2022).
32. Lei, D. et al. Graph convolutional networks reveal network-level functional dysconnectivity in schizophrenia. *Schizophr. Bull.* **48**, 881–892 (2022).
33. Chang, Q. et al. Classification of first-episode schizophrenia, chronic schizophrenia and healthy control based on brain network of mismatch negativity by graph neural network. *IEEE Trans. Neural Syst. Rehabil. Eng.* **29**, 1784–1794 (2021).
34. Li, X. et al. BrainGNN: interpretable brain graph neural network for fMRI analysis. *Med. Image Anal.* **74**, 102233 (2021).
35. Ding, Y., Robinson, N., Tong, C., Zeng, Q. & Guan, C. LGGNet: learning from local-global-graph representations for brain–computer interface. *IEEE Trans. Neural Netw. Learn. Syst.* **35**, 9773–9786 (2024).
36. Pain, S., Roy, S., Sarma, M. & Samanta, D. Detection of alcoholism by combining EEG local activations with brain connectivity features and Graph Neural Network. *Biomed. Signal Process. Control* **85**, 104851 (2023).
37. Kan, X. et al. Brain network transformer. *Adv. Neural Inf. Process. Syst.* **35**, 25586–25599 (2022).
38. Thapaliya, B. et al. Brain networks and intelligence: a graph neural network based approach to resting state fMRI data. *Med. Image Anal.* **101**, 103433 (2025).
39. Zheng, K., Yu, S. & Chen, B. CI-GNN: a Granger causality-inspired graph neural network for interpretable brain network-based psychiatric diagnosis. *Neural Netw.* **172**, 106147 (2024).
40. Tong, W. et al. fMRI-based brain disease diagnosis: a graph network approach. *IEEE Trans. Med. Robot. Bion.* **5**, 312–322 (2023).
41. Zhang, H. et al. Classification of brain disorders in rs-fMRI via local-to-global graph neural networks. *IEEE Trans. Med. Imaging* **42**, 444–455 (2023).
42. Thapaliya, B. et al. Graph-based deep learning models in the prediction of early-stage Alzheimer's. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),* 1–5, https://doi.org/10.1109/EMBC53108.2024.10782267. IEEE. (2024).
43. Azevedo, T. et al. A deep graph neural network architecture for modelling spatio-temporal dynamics in resting-state functional MRI data. *Med. Image Anal.* **79**, 102471 (2022).
44. Thapaliya, B. et al. DSAM: a deep learning framework for analyzing temporal and spatial dynamics in brain networks. *Med. Image Anal.* **101**, 103462 (2025).
45. Chen, T., Hong, R., Guo, Y., Hao, S. & Hu, B. MS^2-GNN: exploring GNN-based multimodal fusion network for depression detection. *IEEE Trans. Cybern.* **53**, 7749–7759 (2023).
46. Liu, S. et al. An objective quantitative diagnosis of depression using a local-to-global multimodal fusion graph neural network. *Patterns* **5**, 101081 (2024).
47. Collins, G. S. et al. Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ* **384**, e074819 (2024).
48. Poirot, M. G. et al. Treatment response prediction in major depressive disorder using multimodal MRI and clinical data: secondary analysis of a randomized clinical trial. *Am. J. Psychiatry* **181**, 223–233 (2024).
49. Long, F. et al. Predicting treatment outcomes in major depressive disorder using brain magnetic resonance imaging: a meta-analysis. *Mol. Psychiatry* https://doi.org/10.1038/s41380-024-02710-6 (2024).
50. Vachez, Y. M. et al. Ventral arkypallidal neurons inhibit accumbal firing to promote reward consumption. *Nat. Neurosci.* **24**, 379–390 (2021).
51. Vollmer, K. M. et al. An opioid-gated thalamoaccumbal circuit for the suppression of reward seeking in mice. *Nat. Commun.* **13**, 6865 (2022).
52. Luo, Y. J. et al. Ventral pallidal glutamatergic neurons regulate wakefulness and emotion through separated projections. *iScience* **26**, 107385 (2023).
53. Ji, Y. W. et al. Plasticity in ventral pallidal cholinergic neuron-derived circuits contributes to comorbid chronic pain-like and depression-like behaviour in male mice. *Nat. Commun.* **14**, 2182 (2023).
54. Gray, J. P., Müller, V. I., Eickhoff, S. B. & Fox, P. T. Multimodal abnormalities of brain structure and function in major depressive disorder: a meta-analysis of neuroimaging studies. *Am. J. Psychiatry* **177**, 422–434 (2020).
55. Talati, A. et al. Putamen structure and function in familial risk for depression: a multimodal imaging study. *Biol. Psychiatry* **92**, 932–941 (2022).
56. Yuan, Z. et al. A corticoamygdalar pathway controls reward devaluation and depression using dynamic inhibition code. *Neuron* **111**, 3837–3853.e3835 (2023).
57. Alexander, L., Hawkins, P. C. T., Evans, J. W., Mehta, M. A. & Zarate, C. A. Jr Preliminary evidence that ketamine alters anterior cingulate resting-state functional connectivity in depressed individuals. *Transl. Psychiatry* **13**, 371 (2023).
58. Tartt, A. N., Mariani, M. B., Hen, R., Mann, J. J. & Boldrini, M. Dysregulation of adult hippocampal neuroplasticity in major depression: pathogenesis and therapeutic implications. *Mol. Psychiatry* **27**, 2689–2699 (2022).
59. Han, H. et al. Implications of neurogenesis in depression through BDNF: rodent models, regulatory pathways, gut microbiota, and potential therapy. *Mol. Psychiatry* https://doi.org/10.1038/s41380-025-03044-7 (2025).
60. Wu, F. et al. A visual cortical-lateral posterior thalamic nucleus circuit regulates depressive-like behaviors in male mice. *Nat. Commun.* **16**, 1395 (2025).
61. Wang, X. Y. et al. The thalamic reticular nucleus-lateral habenula circuit regulates depressive-like behaviors in chronic stress and chronic pain. *Cell Rep.* **42**, 113170 (2023).
62. Xue, L. et al. Shared and unique imaging-derived endo-phenotypes of two typical antidepressant-applicative depressive patients. *Eur. Radiol.* **33**, 645–655 (2023).
63. Tang, H. et al. Electrophysiological predictors of early response to antidepressants in major depressive disorder. *J. Affect Disord.* **365**, 509–517 (2024).

64. Zhang, H. et al. Predicting SSRI-Resistance: Clinical Features and tagSNPs Prediction Models Based on Support Vector Machine. *Front. Psychiatry* **11**, 493 (2020).
65. Endicott, J., Nee, J., Harrison, W. & Blumenthal, R. Quality of Life Enjoyment and Satisfaction Questionnaire: a new measure. *Psychopharmacol. Bull.* **29**, 321–326 (1993).
66. Hu, Y. et al. PhiPipe: a multi-modal MRI data processing pipeline with test-retest reliability and predicative validity assessments. *Hum. Brain Mapp.* **44**, 2062–2084 (2023).
67. Wang, H. et al. Scientific discovery in the age of artificial intelligence. *Nature* **620**, 47–60 (2023).
68. Chung, J., Gulcehre, C., Cho, K. & Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proceedings Advances in Neural Information Processing Systems*. Curran Associates, Inc. (2014).
69. Vaswani, A. Attention is all you need. In *Proceedings Advances in Neural Information Processing Systems.* Curran Associates, Inc. (2017).
70. Gretton, A. et al. Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory*, Vol. 3734, 63–77 (Springer, 2005).

Acknowledgements

This work was supported by STI2030-Major Projects (2021ZD0200600), Beijing Research Ward Excellence Program (BRWEP2024W072120102), Capital's Funds for Health Improvement and Research (CFH2020-4-2125), and Beijing Hospitals Authority Youth Programme (QML20211901).

Author contributions

R.L., A.Y., Y.Z., C.J., and G.W. conceived and designed the study. J.Z., K.Q., H.Q., R.L., L.Z., and J.C. collected the data. R.L., X.H., S.L., and Y.Z. analyzed the data. X.H. and S.L. drafted the manuscript. C.J., A.Y. and Y.Z. revised the draft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01912-8.

Correspondence and requests for materials should be addressed to Cheng Jin, Aihong Yu or Gang Wang.

Reprints and permissions information is available at http://www.nature.com/reprints