Patterns

Self-supervised graph representation learning integrates multiple molecular networks and decodes gene-disease relationships

Highlights

- Integrating multiple molecular networks to improve signal-to-noise ratio
- Self-supervised representation learning at both node level and context level
- Task-specific re-training using graph attention network converges efficiently
- Achieves superior performance to refine disease gene reprioritization

Authors

Yi Wang, Zijun Sun, Qiushun He, Jiwei Li, Ming Ni, Meng Yang

Correspondence

yangmeng1@mgi-tech.com

In brief

Integrating multiple molecular networks is essential to decipher gene function under specific biological context, refine GWAS (genome-wide association study) hits, and offer mechanistic insights via decoding diseases from genome to networks to phenotypes. In this paper, Graphene is introduced as a self-supervised learning framework to aggregate information from biological networks.

Wang et al., 2023, Patterns *4*, 100651
January 13, 2023
https://doi.org/10.1016/j.patter.2022.100651

Patterns

CellPress
OPEN ACCESS

Article

Self-supervised graph representation learning integrates multiple molecular networks and decodes gene-disease relationships

Yi Wang,[1] Zijun Sun,[3] Qiushun He,[1] Jiwei Li,[4] Ming Ni,[1,2] and Meng Yang[1,5,*]
[1]MGI, BGI-Shenzhen, Shenzhen, China
[2]MGI-QingDao, BGI-Shenzhen, Qingdao, China
[3]Computer Center, Peking University, Beijing, China
[4]Department of Computer Science, Zhejiang University, Hangzhou, China
[5]Lead contact
*Correspondence: yangmeng1@mgi-tech.com
https://doi.org/10.1016/j.patter.2022.100651

THE BIGGER PICTURE With the recent progress of high-throughput experimental techniques, physical interactions and functional associations of genes and proteins are accumulating into multiple molecular networks. Effective integration of these networks and extraction of biological insight remains a long-standing challenge. The two-step GNN (graph neural network) approach (Graphene) introduced here offers a self-supervised solution and validates its utility in a range of disease gene sets.

Proof-of-Concept: Data science output has been formulated, implemented, and tested for one domain/problem

SUMMARY

Leveraging molecular networks to discover disease-relevant modules is a long-standing challenge. With the accumulation of interactomes, there is a pressing need for powerful computational approaches to handle the inevitable noise and context-specific nature of biological networks. Here, we introduce Graphene, a two-step self-supervised representation learning framework tailored to concisely integrate multiple molecular networks and adapted to gene functional analysis via downstream re-training. In practice, we first leverage GNN (graph neural network) pre-training techniques to obtain initial node embeddings followed by re-training Graphene using a graph attention architecture, achieving superior performance over competing methods for pathway gene recovery, disease gene reprioritization, and comorbidity prediction. Graphene successfully recapitulates tissue-specific gene expression across disease spectrum and demonstrates shared heritability of common mental disorders. Graphene can be updated with new interactomes or other omics features. Graphene holds promise to decipher gene function under network context and refine GWAS (genome-wide association study) hits and offers mechanistic insights via decoding diseases from genome to networks to phenotypes.

INTRODUCTION

Diseases or traits involve molecules interacting within cellular networks and pathways under certain biological contexts. Understanding functional interdependencies of genes and proteins can provide a system-level view of how genetic alterations dysregulate relevant pathways or biological processes, and further lead to disease phenotypes.[1] A classical insight behind network biology is that genes or proteins presenting similar topological neighborhood patterns are more likely to be correlated, which enables knowledge refinement for known molecules and property inference for unknown ones through "guilt by association" principle. There has been a recent community benchmark effort to evaluate disease module discovery methods on various network configurations.[2] A network-based method has been utilized to reprioritize statistical signals from disease-focused genome-wide association studies (GWAS). For example, the NetWAS[3] framework leverages tissue-specific networks in combination with marginally significant GWAS hits as input for deploying a machine learning model to rank candidate genes.

The NAGA[4] framework harnessed a compositive molecular network to implement a propagation approach to boost GWAS results for eight diseases. iRIGs[5] reprioritized schizophrenia (SCZ) GWAS genes by using a Bayesian framework to integrate multi-omics data and a protein-protein interaction (PPI) network. Buphamalai et al.[6] constructed a multiplex network organized into hierarchical layers spanning different omics levels and revealed that rare diseases also exhibit network signatures similar to complex diseases through propagation-based algorithms.[7] A comprehensive review[8] of network-based disease gene prioritization categorizes existing computational efforts into three major classes, including network diffusion methods, traditional machine learning methods with handcrafted features, and graph representation learning methods. Notably, Set2Gaussian[9] embeds gene sets as a multivariate Gaussian distribution in low-dimensional space based on genes' proximity in the PPI network, manifesting stronger expressive power over traditional network diffusion methods.

The utilities of these network methods strongly rely on the quality and coverage of available molecular networks. Recent advances in high-throughput experimental platforms and computational techniques have enabled characterizing heterogeneous genome-scale networks, including physical interactions (for example, PPI,[10] signaling, and regulatory networks) and functional associations (for example, gene co-expression, genetic dependencies, co-evolution, and phylogenetic patterns). Huang et al.[11] systematically evaluated 21 human interaction networks covering various types of interactions, concluding that ConsensusPathDB,[12] GIANT[13] (now available as Humanbase), and STRING[10] perform best to recover disease gene sets and the larger network as a whole outweighs the drawbacks of potential false positives, and recurrent but nuanced signals can be amplified. Picart et al.[14] also emphasized the merit of introducing a larger network. The ever-growing repositories of interactomes require developing methods to combine these networks while simultaneously tackling inherent noise and incompleteness among them. Huang et al. pioneered a parsimonious composite network (PCNet)[11] with high efficiency. Mashup[15] leverages random walks with restart (RWR)[16] for each network, then optimizes a consistent dimension reduction function to derive compact network integration as low-dimensional vectors for each gene or protein to be plugged into downstream functional tasks. Several other methods have been proposed to integrate multiple networks. Gao et al.[17] used multi-view representation learning to cluster network data. Ma et al.[18] adopted matrix decomposition to integrate heterogeneous networks. Lin et al.[19] combined node2vec[20] and matrix factorization to analyze cancer attributed networks. DeepMNE-CNN[21] developed a semi-supervised autoencoder method to integrate RWR-derived embeddings from multiple networks and predict gene function using convolutional network.

Graph neural network (GNN) has recently emerged to incorporate graph structures into a deep learning framework.[22] To represent genes as nodes and their interactions as edges, GNN naturally captures the interdependent relationships of comprised molecules within networks, and node embeddings are learned by iteratively updating the information aggregated from its adjacent neighbors. According to the different ways with which GNN propagates information, the architectures of GNN include graph convolutional networks (GCNs),[23] GraphSAGE,[24] GAT,[25] GIN,[26] etc. In recent years, GNN had demonstrated effectiveness in biologically related tasks, such as drug-target interactions[27] and disease identification.[28] For example, EMOGI[29] leverages GCNs to integrate topologic features from PPI networks with multi-omics pan-cancer data to propose novel cancer genes. Furthermore, multimodal GNNs incorporating more than one type of node enables multi-relational link prediction. Decagon[30] constructed a heterogeneous gene-drug network to predict polypharmacy side effects via decoding links between drug pairs.

Self-supervised learning (SSL) has recently provided a promising paradigm toward human-level intelligence and achieved great success in the domains of natural language processing and computer vision, such as BERT,[31] SimCLR,[32] and MAE.[33] SSL firstly pre-trains a model on a well-designed pretext task, then fine-tunes it on a specific downstream task of interest. Biology networks contain tremendous intrinsic information, and applying SSL to network biology shows promise to directly learn from interacted biological molecules. Due to the non-Euclidean data structure, graph SSL has several particular characteristics for which pre-training can be implemented at the level of individual nodes and entire graphs to derive useful local and global representations simultaneously.[34] A recent review article[35] divided the pre-training task into four categories, including generative, contrastive, and auxiliary property-based, as well as their hybridizations. Avoiding negative generalizability during knowledge transfer from pre-training task to downstream objectives is the key consideration for self-supervised graph representation learning.[36]

Inspired by the recent progress of self-supervised GNN,[34] we propose Graphene, a two-step graph representation learning method for gene function analysis. We first integrate multiple molecular networks and then pre-train a GCN to derive initialized embeddings for each gene or protein. Then we re-train the network via GAT model architecture and achieve state-of-the-art performance to recover pathway and disease genes. The integration is simply done through taking the unions of edges derived from different networks after aligning the nodes' identities (see methods). The generalizability of gene embeddings learned from GWAS hits is directly tested by another two independently curated disease gene sets (DisGeNET[37] and UK Biobank[38]) without further model training. Tissue-specific patterns are recapitulated for a broad range of diseases. Reprioritized genes show biologically relevant functional enrichment in related pathways. We also show that attention weights between gene nodes learned from the GAT network offer natural hints on regulatory relationships. Shared gene modules are identified among several common psychiatric disorders, offering functional evidence and recapitulating previous mechanistic insights. In brief, we demonstrate that pre-training GNN on molecular networks in a self-supervised manner provides strategic adaptability to a series of downstream tasks, including pathway gene recovery, disease gene prioritization, module identification, and comorbidity validation. Prioritizing disease-related markers can also benefit from explicitly adding disease nodes. For example, Zhang et al.[39] integrated a microRNA network and disease phenotype network to prioritize disease relevant microRNA. In the comorbidity prediction task, we also demonstrate how to incorporate disease nodes to build a heterogeneous GNN, followed by

adding a decoder function and re-training the network, which achieves superior accuracy.

RESULTS

Overview of Graphene

As shown in Figure 1A, we use four molecular networks to pre-train Graphene, including 142 tissue-specific gene networks from Humanbase, a PPI network from STRING (9606 v11), a recently released systematic proteome-wide reference, namely the Human Reference Interactome (HuRI),[40] and a well-integrated composite network, PCNet. These networks are combined by unifying their edges and nodes (see methods and all network datasets in Table S1) and result in a giant network comprising 19,324 gene nodes and 16,142,804 interconnected edges. We adopt node recovery and context-prediction as two pretext tasks for Graphene pre-training[34] (methods). In particular, we randomly mask 15% of nodes and predict the identifications of masked nodes from transformation of neighborhood representations, defined as a multi-class classification problem through cross-entropy loss. For context prediction, the k-hop neighborhood contains all nodes that are k-hops away from the center node. Nodes shared between the neighborhood and the context graph are referred to as context anchor nodes, providing the connectivity information between the neighborhood and context graphs. Then negative sampling[41] is used to jointly learn both neighborhood and context graph-derived embeddings, casting it as a binary classification problem whether particular context graph and neighborhood belong to the same center node or not. These two auxiliary tasks enable the integration of four molecular networks in a self-supervised manner. We consider GCN and GAT as two pre-training GNN architectures to aggregate neighborhood features. In our experiments of model pre-training, we find that GCN produces more flexible embeddings than GAT, which is beneficial to the downstream re-training process. Embedding size is set as 100. The number of layers for GCN is set as 5. We use one Tesla V100 GPU and draw lessons from the previous report[34] to pre-train Graphene for 100 epochs in around 150 h. The downstream tasks of disease gene reprioritization and gene set member identification can be completed in about 300 s (1,000 epochs) on a Quadro RTX 6000 GPU (Table S2), which is much more efficient than other competing methods.

At the downstream re-training stage, we borrow all pre-trained node embedding as model initialization and adopt two to three GAT layers to drive node embeddings for downstream tasks due to GAT's faster convergence speed (see Table S2) during re-training. These node representations are then fed into one multiple-perceptron classification layer to predict node labels. We use Reactome[42] and NCI[43] as validation datasets for membership recovery task of pathway gene sets (Figure 1B). Only half of the nodes' pathway labels are kept for training, and the remaining members are recovered for each pathway. We use the GWAS Catalog[44] dataset, composed of 202 common diseases, as a training set for the task of disease-gene reprioritization (Figure 1C). It is noted that the re-training process of the disease gene prioritization task is different from the pathway member recovery setting in aspects of train-validation ratio and mask split (methods). Then DisGeNET and UK Biobank (171 aligned disease nomenclatures with GWAS for DisGeNET and 81 diseases for UK Biobank) are used as hold-out test set without further model training for independent cross-dataset evaluation. The re-ranked genes by Graphene can then be used for disease-relevant function module identification and tissue specificity analysis. We also construct a heterogeneous graph via explicitly adding disease nodes to explore the comorbidity relationship between disease pairs, where a decoder function is introduced to predict the edge labels between two disease nodes (Figure 1D). Detailed model architectures for each stage can be found in Figure S1, and illustrations of Graphene implementation can be found in the methods.

Graphene improves member identification for the pathway gene set

Publicly available pathway gene sets related to certain biological processes contain abundant noise due to the inherent nature of high-throughput experiments. We first sought to assess whether re-training Graphene could accurately denoise and recover pathway gene sets. Initialized with pre-trained embeddings, we use a two-layer GAT architecture followed by one classification layer to learn domain-specific representations for Reactome and NCI pathway gene sets. We adopt the same train-test ratio for Set2Gaussian where only half of those membership labels are used in the re-training stage. Evaluated on an NCI dataset using the same metric (mean area under the precision recall curve [mean AUPRC]), Graphene outperforms Set2Gaussian and the simple mean pooling method across all three levels of pathway sets (mean AUPRC = 0.29, 0.31, and 0.29 for small (3–10), medium (11–30), and large (31–1,000), respectively) (Figure 2A). For the purpose of comparison, we also use random initial input embeddings to train Graphene with the same model architecture and obtain inferior performance. Detailed comparison results can be found in Table S3. For the Reactome dataset, Graphene achieves mean an AUPRC of 0.58 and 0.69 for medium (11–30) and large (31–1,000) sets (Figure 2B), outperforming Set2Gaussian. Graphene's GNN architecture effectively propagates information across the graph and facilitates knowledge transfer using a two-step training strategy. This task is run with five repetitions (Figure S5).

Graphene achieves superior performance for disease gene reprioritization with tissue specificity

As potential disease genes converge on interacting molecules in functional networks, we next apply Graphene to GWAS hits to examine how integration of multiple networks and pre-training can benefit decoding gene-disease relationships. We collect association signals for 202 diseases downloaded from the GWAS Catalog and leverage 60% of labels to re-train Graphene on the disease gene recovery task, which is compatible with canonical GWAS workflow. NAGA,[4] which uses RWR as its propagation scheme, together with GenePanda[45] and N2V,[20] are chosen as benchmark methods. NAGA reported stronger performance over other network-based methods, including NetWAS[3] and GWAB.[46] To keep consistent with NAGA, we use the DisGeNET dataset as independent evaluation. In other words, we train, validate on GWAS Catalog disease gene sets, and test on the DisGeNET dataset. DisGeNET is a comprehensive source from expert curations, GWAS catalogs, animal models,

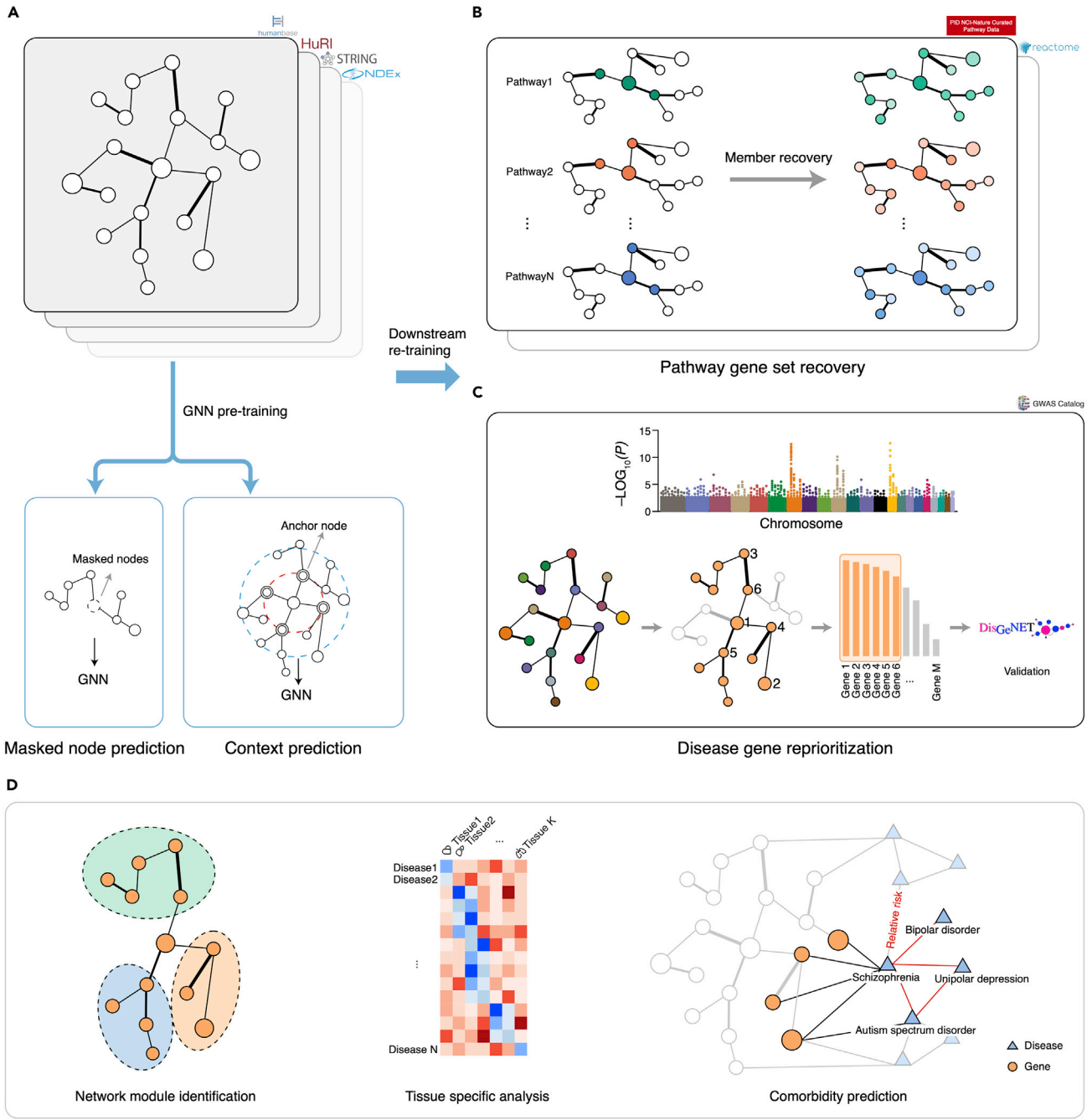


Figure 1. Overview of Graphene workflow

(A) Graphene pre-training includes two pretext tasks, i.e., masked node recovery and context prediction. Four molecular networks are used for self-supervised learning, including HumanBase, STRING (9606 v11), HuRI, and PCNet; nodes stand for genes or proteins, edges stand for the presence of connections between genes in a specific network.

(B) Graphene re-training for gene set member recovery as downstream task. NCI and Reactome are used for pathway gene set recovery.

(C) Graphene re-training for disease gene reprioritization. GWAS signals are used for re-training and DisGeNET dataset is used as independent test set.

(D) Functional analysis include module identification, tissue specificity analysis, and comorbidity prediction. Disease nodes are introduced into Graphene to construct bipartite network. Edges between disease pairs stand for relative risks.

and scientific literature, and developed to support mechanistic studies on human diseases. Like the settings in NAGA, we use the area under the receiver operating characteristic curve (AUROC) as an evaluation metric. Graphene achieves a mean AUROC of 0.76, outperforms NAGA (mean AUROC = 0.71), GenePanda (mean AUROC = 0.59), and N2V (mean

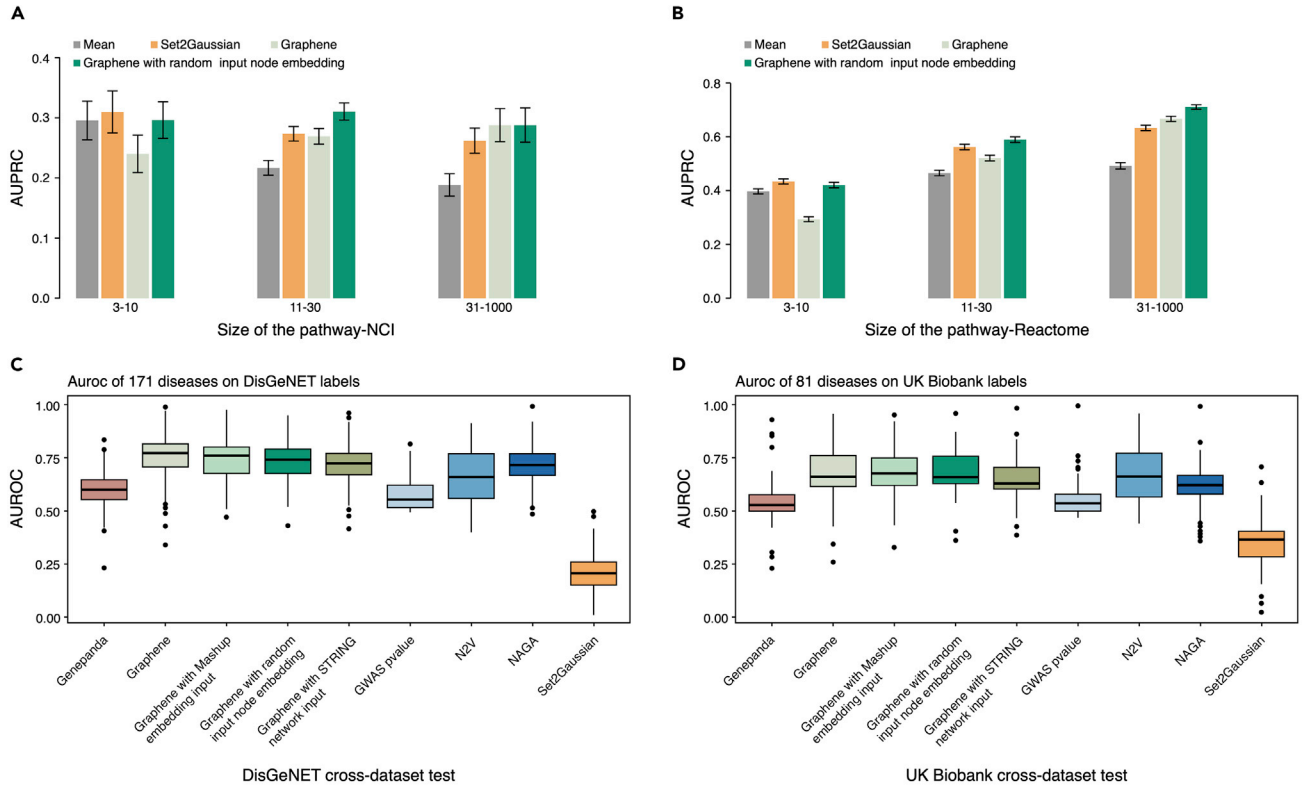


Figure 2. Graphene achieves superior performance in pathway gene set recovery and disease gene reprioritization

(A and B) Application of Graphene downstream re-training for pathway gene set member recovery (NCI [A], Reactome [B]) in comparison with Set2Gaussian, Graphene with random input node embeddings, and mean pooling. Boxplot shows the comparison of area under precision recall curve (AUPRC). Error bars represent the 95% confidence interval.

(C) Comparison of AUROC results on 171 diseases from DisGeNET dataset among nine methods (Graphene, Graphene with Mashup embedding input, Graphene with random input node embedding, Graphene with STRING network input, GWAS p value, NAGA, Set2Gaussian, and GenePanda, N2V).

(D) Comparison of AUROC results on 81 diseases from UK Biobank dataset among nine methods (Graphene, Graphene with Mashup embedding input, Graphene with random input node embedding, Graphene with STRING network input, GWAS p value, NAGA, Set2Gaussian, and GenePanda, N2V). In the boxplot, the center line and box limits denote the median and upper/lower quartiles, respectively. 1.5× interquartile ranges are displayed as whiskers.

AUROC = 0.67) (Figure 2C). Graphene initialized with Mashup embeddings ranked second for the DisGeNET task. Set2Gaussinan (mean AUROC = 0.2) is specifically developed on pathway-level gene sets and its low-dimensional embedding cannot effectively transfer to the disease domain. AUROC results of Graphene for all DisGeNET diseases can be found in Figure S2. In addition, we use UK Biobank summary statistics to check whether the result of GWAS-trained Graphene will generalize to other independent gene-disease association databases. The results show that all four different settings of Graphene exhibit better performance (mean AUROC = 0.68, 0.67, 0.65, and 0.67) than the other four methods, i.e., GWAS (p = 0.55), GenePanda (p = 0.54), NAGA (p = 0.62), and Set2Gaussian (p = 0.34). N2V also achieves a relatively high mean AUROC (p = 0.66), which is comparable with Graphene (Figure 2D). We also show that original GWAS p values cannot compete with a network-based denoising method to recover UK Biobank associations. Notably, NAGA, N2V, and GenePanda can only evaluate one disease at a time, whereas Graphene can test on all diseases in a batch-wise manner. The validation on 202 GWAS diseases is repeated 5 times during downstream re-training, as shown in Figure S5. We also train Graphene on a single network, i.e., STRING (which is the largest of the four individual networks that Graphene has integrated), to illustrate how integrating multiple networks rather than using single input can benefit disease gene prioritization task.[14] DisGeNET and UK Biobank results are shown in Figures 2C and 2D, respectively (Graphene with STRING network input).

We then investigate whether top prioritized genes for a given disease (TPGs) identified by Graphene can reveal tissue specificity in network wiring for relevant diseases. We use expression data from the Genotype-Tissue Expression (GTEx) project[47] and adopt Jensen-Shannon (JS) divergence[48] to measure the tissue specificity of each gene in each tissue. By implementing one-sided Wilcoxon rank-sum test and Bonferroni correction, we test the significance levels of tissue specificity for 300 TPGs against last-ranked 1,000 genes after Graphene reprioritization on GWAS hits. Taking five common diseases as examples (Figure 3A), we show that 300 TPGs of BIP and SCZ have significantly enriched expression level in brain tissues ($p_{adjusted}$ = 2.2×10^{-18} for BIP, 3.4×10^{-27} for SCZ in cortex; $p_{adjusted}$ = 5.4×10^{-10} for BIP, 1.9×10^{-19} for SCZ in the spinal cord) compared with other tissue types. The 300 TPGs of rheumatoid arthritis are observed to exhibit an enriched expression pattern in

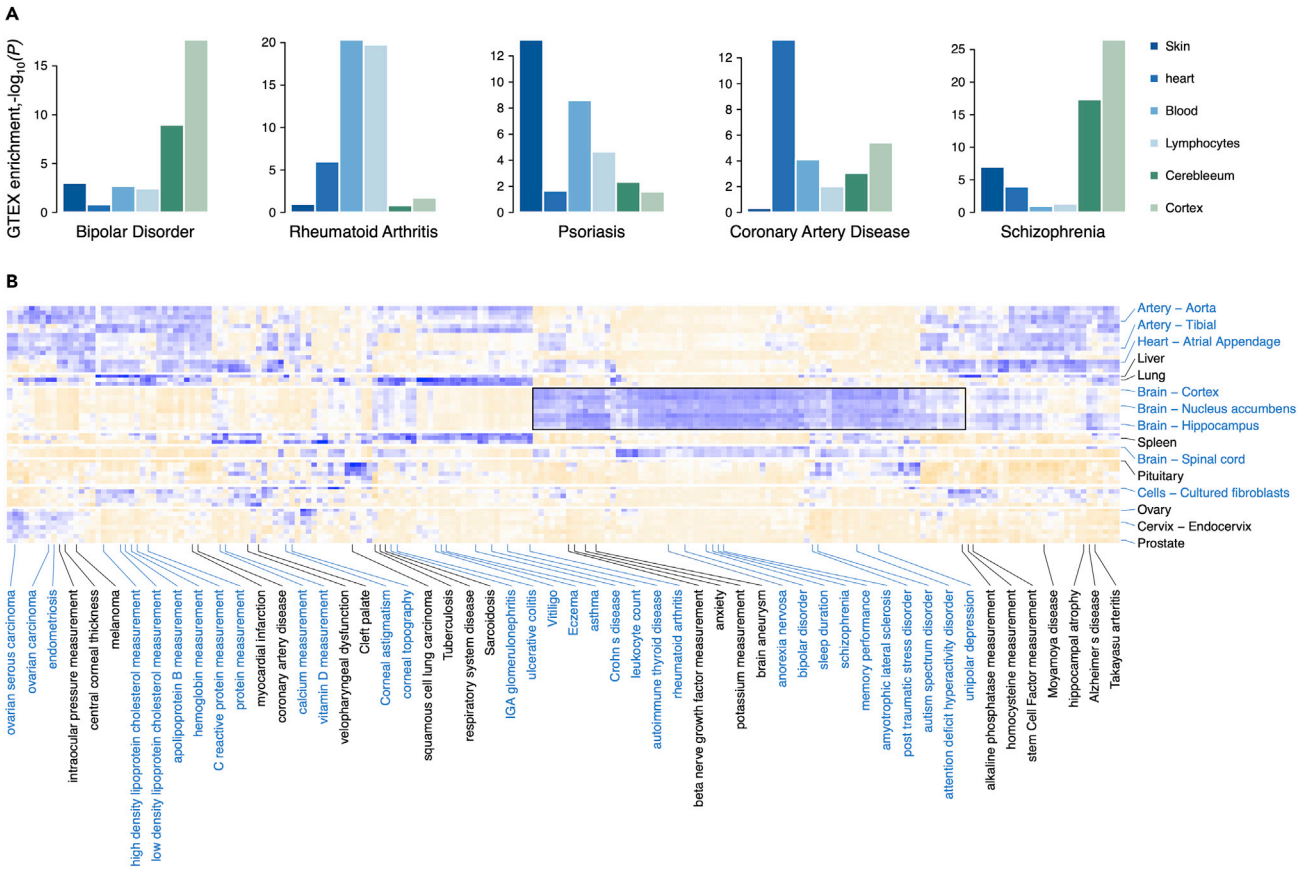


Figure 3. Tissue specificity of top prioritized genes identified by Graphene

(A) Tissue specificity of genes reprioritized by Graphene from GWAS hits. Tissue enrichment scores of five diseases on six different tissue types are plotted for illustration. Details of computing tissue enrichment score are described in methods.

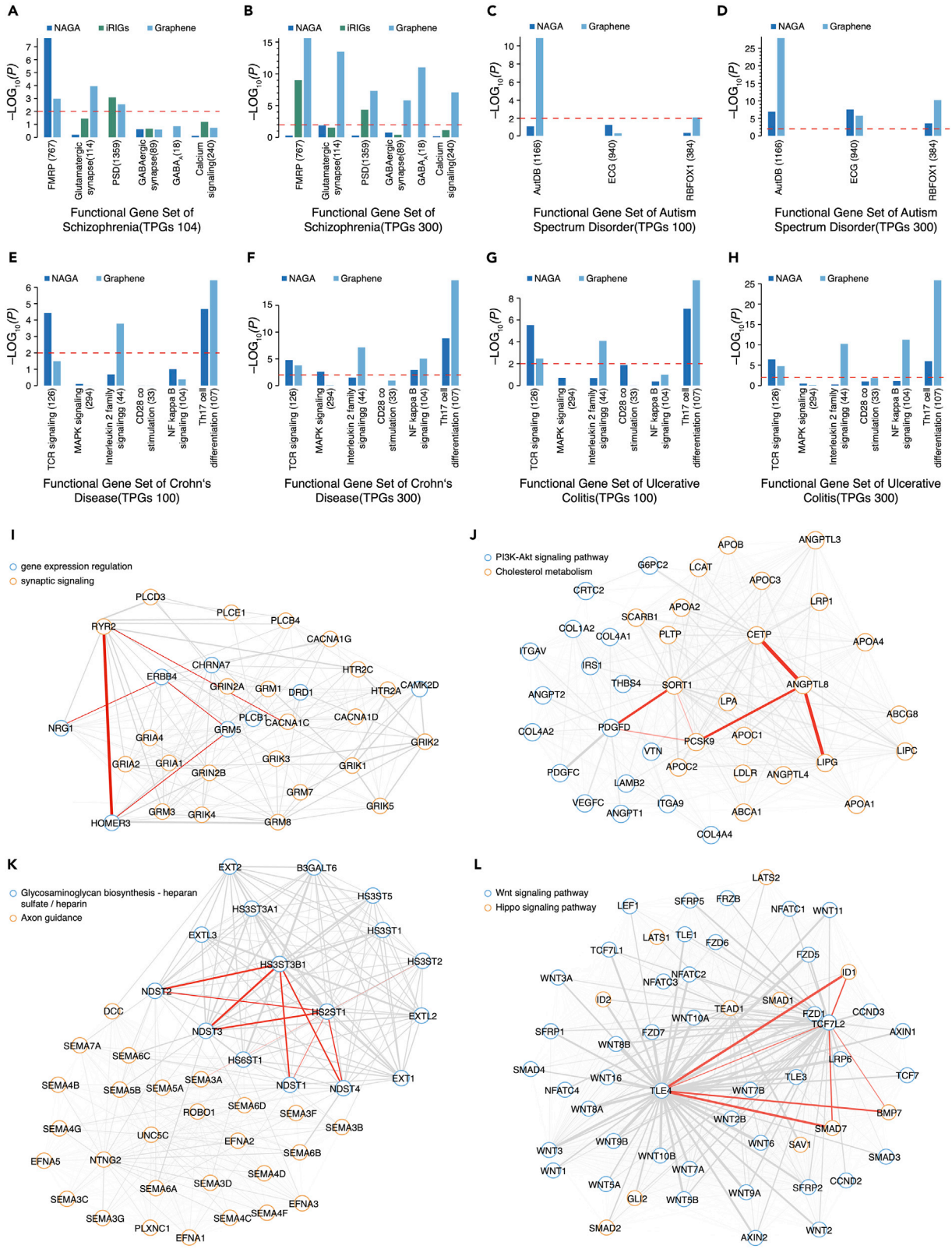
(B) Tissue specificity of genes reprioritized by Graphene on 202 GWAS diseases across 53 tissues in GTEx. Deep blue represents top predicted risk genes highly expressed in corresponding tissues. Wilcoxon rank-sum test is adopted using 300 TPGs and 1,000 last-ranked background genes predicted by Graphene. TPGs, top prioritized genes.

blood ($p_{adjusted}$ = 4.6 × 10^{-21}) and lymphocytes ($p_{adjusted}$ = 1.7 × 10^{-20}) over other unrelated tissue types, such as the cerebellum ($p_{adjusted}$ = 0.15) and skin ($p_{adjusted}$ = 0.1). Also, 300 TPGs show expression enrichment in heart tissue ($p_{adjusted}$ = 3.1 × 10^{-14}) for coronary artery disease and in skin tissue ($p_{adjusted}$ = 3.9 × 10^{-13}) for psoriasis.

The overall heatmap shows clear tissue enrichment differences among various diseases (Figure 3B). Particularly, TPGs of mental diseases are enriched in brain-related tissues. For comparison, original disease-gene mappings from the GWAS Catalog are used as baseline, which present no clear clustering pattern (Figure S3). Although the Humanbase network is incorporated during the Graphene pre-training stage, the tissue information is not explicitly included in the training process. Re-training on GWAS hits can guide the network in recovering tissue specificity. In brief, Graphene effectively denoises the GWAS signals validated by the above observations regarding disease-relevant tissue specificity. In this case, Graphene provides a convenient way to reprioritize GWAS risk genes through injecting molecular network topology derived from graph representation learning.

Graphene effectively characterizes functional enrichment pattern of prioritized disease-associated genes

Dysregulated genes underlying diseases are frequently involved in context-specific biological processes. We further evaluate how TPGs uncover functional modules via gene set enrichment analysis (GSEA). Schizophrenia (SCZ) and autism spectrum disorder (ASD) are both complex mental disorders, representing paradigmatic challenge to illuminate disease biology. iRIGs jointly models multi-omics data for each gene together with their network-based interactions to prioritize GWAS risk loci and assess several gene sets, which have been widely and repeatedly implicated in SCZ. We choose six functional gene sets to evaluate the quality of Graphene TPGs against 104 high-confidence risk genes (HRGs) by iRIGs and NAGA results. Functional gene set includes fragile X mental retardation protein (FMRP) targets[49] (n = 767), postsynaptic density (PSD) proteins[50] (n = 1,359), $GABA_A$ receptor complex,[51] and another 3 KEGG pathways,[52] i.e., calcium signaling pathway[53] (n = 240), glutamatergic synapse[54] (n = 114), and GABAergic synapse (n = 89) (see methods). When using 300 TPGs, Graphene recovers far more



(legend on next page)

significantly enriched signals than an equal number of genes ranked by NAGA and iRIGs HRGs in all 6 gene sets (Figure 4B) ($p_{adjusted}$ = 2.4 × 10^{-16} for FMRP, $p_{adjusted}$ = 4.9 × 10^{-8} for PSD; $p_{adjusted}$ = 9.1 × 10^{-12} for $GABA_A$; $p_{adjusted}$ = 8.4 × 10^{-8} for calcium signaling, $p_{adjusted}$ = 3.3 × 10^{-14} for glutamatergic synapse, and $p_{adjusted}$ = 1.5 × 10^{-6} for GABAergic synapse). Enrichment results using 104 Graphene TPGs outperforms equal numbers of genes prioritized by NAGA in all gene sets and surpasses iRIGs HRGs except for the FMRP gene set (Figure 4A). In addition, we analyzed the ASD scenario in 3 ASD-relevant gene sets (Figure 4D). Using the top-ranked 300 genes, Graphene achieves more significant enrichment than NAGA in the target gene set of RBFOX1 RNA binding protein[55] (n = 384, $p_{adjusted}$ = 2.4 × 10^{-11}) and the gene set from the AutDB database[56] (n = 1,166, $p_{adjusted}$ = 4.9 × 10^{-29}), while exhibiting slightly weaker signals in evolutionarily constrained genes (ECGs)[57] (n = 940, $p_{adjusted}$ = 7.1 × 10^{-7}). However, when we test 100 TPGs, Graphene still shows enrichment signals in AutDB (n = 1,166, $p_{adjusted}$ = 1.2 × 10^{-11}) and RBFOX1 (n = 384, $p_{adjusted}$ = 6.6 × 10^{-3}), while the top 100 genes identified by NAGA fail to reach significance level (Figure 4C). To further validate whether Graphene-derived gene sets can identify enriched biological insights as in other curated knowledgebase or populational studies, we evaluate TPGs of two types of inflammatory bowel disease (IBD) (ulcerative colitis and Crohn disease) identified by Graphene on six previously reported pathways related to immune system signal transduction and T cell activation[42,58] (methods). We demonstrate that 100 TPGs of Graphene signals are significantly enriched ($p < 0.01$) in Th17 cell differentiation pathway and interleukin-2 family signaling pathway, and 300 TPGs of Graphene further recapitulate enriched signals in NF-κB signaling and TCR signaling pathways (Figures 4E–4H).

It is essential to translate GWAS hits to uncover underlying biological mechanisms. EMOGI adapts a layer-wise relevance propagation (LRP) rule[59] to the GCN network to calculate importance scores of PPI partners. Graphene uses the GAT network for downstream functional analysis, so we utilize attention weights to extract important gene-gene interactions under certain disease contexts. For illustration purpose, we check part of 300 SCZ TPGs identified by Graphene that are enriched in glutamatergic synapse and calcium signaling pathways. Two main Gene Ontology (GO) terms, synaptic signaling and gene expression regulation, emerge as key modules. Visualized by the width of the edges scaling with the attention weights for the Graphene model (Figure 4I), we take the following examples to illustrate several important interactions, highlighted in red. RYR2 encodes ryanodine receptor protein of the calcium channel and calcium release is triggered by its activation of the L-type calcium channel CACNA1C.[60,61] Among all RYR proteins widely expressed in the cerebellum and hippocampus (RYR1, RYR2, and RYR3), RYR2 is the most abundant.[61,62] HOMER3 encodes a PSD scaffolding protein that binds and crosslinks to cytoplasmic regions of GRM5, and RYR2[63] assists surface receptors to couple with intracellular calcium release. SCZ GWAS hits on ERBB4 and GRM5 loci were discovered by Greenwood et al.[64] In addition, NRG1 encodes a membrane glycoprotein that mediates cell-cell signaling, and its receptor ERBB4[65] is found to be expressed in GABAergic neurons.[66,67] All prioritized genes connected by attention weights can be found in Figure S4. We show large attention weights representing strong interconnections naturally provide insights about underlying regulatory or interplay mechanisms of complex mental diseases and equip the Graphene model with certain interpretability. To better illustrate the potential utilities of attention weights, we add examples for another three diseases, shown in Figures 4J–4L. Identified TPGs of coronary artery disease are enriched in "cholesterol metabolism" and "PI3K-Akt signaling pathway" (Figure 4J). Sortilin (SORT1) might bind components of the platelet-derived growth factor,[68] whose function can be enhanced by PCSK9.[69] SORT1 is a high-affinity sorting receptor for PCSK9.[70] PathCards[71] and GWAS[72] also show correlations among ANGPTL8, CETP, and LIPG. For hippocampal atrophy's TPGs, two major pathways identified by attention weights are "lycosaminoglycan biosynthesis-heparan sulfate/heparin" and "axon guidance" (Figure 4K). Heparan sulfate is reported to be related to hippocampal atrophy[73] and its synthesis and modification involve NDST1-4, the HS3ST family, and HS2ST1. NDST enzymes may affect the potential functional relation between NDSTs and HS2ST1.[74,75] In addition, HS3ST and NDST have very similar sulfotransferase domain.[76] Studies show that semaphorin-3a (Sema3a)-induced axonal growth cone collapse depends on HS3ST, indicating that activities of the semaphorin family rely on HS modifications.[77–79] TPGs of alopecia mainly cluster into two pathways termed the "Wnt signaling pathway" and the "Hippo signaling pathway" (Figure 4L). A previous study indicated that Wnt/β-catenin and Hippo signaling pathways played important roles in hair follicle regeneration[80] and development of alopecia.[81] TLE4 is involved in the negative regulation of the canonical Wnt signaling pathway. It can suppress Smad7 and activate the expression of bone morphogenetic protein (BMP) signaling, and enhance and sustain the upregulation of the endogenous ID1 gene induced by BMP7.[82] Through interacting with TCF7L2, the co-repressor TLEs repress transactivation.[83] The TCF/LEF family interacts with Smad families to coordinate the transcription of target genes,[84] while it may repress BMP/SMAD signaling with elevated expression of BMP signaling targets, such as Id1, Id2, and Id3.[85]

Figure 4. Graphene identifies functional enrichment pattern of mental disorders and discovers relevant molecular interactions

(A and B) Enrichment of 104 TPGs and 300 TPGs identified by Graphene in six schizophrenia (SCZ)-related functional gene sets in comparison with equal number of genes ranked by iRIGs and NAGA.

(C and D) Enrichment of 100 TPGs and 300 TPGs identified by Graphene in three autism spectrum disorder (ASD) relevant functional gene sets in comparison with equal numbers of genes ranked by NAGA.

(E–H) Enrichment of 100 TPGs and 300 TPGs identified by Graphene on six IBD (ulcerative colitis, Crohn disease) relevant gene sets in comparison with equal numbers of genes ranked by NAGA.

(I–L) Attention weights extracted from 300 Graphene TPGs of four different diseases exhibit important molecular interactions within functional pathways. (I) SCZ, (J) coronary artery disease, (K) hippocampal atrophy, (L) alopecia.

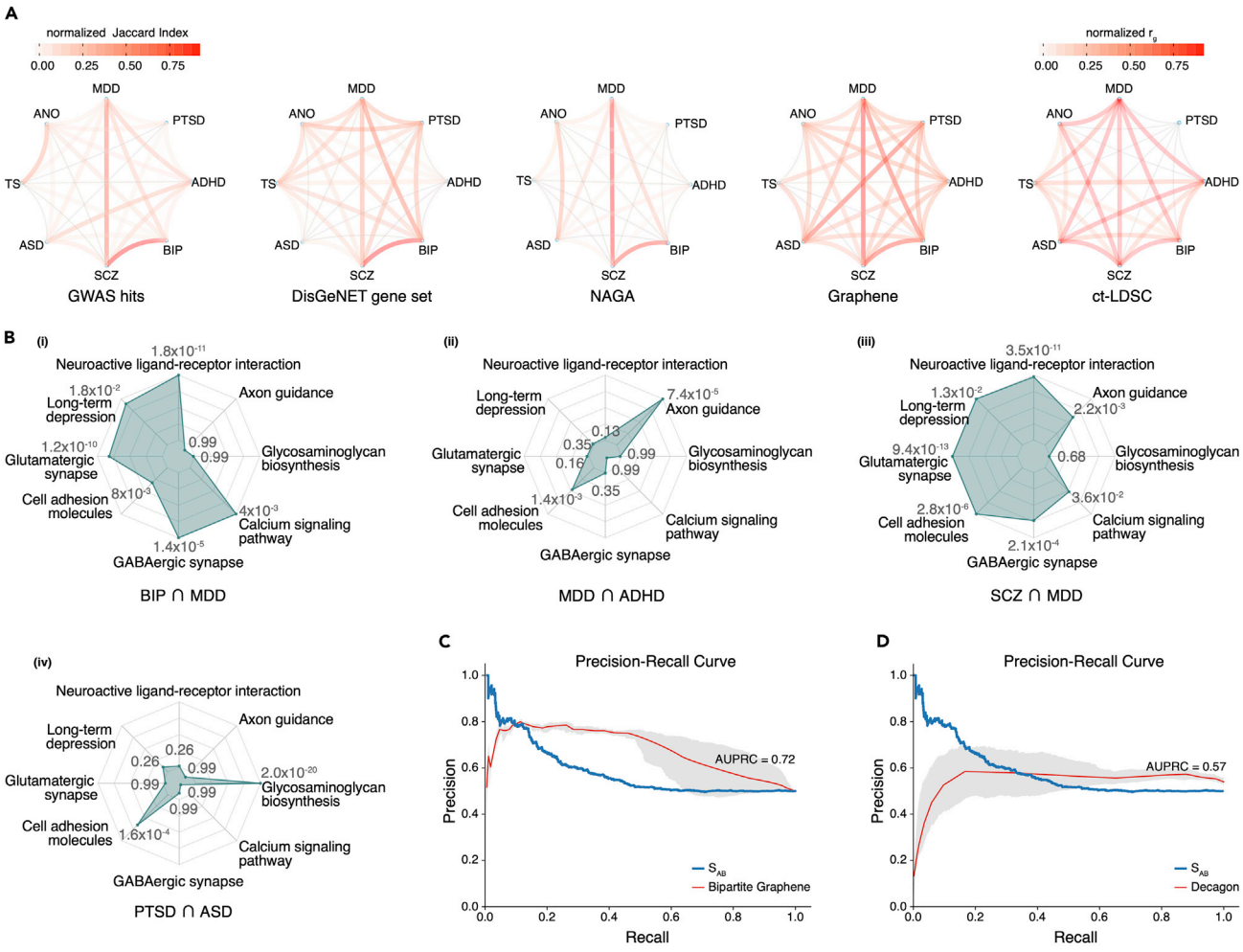


Figure 5. Graphene identifies strongly shared heritability of mental illnesses and boosts performance for comorbidity prediction

(A) Comparison plot of genetic correlations among eight mental diseases identified by original GWAS hits, DisGeNET gene sets, ct-LDSC correlation score, NAGA, and Graphene. The gradational color between disease pairs represents normalized Jaccard Index except ct-LDSC. Eight mental diseases include unipolar depression (MDD), post-traumatic stress disorder (PTSD), attention deficit hyperactivity disorder (ADHD), bipolar disorder (BIP), schizophrenia (SCZ), autism spectrum disorder (ASD), Tourette syndrome (TS), and anorexia nervosa (ANO).

(B) Functional enrichment score of overlapping Graphene TPGs for four disease pairs (BIP and MDD, MDD and ADHD, SCZ and MDD, and PTSD and ASD). Eight KEGG pathways associated with mental disorders were used for evaluation.

(C) Precision-recall (PR) curve of Bipartite Graphene for comorbidity prediction in comparison with original disease separation score (s_{AB}).

(D) PR curve of Decagon for comorbidity prediction in comparison with disease separation score (s_{AB}). Ten-fold cross validation is implemented for (C) and (D).

Graphene discovers both shared heritability and distinct genetic underpinnings of multiple psychiatric disorders

Mental disorders usually share similar symptoms with epidemiological comorbidity, posing difficulties for diagnosis and treatments.[86] Illuminating genetic underpinnings can provide evidence about intercorrelated psychopathology and raise the need to refine current clinical psychiatric diagnostics. We investigate whether Graphene TPGs of eight common mental diseases can reveal their genetic intercorrelations. Following CC-GWAS[87] definition, we measure the pairing correlation of every two diseases via computing the normalized Jaccard Index of their prioritized gene set (100–500 min-max normalization for each disease). We also leverage similar methods to compare the correlation results obtained from original GWAS hits, DisGeNET, and NAGA. The values of normalized genetic correlation (defined as the r_g value) computed by cross-trait LD score regression (ct-LDSC)[88] are directly retrieved from the CC-GWAS paper.[87] We compare different correlation patterns derived from all these strategies (Figure 5A). Overall, ct-LDSC and Graphene exhibit stronger intercorrelations among these mental disorders compared with original GWAS hits, NAGA, and DisGeNET. Considering two closely related depressive disorders as an example, i.e., unipolar depression (MDD) and BIP, their GWAS hits correlation (0.32) is much lower than ct-LDSC (0.5), DisGeNET (0.76), and Graphene (0.94), again demonstrating the importance to refine GWAS signals. We extract the overlapping Graphene TPGs for BIP and MDD (overlapping genes include KCND2, RIMS1, KCNA4, and RGS8) and implement GSEA on eight mental illness-relevant KEGG

pathways (Figure 5B-i). We observe that these genes have functional enrichment in neuroactive ligand-receptor interaction ($p_{adjusted} = 1.8 \times 10^{-11}$), glutamatergic synapse ($p_{adjusted} = 1.2 \times 10^{-10}$), and GABAergic synapse ($p_{adjusted} = 1.4 \times 10^{-5}$). Moreover, both Graphene (0.54) and ct-LDSC (0.49) report higher correlation between ASD and MDD against GWAS (0.15), NAGA (0.26), and DisGeNET (0.06). A similar trend is also observed between anorexia nervosa (ANO) and SCZ, where GWAS (0.05), DisGeNET (0.2), and NAGA (0.05) show relatively lower correlation than ct-LDSC (0.37) and Graphene (0.31). SCZ and MDD are identified as strongly correlated disease pairs among all approaches, their overlapping Graphene TPGs (including CTNNA3, HLA-G, CSRNP3, GRIN2B, and RELN) manifest functional enrichment in seven KEGG pathways (Figure 5B-iii), including neuroactive ligand-receptor interaction ($p_{adjusted} = 3.5 \times 10^{-11}$), glutamatergic synapse ($p_{adjusted} = 9.4 \times 10^{-13}$), GABAergic synapse ($p_{adjusted} = 2.1 \times 10^{-4}$), calcium signaling ($p_{adjusted} = 3.6 \times 10^{-2}$), axon guidance ($p_{adjusted} = 2.2 \times 10^{-3}$), cell adhesion molecules ($p_{adjusted} = 2.8 \times 10^{-6}$), and long-term depression ($p_{adjusted} = 1.3 \times 10^{-2}$). For MDD and ADHD, their overlapping Graphene TPGs (including NALCN, NRXN3, NRG3, and LRP1B) are enriched in axon guidance ($p_{adjusted} = 7.4 \times 10^{-5}$) and cell adhesion molecules ($p_{adjusted} = 1.4 \times 10^{-3}$) (Figure 5B-ii). Another interesting discovery from Graphene is the relatively stronger correlation between post-traumatic stress disorder (PTSD) and ASD, and their overlapping Graphene TPGs (including CBLN4, BRINP1, and GLCE) are enriched in glycosaminoglycan biosynthesis ($p_{adjusted} = 2.0 \times 10^{-20}$) and cell adhesion molecules ($p_{adjusted} = 1.6 \times 10^{-4}$) (Figure 5B-iv). Brinp1 has been reported to be associated with both ASD[89] and PTSD.[90] Several cognitive and behavioral mechanisms might be shared between PTSD and ASD, such as increased rumination, cognitive rigidity, avoidance, anger, and aggression. Understanding the shared genetics can help explore the common mechanisms underlying paired mental disorders. Correlation values of each disease pair extracted by the above five methods are listed in Tables S4–S8. Considering ct-LDSC derived scores as gold standard, we also calculate the Spearman correlation coefficients of all paired similarities extracted by Graphene, NAGA, GWAS, and DisGeNet with ct-LDSC (PTSD is not included in ct-LDSC), and the result (Table S9) shows that Graphene and NAGA denoise the underlying signals and achieve more similarly shared genetics with ct-LDSC than GWAS and DisGeNet.

As opposed to the shared genetic correlation, we also investigate whether the genetic differences between two diseases can reveal their distinct parthenogenesis mechanisms. CC-GWAS[87] leverages allele frequency differences to identify differential genetic components between cases of two disorders. We also check the non-overlapping TPGs between two diseases identified by Graphene. For ANO versus Tourette syndrome (TS) and SCZ against TS, POU3F2[91,92] encodes neural transcription factors involved in neuronal differentiation. For SCZ against MDD, KCNV1[93] encodes a member of the potassium voltage-gated channel subfamily V as an essential function in the brain. For SCZ against TS, NFIB[94] is a transcriptional activator, essential for neuron axon genesis and other CNS. All overlapping and non-overlapping genes between the aforementioned mental disease pairs can be found in Tables S10 and S11, respectively.

Leveraging heterogeneous graph to re-train Graphene enables comorbidity prediction of disease pairs

All the above disease-gene association analyses are based on homogeneous GNNs, where only the gene node presents, and diseases are used as node attributes or labels. Constructing a multimodal graph where two or more types of nodes exist, more diverse inter-node relationships can be modeled. Decagon builds a bipartite graph to represent protein-drug interactions and model polypharmacy side effects as edges between paired drug nodes. Inspired by Decagon, we further introduce disease node into Graphene to model comorbidity relationship as links between disease nodes. Disease-associated genes or proteins interacting with each other tend to cluster into neighborhood structure as disease modules. If two diseases partially share overlapping modules, the local perturbation of functional pathways of one disease can lead to similar disruption in another, displaying as shared clinical and pathobiological features. Menche et al.[95] integrated disease-gene annotations from Online Mendelian Inheritance in Man (OMIM)[96] and GWAS data from the Phenotype-Genotype Integrator database (PheGenI),[97] obtaining 299 diseases and 3,173 associated genes, and used 30 million individuals aged 65 and older to determine relative risk (RR) for each disease pair as comorbidity metric. They also developed a network-based separation measurement of a disease pair defined as s_{AB} by comparing the shortest distances between proteins within each disease based on their constructed interactome, which is a network of 13,460 protein nodes and 141,296 links. They found that s_{AB} can be used as a metric to discriminate the degree of RR between disease pair (RR $\geq$ 10 for $s_{AB} < 0$ versus random expectation of RR $\approx$ 1 for $s_{AB} > 0$, see methods). Akram et al.[98] developed a weighted geometric embedding algorithm on this dataset and predicted comorbidity with performance of AUROC = 0.76 at threshold RR = 1 in a supervised manner. To test the decoder utility of bipartite Graphene to predict RR, we reconstruct Graphene through adding the same 299 disease nodes, re-training on the same gene-disease associations data, and training Graphene on paired disease RR values as edge labels in 10-fold cross validation setting. Similar training procedures are implemented for Decagon architecture. As shown in Figure 5C, Bipartite Graphene achieves a mean AUPRC of 0.72 and a mean AUROC of 0.79 (training for 20 epochs), significantly surpassing Decagon (mean AUPRC = 0.57, mean AUROC = 0.67 for 30 epochs of training, Figure 5D). Graphene's GAT decoder and pre-training setting show stronger performance to predict disease separation than Decagon's end-to-end supervised training with GCN decoder.

DISCUSSION

We present Graphene, an integrative GNN framework to decode gene function under network-defined context.

Graphene integrates multiple interactome networks from heterogeneous sources via a graph SSL approach. Then the informative gene embeddings are used as model initialization to infer functional properties of genes or proteins. We successfully demonstrate the wide applicability of Graphene in pathway gene recovery, disease-gene reprioritization, module identification, and comorbidity prediction. Several benchmark experiments have been performed to validate substantial improvements of Graphene over previous methods in each application.

The parameters sharing the scheme of pre-training GCN allow Graphene to encode both node attributes and its diverse neighborhood or context, leading to stronger expressive power over traditional network diffusion-based methods. During the re-training stage, GAT architecture guides Graphene to search task-specific connectivity patterns across the network and reprioritize all genes with fast convergence speed. We have shown that the emerging pre-training and re-training paradigm in deep learning community can be applied to complex biological networks and effectively transfer knowledge to downstream functional analysis. In this paper, we only implement node-level pre-training, and we plan to incorporate a graph-level pre-training task as a supplement to further capture global-level representations.

We also showcase that Graphene can re-rank GWAS hits and validate superior disease gene recovery performance on an independent hold-out DisGeNET dataset. Population-wide GWAS have identified a large number of disease-associated loci with genome-wide significance, although only contributing small amount of the heritability. There is an ongoing debate whether GWAS hits can reveal disease etiology and imply therapeutic targets; in particular, most signals do not match with core genes. The ''Omnigenic'' model[1] has been raised to explain those genomic regions that fell below statistical significance for association increase disease susceptibility through cumulative weak effects in relevant tissues. These weak effects are broadly distributed across network modules and function together in certain biological processes, pathways, and more complex networks. Indeed, disease genes are not scattered randomly but organized into disease-specific modules. Therefore, molecular networks can serve as functional map to refine GWAS hits, re-rank risk genes and guide the discovery of additional candidate genes. Developing a powerful network-based method based on large-scale, cross-tissue interactome datasets is essential to understand pathophysiological processes. Although we only use generic networks as integration inputs where tissue or context labels are not explicitly incorporated into Graphene during both pre-training and downstream re-training process, we recapitulate tissue specificity of reprioritized GWAS signals based on the GTEx dataset. In the future, we expect explicitly incorporating multi-view labels during network integration of GNN pre-training can equip the model with tissue awareness and further boost learning effectiveness. As an ever greater number of biological interactome are mapped, the Graphene framework presented here is easily expandable by adding newly discovered networks into the GNN model and is thereby adaptable to various functional analysis.

We showcase TPGs identified by Graphene revealing stronger functional enrichment in SCZ- and ASD-relevant pathways over previous methods. We also demonstrate certain model interpretability by extracting significant gene-gene attention weights from the GAT network to pinpoint important gene-wise interaction partners. Moreover, Graphene provides genetic underpinnings of shared heritability among eight common mental disorders by investigating their overlapping TPGs. The non-overlapping TPGs also offer some hints regarding distinct pathogenesis mechanisms between disease pairs. By adding disease nodes into Graphene to build a heterogeneous bipartite network, Graphene achieves excellent performance for comorbidity prediction via link prediction. Due to the fact that 299 diseases used for evaluation are far fewer than the number of genes, learning effective disease-disease edge embedding is non-trivial. Our GAT decoder outperforms Decagon's GCN decoder, again demonstrating the importance of GNN architecture choices at different stages.

In the absence of a gold standard disease gene set, Graphene serves as a ready-to-use tool to refine any novel GWAS findings and retrieve candidate genes for detailed follow-up investigation. Since GWAS are based on population-level genotype-phenotype information, which is different from those networks used as input to Graphene, we foresee our tool can offer orthogonal evidence to discover biologically relevant modules and elucidate underlying disease mechanisms. Based on the robustness for gene prioritization, Graphene can also be extended to develop target gene panels for diagnosis of inherited disease or risk evaluation panel for complex traits. In addition, for a cohort where individual-level omics data are available, Graphene can concatenate variant information and other multi-omics features together with pre-trained gene embeddings, as in EMOGI,[29] and enable patient-level disease classification during the downstream re-training stage, thus providing a potential analysis tool for applications in precision medicine. Considering recent progress in applying graph SSL for information retrieval and recommendation system, we plan to further explore causal inference-based learned GNN to interpret large biological networks in the future.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Meng Yang, yangmeng1@mgi-tech.com.

Materials availability

This study did not generate new unique reagents.

Data and code availability

All datasets in this study were published previously, and their availabilities are described in Table S1. Graphene is written in Python using the Pytorch library. The source code has been deposited at Zenodo under the https://doi.org/10.5281/zenodo.7233857.

Methods

In this work, we first design two auxiliary tasks to pre-train GNN to integrate four molecular networks and re-train the network for downstream investigations, including pathway gene set recovery, disease gene reprioritization, and other functional studies. In the following sections, we describe each of the proposed components in details.

Pre-training the GNN

Sources of pre-training molecular networks. We combine four different sources of networks freely accessible to build a single network for pre-training. We assign the presence of edge connection between two nodes as long as there exists interaction in any single network. HumanBase, as a tissue-specific gene network, is built on a collection of datasets covering thousands of experiments from 14,000 distinct publications. Incorporating HumanBase might help inject tissue specificity into our combined network and we download 142 gold standard tissue networks from Humanbase (https://hb.flatironinstitute.org/download). The tissue label is not explicitly included. We download STRING9606 v11 (https://string-db.org), which contains experimentally derived protein-protein interactions through literature curation, scientific text mining, calculation from genomic features, and other model organisms. We also collected 52,548 connections from the Human Reference Interactome (HuRI) (http://www.interactome-atlas.org/download), which is a systematic proteome-wide reference that links genomic variation to phenotypic outcomes. In addition, PCnet itself is a composite network that can boost performance and serves as supplementary to the other three networks, and 2,610,605 connections were downloaded from the Network Data Exchange (NDEx) database (http://www.ndexbio.org). We convert each network to a set of tuples, and each tuple consists of two nodes, interconnected by an edge between them. The node ID of each node is Entrez ID. We then take the union of four sets to generate a unified network of 19,324 gene nodes and 16,142,804 edges. The edges are equally weighted.

Model structure for pre-training. A schematic diagram of model architectures can be found in Figure S1 and we illustrate in formula form below. Our graph was denoted by $G = \{V, E\}$ with N nodes $v_i \in V$, edges $(v_i, v_j) \in E$, a binary adjacency matrix $A \in A^{N\times N}$. We randomly initialized node feature vector matrix X_{vi} for $v_i \in V$ as the input to GNN:

$$X_{v_i} = Embedding(V, embedsize), \quad \text{(Equation 1)}$$

where $X_{vi} \in R^{1\times De}$, where D_e represents the embedding size. Node representations were updated at each layer by:

$$H^{(l+1)} = \sigma\left(\tilde{D} - \frac{1}{2}\tilde{A}\tilde{D} - \frac{1}{2}H^{(l)}W^{(l)}\right), \quad \text{(Equation 2)}$$

where $\tilde{A} = A + I_N$ is the adjacency matrix of the graph G, I_N is the identity matrix, D is a trainable weight matrix. The equation adopts ReLU activation $(\sigma(\cdot))$ with a certain number of hidden units. We devised two pre-training auxiliary tasks, context prediction and masked node recovery, as follows.

Context prediction. We performed this task by negatively sampling neighborhood and context representations. The above node update scheme provided us with neighborhood representation $h_{v_i}^k$ of center node v_i. Furthermore, we defined context representation $c_{v_i}^G$ by calculating the mean sum of representations of anchor nodes $v_j \in A_{\text{anchor}}$ that are k hops adjacent to the center node:

$$c_{v_i}^G = MEAN\left(\sum_{v_j \in A_{anchor}} h_{v_j}^k\right). \quad \text{(Equation 3)}$$

With these two representations, the learning objective of Context Prediction was a binary classification of whether a particular neighborhood $h_{v_i}^k$ of v_i and a particular context $c_{v_j}^G$ of v_j belong to the same node:

$$y' = \sigma\left(h_{v_i}^{(k)T} c_{v_j}^G\right) = \begin{cases} 0 & (i \neq j) \\ 1 & (i = j) \end{cases}, \quad \text{(Equation 4)}$$

where $\sigma(\cdot)$ is a sigmoid function. During training, we chose either a positive pair of $h_{v_i}^k$ and $c_{v_j}^G$ $(i = j)$ or a random negative pair $(i \neq j)$ with positive/negative sampling ratio 1:1, and we used binary cross entropy loss:

$$\mathcal{L}_c = -y \log(y') - (1 - y)\log(1 - y'). \quad \text{(Equation 5)}$$

Node prediction. We cast the masked node recovery as a classification task. We masked the node and let the pre-train model predict those nodes. First, we masked a node in the graph by replacing its node embedding with a mask embedding. Second, we applied a pre-training graph model to obtain a corresponding node hidden state $h_{v_i}^{(k)}$, which is consistent with Equation 2. Finally, we applied *FC* (fully connected layer) on $h_{v_i}^{(k)}$ to predict the node:

$$p_{v_i}^{node} = softmax\left(W^{node} \cdot h_{v_i}^{(k)} + b^{node}\right). \quad \text{(Equation 6)}$$

$p_{v_i}^{node} \in R^N$ is a vector that represents the probability of each node. W^{node} is weight matrix and b^{node} donates the bias matrix. We use cross entropy loss to optimize the entire pre-train model:

$$\mathcal{L}_{NodeMask-v_i} = -\log\left(p_{v_i}^{node}[i]\right). \quad \text{(Equation 7)}$$

As the ground truth label is a one-hot vector, the cross entropy loss can be simplified to the above format. $p_{v_i}^{node}[i]$ indicates the *i-th* item in vector p_{vi}.

Re-training for pathway gene set recovery

Sources of pathway gene sets. Two widely used public gene sets were considered in this task, i.e., the National Cancer Institute Pathway Interaction Database (NCI) and The Reactome Knowledgebase (Reactome). We downloaded all 211 NCI pathways from NDEx (http://www.ndexbio.org) composed of human molecular signaling, regulatory events, and key cellular processes. Reactome is a free and open source database of biological pathways in intermediary metabolism, signaling, innate and adapted immunity, transcriptional regulation, apoptosis, and various diseases. We downloaded the Reactome Pathways Gene Set file, which contained 2,408 sets (https://reactome.org/download-data). We removed those pathways containing fewer than 3 genes and finally obtained a Reactome label file of 2,035 pathways.

Sources of disease gene sets. Disease-gene associations for re-training are downloaded from GWAS Catalog v1.0.2 (https://www.ebi.ac.uk/gwas/docs/file-downloads), which is a publicly available resource of GWAS. We obtained 3,954 diseases grouped by mapped traits with gene $p < 5 \times 10^{-5}$. To unify the nomenclature with downstream DisGeNET datasets, we chose diseases/traits that have identical names in DisGeNET (https://www.disgenet.org/downloads) and deleted those traits/diseases with associated genes less than 30 and we finally obtained 202 traits/GWAS disease. Most of these 202 chosen diseases were among the most common disorders cataloged in both GWAS and DisGeNET, and 171 of them have curated gene lists in DisGeNET (for detailed IDs, see Figure S2). A total of 171 DisGeNET gene sets was used as a hold-out test set for disease gene reprioritization.

Model structure for gene set member recovery. For each database above, suppose we had M gene sets and each set corresponded to N human genes. We arranged these datasets into a target matrix $S = \{s_{ij}\}_{M\times N}$, where s_{ij} is a binary value indicating whether the gene v_j was the member of *i-th* gene set. Our aim was to predict the presence possibility of gene v_j in a given gene set m_i. The proposed downstream re-training of the GNN model consists of three modules: the embedding layer, the GAT layers, and the classification layer. Input gene embeddings were extracted from the above pre-trained network.

The pre-trained node embeddings can be represented as $H = \{h_1, h_2, \ldots, h_N\}$, where $h_i \in R^K$ and K represent embedding size. The embedding layer accepts graph node embeddings as initializations. Each node embedding is represented as a K-dimensional vector and the weights are initialized by our pre-trained node embeddings, i.e., the *i-th* node embedding is h_i. Then we map these node embeddings into F-dimensional vectors through a fully connected layer:

$$H^{emb} = f\left(W^{emb} \cdot H + b^{emb}\right) = \left\{h_1^{emb}, h_2^{emb}, \cdots, h_N^{emb}\right\}. \quad \text{(Equation 8)}$$

Here, $h_i^{emb} \in R^F$ is the output embedding, $W^{emb} \in R^{F\times K}$ is weight matrix, $b^{emb} \in R^F$ is the bias vector, and $H^{emb} \in R^{N\times F}$ represents the output of embedding layer.

Then, the GAT layers take the output from the embedding layer H^{emb} as input and aggregate the node information through a graph structure. We use the following formula to obtain the edge weight α_{ij} between nodes v_i and v_j:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\alpha^\top \left[W^{GAT} h_i^{emb} \parallel W^{GAT} h_j^{emb}\right]\right)\right)}{\sum_{l\in O_i} \exp\left(\text{LeakyReLU}\left(\alpha^\top \left[W^{GAT} h_i^{emb} \parallel W^{GAT} h_l^{emb}\right]\right)\right)}. \quad \text{(Equation 9)}$$

$W^{GAT} \in R^{F'\times F}$ is a weight matrix applied to every node transforming the dimensionality from F to F', $a \in R^{2F'}$ is a learnable vector. We applied *LeakyReLU* as the activation function. O_i is the set of the neighboring nodes

close to v_i in graph G. With weight α_{ij}, we can obtain the final output feature of every node produced by the GAT layer:

$$h_i^{GAT} = \prod_{t=1}^{T} \sigma\left(\sum_{j \in O_i} \alpha_{ij}^t W_t h_j^{emb}\right). \quad \text{(Equation 10)}$$

We employed the multi-head attention mechanism, where T is the number of heads and $\prod$ represents concatenation. W^t, $t = 1, \cdots, T$ is a weight matrix and σ is a nonlinearity activation function. $h_i^{GAT} \in R^{TF'}$ is the produced vector for node v_i. The output of the GAT layer is $H^{GAT} = \{h_1^{GAT}, h_2^{GAT}, \ldots, h_N^{GAT}\}$.

Then the classification layer takes H^{GAT} as input and derives final classification. This layer applies average pooling to H^{GAT} over all heads and then uses the sigmoid function for classification:

$$h_i^{out} = sigmoid\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{j \in O_i} \alpha_{ij}^t W_t^{out} h_j^{GAT}\right), \quad \text{(Equation 11)}$$

where $W^{out}{}_t \in R^{M \times TF'}$, $t = 1, \cdots, T$, and M is the number of gene sets. $h_i^{out} \in R^M$ is the output probability vector of node v_i. The output of the classification layer can be represented as a matrix $H^{out} = [h_1^{out}, h_2^{out}, \ldots, h_N^{out}] \in R^{N \times M}$. Each element H_{ij}^{out} in this matrix means the probability of the gene v_i is the member of gene set j. Then we can use the binary cross entropy loss to optimize the full network:

$$\mathcal{L}_{GenSet} = \sum_{i=1}^{N}\sum_{j=1}^{M} -s_{ij} \log H_{ij}^{out} - (1 - s_{ij}) \log\left(1 - H_{ij}^{out}\right). \quad \text{(Equation 12)}$$

During the re-training stage, we randomly masked the labels of half of all nodes and used the other half as a training set to enforce the model to predict the probabilities of all genes.

We used the same model architecture as described above for gene set member recovery. The embedding layer also takes pre-trained node embeddings H as input, and the GAT layer employs G_{gg} as the graph. The output of the classification layer represents the importance probability of the gene v_j to disease d_i. We used the binary cross entropy as loss function to train the model, given the ground-truth label matrix $D = \{d_{ij}\}^{Q \times N}$.

Disease comorbidity prediction

Source of disease-disease comorbidity and disease-gene associations. For this task we adopted RR (from 0 to <9,000) of disease-disease comorbidity for each pair of diseases that were determined using the disease history records of 30 million individuals aged 65 years or older (U.S. Medicare). There were 6,269 disease pairs with comorbidity value RR $\geq$ 1 as positive pair and the rest were negative. For convenience of comparison, we used the disease-gene associations through integrating OMIM (www.ncbi.nlm.nih.gov/omim) and GWAS (www.ncbi.nlm.nih.gov/gap/PheGenI), using a p value cutoff of 5×10^{-8}.

Bipartite model structure for disease comorbidity prediction. We constructed Bipartite Graphene by replacing GCN layer of Decagon model's decoder with a GAT layer. In addition to gene-gene graph G and disease-gene association matrix D, a disease-disease relationship matrix C was required. Disease-disease relationships were calculated by the Jaccard Index between those 299 diseases chosen above. Then, we learned hidden states of each node from their neighborhood consisting of heterogeneous node types. Finally, we made predictions between two nodes via an edge decoding function. Then comorbidity can be considered as links between two disease nodes. We trained a model to learn the relationships between disease pairs and then predicted those test links in 10-fold cross validation. Formally, the Bipartite Graphene model takes the following form:

$$h_{i,x}^{BGAT} = \sum_{l}\sum_{j \in O_i^l} u_{i,j} W_l^x z_{j,x} + W_{b_{i,s}}^x z_{i,x}, z_{i,x+1} = \varnothing(h_{i,x}). \quad \text{(Equation 13)}$$

$z_{i,x} \in R^{ux}$ represents the hidden state of node v_i in the x-th GAT layer. $h_{i,x}^{BGAT}$ is the feature vector that aggregates information from v_i's neighborhoods, l is the type of node links, and O_i^l is the neighborhood set of node v_i with regard to type l. W_l^x and $W_{b_{i,s}}^x$, are the weight matrices at layer x, and b_j is the type of the node. $u_{i,j}$ is a normalization constant, which can be formulated as $u_{i,j} = 1/\sqrt{|Q_i| \cdot |Q_j|}$. φ indicates the activation function $ReLU$.

Since we have different types of nodes and links, the computation of the graph propagation can vary according to different types of the neighborhood. We used GAT architecture to aggregate and propagate node representation $z_{i,x}$ into the node representation $z_{i,x+1}$ for the next layer. The final representation of node v_i is $z_{i,x}$, where x is the number of GAT layers. For the edge decoding model, the probability of a link between disease j and disease i can be described as:

$$P(d_i, d_j) = \sigma\left(z_{d_i,x}^{\top} W_c z_{d_j,x}\right). \quad \text{(Equation 14)}$$

$z_{di,x}$ is disease node representation for d_i, $z_{dj,x}$ is disease node representation for d_j. W_c is the weight matrix to capture the relationships between disease pairs. σ is the sigmoid function, so $P(d_i, d_j)$ will be a real value within range (0, 1) indicating the co-occurrence coefficient between d_i and d_j.

During the training stage, we select edges where $c_{ij} \geq 0.9$ are positive samples, and recorded the index (i, j) into the positive set S_p. For negative samples, we still employ negative sampling given a positive edge c_{ij}, we randomly sampled one negative edge c_{ir}, where $c_{ir} < 0.9$, and recorded the sampled negative index (i, r) into the negative set S_n. The training objective is thus:

$$\mathcal{L}_{Comorbid} = \sum_{(i,j) \in S_p \cup S_n} -c_{ij} \log P(d_i, d_j) - (1 - c_{ij}) \log(1 - P(d_i, d_j)). \quad \text{(Equation 15)}$$

Experimental setting and hyperparameters choice. The dimensionality of pre-trained node embeddings K is set to 100. For the gene set member recovery task, the dimensionality of the embedding layer output is set to 256. The number of heads T is 8, the number of GAT layers is 2. The output representation dimensionality of each head in the first GAT layer is 128. We set the learning rate to $1e^{-3}$. During the pathway gene set recovery experiments, we followed the setting of Set2Gaussian to retrieve 50% of the gene set members as test data and used the remaining 50% as the training data. For the disease gene reprioritization task, we randomly masked 40% of the associations for disease-gene matrix D as test set and used the other 60% of data for training. We set the attention dropout to 0.3, and the learning rate to $5e^{-3}$. The hidden size of the GAT layer is 128 per head. We train the model for 7,100 epochs. For comorbidity prediction, we randomly hid 10% of edges of the comorbid disease matrix C as test set and used the remaining 90% as the training set (10-fold cross validation). We trained the bipartite Graphene model for 20 epochs (30 epochs for Decagon), with batch size of 512 and learning rate of $1e^{-3}$. The threshold of the input relative risk is 1.0.

GSEA. GSEApy (https://github.com/zqfang/GSEApy) API was used for enrichment analysis, where the p value was computed using the hypergeometric test and the $p_{adjusted}$ value using the Benjamini-Hochberg method for correction. The $p_{adjusted}$ value was reported. The following gene sets were included for SCZ: FMRP targets, PSD genes, $GABA_A$ receptor, calcium signaling, and glutamatergic synapse of KEGG. ASD related gene sets include database AutDB, ECGs, and targets of RBFOX1. In downstream analysis of disease gene prioritization, the following KEGG pathways were used for correlation analysis for eight mental disorders: Neuroactive ligand-receptor interaction, long-term depression, glutamatergic synapse, cell adhesion molecules, GABAergic synapse, calcium signaling pathway, glycosaminoglycan biosynthesis, and axon guidance. For two IBDs, i.e., ulcerative colitis and Crohn disease, we chose three pathways involved in immune system and signal transduction (mitogen-activated protein kinase) signaling, NF-κB signaling, Th17 cell differentiation) from KEGG (https://www.genome.jp/kegg/pathway.html) and we chose another three pathways of T cell activation (TCR signaling, CD28 co-stimulation, interleukin-2 family signaling) from Reactome.

Tissue specificity analysis. For the tissue-specificity analysis, we downloaded gene-level TPM (transcripts per kilobase million) data containing 53 tissues from GTEx portal (https://www.gtexportal.org/home/datasets) and adopted the JS divergence to measure the tissue specificity of each gene in each tissue. JS divergence is an entropy measurement that quantifies the similarity between a gene's expression pattern e and an extreme pattern where a gene is expressed in only one tissue e^t, and their JS divergence to be

$$JS(e, e^t) = H\left(\frac{e + e^t}{2}\right) - \frac{H(e) + H(e^t)}{2}, \quad \text{(Equation 16)}$$

where the entropy of a discrete probability distribution is denoted as H:

$$e = (e_1, e_2 \cdots e_n), 0 \le e_i \le 1 \; and \; \sum_{i=1}^{n} e_i = 1$$

$$e^t = (e_1^t, e_2^t \cdots e_n^t) \; and \; e_i^t = \begin{cases} 0, & (i \neq t) \\ 1, & (i = t) \end{cases} \quad \text{(Equation 17)}$$

$$H(p) = -\sum_{i=1}^{n} e_i \log(e_i).$$

The distance between two tissue expression patterns, e and e^t is defined as:

$$JS_{dist(e,e^t)} = \sqrt{JS(e, e^t)}. \quad \text{(Equation 18)}$$

Then the tissue-specific expression pattern of gene e with respect to tissue t can be defined as

$$JS_{sp}(e|t) = 1 - JS_{dist}(e, e^t). \quad \text{(Equation 19)}$$

Finally, Wilcoxon rank-sum test was adopted to calculate the overall expression pattern of genes relating to one disease.

Gene classification according to GO annotation. In the task of SCZ disease module identification (Figure S4), we devised a way of classifying module genes according to GO annotation. Like other mental diseases, the following functions played important roles: gene expression regulation: GO:0010468, GO:0032774, GO:0051252; synaptic signaling: GO:0099536, GO:0007154, GO:0023052, GO:0005737, GO:0007267; ion transport: GO:0006811, GO:0006810; cytoskeleton organization: GO:0070507, GO:0032886, GO:0000226, GO:0007010, GO:0006996; nervous system development: GO:0048854, GO:0009887, GO:0007399, GO:0050877, and so on. Each function class contains a certain amount of GO annotations. We classified a gene by searching the GO annotation hierarchy tree and see if the gene itself has any annotation belonging to a certain function or if any close ancestor of it does.

TPGs chosen for Jaccard Index calculation among eight mental disorders. The number of TPGs used for Jaccard Index calculation of eight mental disorders was chosen according to their GWAS association genes in training, and then normalized to a range from 100 to 500 (min-max normalization).

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100651.

ACKNOWLEDGMENTS

This research is supported by the Ministry of Science and Technology of the People's Republic of China's program titled "Science & Technology Boost Economy 2020" (SQ2020YFF0426292).

AUTHOR CONTRIBUTIONS

M.Y. conceived the problem and designed all detailed studies. Y.W., Z.J.S., and Q.S.H. performed analysis. M.N. coordinated the resources and facilitated insightful discussions. J.W.L. provided suggestions on pre-trained models. M.Y. and Y.W. wrote the manuscript.

DECLARATION OF INTERESTS

M.N. declares the following competing interests: stock holdings in MGI, BGI-Shenzhen.

Received: April 4, 2022
Revised: May 19, 2022
Accepted: November 7, 2022
Published: December 6, 2022

REFERENCES

1. Wong, A.K., Sealfon, R.S.G., Theesfeld, C.L., and Troyanskaya, O.G. (2021). Decoding disease: from genomes to networks to phenotypes. Nat. Rev. Genet. *22*, 774–790. https://doi.org/10.1038/s41576-021-00389-x.
2. Choobdar, S., Ahsen, M.E., Crawford, J., Tomasoni, M., Fang, T., Lamparter, D., Lin, J., Hescott, B., Hu, X., Mercer, J., et al. (2019). Assessment of network module identification across complex diseases. Nat. Methods *16*, 843–852. https://doi.org/10.1038/s41592-019-0509-5.
3. Greene, C.S., Krishnan, A., Wong, A.K., Ricciotti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., et al. (2015). Understanding multicellular function and disease with human tissue-specific networks. Nat. Genet. *47*, 569–576. https://doi.org/10.1038/ng.3259.
4. Carlin, D.E., Fong, S.H., Qin, Y., Jia, T., Huang, J.K., Bao, B., Zhang, C., and Ideker, T. (2019). A fast and flexible framework for network-assisted genomic association. iScience *16*, 155–161. https://doi.org/10.1016/j.isci.2019.05.025.
5. Wang, Q., Chen, R., Cheng, F., Wei, Q., Ji, Y., Yang, H., Zhong, X., Tao, R., Wen, Z., Sutcliffe, J.S., et al. (2019). A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. Nat. Neurosci. *22*, 691–699. https://doi.org/10.1038/s41593-019-0382-7.
6. Buphamalai, P., Kokotovic, T., Nagy, V., and Menche, J. (2021). Network analysis reveals rare disease signatures across multiple levels of biological organization. Nat. Commun. *12*, 6306. https://doi.org/10.1038/s41467-021-26674-1.
7. Cowen, L., Ideker, T., Raphael, B.J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. Nat. Rev. Genet. *18*, 551–562. https://doi.org/10.1038/nrg.2017.38.
8. Ata, S.K., Wu, M., Fang, Y., Ou-Yang, L., Kwoh, C.K., and Li, X.-L. (2021). Recent advances in network-based methods for disease gene prediction. Briefings Bioinf. *22*, bbaa303. https://doi.org/10.1093/bib/bbaa303.
9. Wang, S., Flynn, E.R., and Altman, R.B. (2020). Gaussian embedding for large-scale gene set analysis. Nat. Mach. Intell. *2*, 387–395. https://doi.org/10.1038/s42256-020-0193-2.
10. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452. https://doi.org/10.1093/nar/gku1003.
11. Huang, J.K., Carlin, D.E., Yu, M.K., Zhang, W., Kreisberg, J.F., Tamayo, P., and Ideker, T. (2018). Systematic evaluation of molecular networks for discovery of disease genes. Cell Syst. *6*, 484–495.e5. https://doi.org/10.1016/j.cels.2018.03.001.
12. Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2009). ConsensusPathDB—a database for integrating human functional interaction networks. Nucleic Acids Res. *37*, D623–D628. https://doi.org/10.1093/nar/gkn698.
13. Wong, A.K., Krishnan, A., and Troyanskaya, O.G. (2018). Giant 2.0: genome-scale integrated analysis of gene networks in tissues. Nucleic Acids Res. *46*, W65–W70. https://doi.org/10.1093/nar/gky408.
14. Picart-Armada, S., Barrett, S.J., Willé, D.R., Perera-Lluna, A., Gutteridge, A., and Dessailly, B.H. (2019). Benchmarking network propagation methods for disease gene identification. PLoS Comput. Biol. *15*, e1007276. https://doi.org/10.1371/journal.pcbi.1007276.
15. Cho, H., Berger, B., and Peng, J. (2016). Compact integration of multi-network topology for functional analysis of genes. Cell Syst. *3*, 540–548.e5. https://doi.org/10.1016/j.cels.2016.10.017.
16. Tong, H., Faloutsos, C., and Pan, J.y. (2006). Fast random walk with restart and its applications. In Sixth International Conference on Data Mining, 2006 (IEEE), pp. 613–622.

17. Gao, X., Ma, X., Zhang, W., Huang, J., Li, H., Li, Y., and Cui, J. (2022). Multi-view clustering with self-representation and structural Constraint. IEEE Trans. Big Data *8*, 882–893. https://doi.org/10.1109/TBDATA.2021.3128906.

18. Ma, X., Sun, P., and Gong, M. (2022). An integrative framework of heterogeneous genomic data for cancer Dynamic modules based on matrix decomposition. IEEE ACM Trans. Comput. Biol. Bioinf. *19*, 305–316. https://doi.org/10.1109/TCBB.2020.3004808.

19. Lin, Q., Lin, Y., Yu, Q., and Ma, X. (2020). Clustering of cancer attributed networks via integration of graph embedding and matrix factorization. IEEE Access *8*, 197463–197472. https://doi.org/10.1109/ACCESS.2020.3034623.

20. Grover, A., and Leskovec, J. (2016). node2vec: scalable feature learning for networks. In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining, 2016 (Association for Computing Machinery), pp. 855–864.

21. Peng, J., Xue, H., Wei, Z., Tuncali, I., Hao, J., and Shang, X. (2021). Integrating multi-network topology for gene function prediction using deep neural networks. Briefings Bioinf. *22*, 2096–2105. https://doi.org/10.1093/bib/bbaa036.

22. Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. Adv. Neural Inf. Process. Syst. *29*.

23. Kipf, T.N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1609.02907.

24. Hamilton, W., Ying, Z., and Leskovec, J. (2017). Inductive representation learning on large graphs. Adv. Neural Inf. Process. Syst. *30*.

25. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1710.10903.

26. Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks?. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.00826.

27. Torng, W., and Altman, R.B. (2019). Graph convolutional neural networks for predicting drug-target interactions. J. Chem. Inf. Model. *59*, 4131–4149. https://doi.org/10.1021/acs.jcim.9b00628.

28. Xu, H., Wang, H., Yuan, C., Zhai, Q., Tian, X., Wu, L., and Mi, Y. (2020). Identifying diseases that cause psychological trauma and social avoidance by GCN-Xgboost. BMC Bioinf. *21*, 504. https://doi.org/10.1186/s12859-020-03847-1.

29. Schulte-Sasse, R., Budach, S., Hnisz, D., and Marsico, A. (2021). Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. Nat. Mach. Intell. *3*, 513–526. https://doi.org/10.1038/s42256-021-00325-y.

30. Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics *34*, i457–i466. https://doi.org/10.1093/bioinformatics/bty294.

31. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.

32. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations (PMLR), pp. 1597–1607.

33. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. (2022). Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022 (IEEE), pp. 16000–16009.

34. Hu, W., Liu, B., Gomes, J., Zitnik, M., Liang, P., Pande, V., and Leskovec, J. (2019). Strategies for pre-training graph neural networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1905.12265.

35. Liu, Y., Jin, M., Pan, S., Zhou, C., Zheng, Y., Xia, F., et al. (2022). Graph self-supervised learning: a survey. In IEEE Transactions on Knowledge and Data Engineering (IEEE). 1–1.

36. Rosenstein, M., Marx, Z., Kaelbling, L., and Dietterich, T. (2005). To Transfer or Not to Transfer (NIPS).

37. Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L.I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. *45*, D833–D839. https://doi.org/10.1093/nar/gkw943.

38. McInnes, G., Tanigawa, Y., DeBoever, C., Lavertu, A., Olivieri, J.E., Aguirre, M., and Rivas, M.A. (2019). Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. Bioinformatics *35*, 2495–2497. https://doi.org/10.1093/bioinformatics/bty999.

39. Zeng, X., Zhang, X., and Zou, Q. (2016). Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. Briefings Bioinf. *17*, 193–203. https://doi.org/10.1093/bib/bbv033.

40. Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402–408. https://doi.org/10.1038/s41586-020-2188-x.

41. Ying, R., He, R., Chen, K., Eksombatchai, P., Hamilton, W.L., and Leskovec, J. (2018). Graph convolutional neural networks for web-scale recommender systems. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (Association for Computing Machinery), pp. 974–983.

42. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., et al. (2018). The reactome pathway knowledgebase. Nucleic Acids Res. *46*, D649–D655. https://doi.org/10.1093/nar/gkx1132.

43. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K.H. (2009). PID: the pathway interaction database. Nucleic Acids Res. *37*, D674–D679. https://doi.org/10.1093/nar/gkn653.

44. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47*, D1005–D1012. https://doi.org/10.1093/nar/gky1120.

45. Yin, T., Chen, S., Wu, X., and Tian, W. (2017). GenePANDA—a novel network-based gene prioritizing tool for complex diseases. Sci. Rep. *7*, 43258. https://doi.org/10.1038/srep43258.

46. Shim, J.E., Bang, C., Yang, S., Lee, T., Hwang, S., Kim, C.Y., Singh-Blom, U.M., Marcotte, E.M., and Lee, I. (2017). GWAB: a web server for the network-based boosting of human genome-wide association data. Nucleic Acids Res. *45*, W154–W161. https://doi.org/10.1093/nar/gkx284.

47. GTEx, C., Ardlie, K.G., Deluca, D.S., Segrè, A.V., Sullivan, T.J., Young, T.R., Gelfand, E.T., Trowbridge, C.A., Maller, J.B., Tukiainen, T., et al. (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science *348*, 648–660. https://doi.org/10.1126/science.1262110.

48. Fuglede, B., and Topsoe, F. (2004). Jensen-Shannon divergence and Hilbert space embedding. In International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings., 2004 (IEEE), p. 31.

49. Ascano, M., Mukherjee, N., Bandaru, P., Miller, J.B., Nusbaum, J.D., Corcoran, D.L., Langlois, C., Munschauer, M., Dewell, S., Hafner, M., et al. (2012). FMRP targets distinct mRNA sequence elements to regulate protein expression. Nature *492*, 382–386. https://doi.org/10.1038/nature11737.

50. Bayés, À., van de Lagemaat, L.N., Collins, M.O., Croning, M.D.R., Whittle, I.R., Choudhary, J.S., and Grant, S.G.N. (2011). Characterization of the proteome, diseases and evolution of the human postsynaptic density. Nat. Neurosci. *14*, 19–21. https://doi.org/10.1038/nn.2719.

51. Pocklington, A.J., Rees, E., Walters, J.T.R., Han, J., Kavanagh, D.H., Chambert, K.D., Holmans, P., Moran, J.L., McCarroll, S.A., Kirov, G., et al. (2015). Novel findings from CNVs implicate Inhibitory and Excitatory signaling Complexes in schizophrenia. Neuron *86*, 1203–1214. https://doi.org/10.1016/j.neuron.2015.04.022.

52. Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30. https://doi.org/10.1093/nar/28.1.27.

53. Ripke, S., Neale, B.M., Corvin, A., Walters, J.T.R., Farh, K.-H., Holmans, P.A., Lee, P., Bulik-Sullivan, B., Collier, D.A., Huang, H., et al. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature *511*, 421–427. https://doi.org/10.1038/nature13595.

54. Volk, L., Chiu, S.-L., Sharma, K., and Huganir, R.L. (2015). Glutamate synapses in human cognitive disorders. Annu. Rev. Neurosci. *38*, 127–149. https://doi.org/10.1146/annurev-neuro-071714-033821.

55. Weyn-Vanhentenryck, Sebastien, M., Mele, A., Yan, Q., Sun, S., Farny, N., Zhang, Z., Xue, C., Herre, M., Silver, P.A., et al. (2014). HITS-CLIP and integrative modeling define the Rbfox Splicing-regulatory network linked to brain development and autism. Cell Rep. *6*, 1139–1152. https://doi.org/10.1016/j.celrep.2014.02.005.

56. Basu, S.N., Kollu, R., and Banerjee-Basu, S. (2009). AutDB: a gene reference resource for autism research. Nucleic Acids Res. *37*, D832–D836. https://doi.org/10.1093/nar/gkn835.

57. Samocha, K.E., Robinson, E.B., Sanders, S.J., Stevens, C., Sabo, A., McGrath, L.M., Kosmicki, J.A., Rehnström, K., Mallick, S., Kirby, A., et al. (2014). A framework for the interpretation of de novo mutation in human disease. Nat. Genet. *46*, 944–950. https://doi.org/10.1038/ng.3050.

58. Liu, J.Z., van Sommeren, S., Huang, H., Ng, S.C., Alberts, R., Takahashi, A., Ripke, S., Lee, J.C., Jostins, L., Shah, T., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nat. Genet. *47*, 979–986. https://doi.org/10.1038/ng.3359.

59. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise Explanations for non-linear classifier Decisions by layer-wise relevance propagation. PLoS One *10*, e0130140. https://doi.org/10.1371/journal.pone.0130140.

60. Mouton, J., Ronjat, M., Jona, I., Villaz, M., Feltz, A., and Maulet, Y. (2001). Skeletal and cardiac ryanodine receptors bind to the Ca2+-sensor region of dihydropyridine receptor α1C subunit. FEBS (Fed. Eur. Biochem. Soc.) Lett. *505*, 441–444. https://doi.org/10.1016/S0014-5793(01)02866-6.

61. Martin, C., Chapman, K.E., Seckl, J.R., and Ashley, R.H. (1998). Partial cloning and differential expression of ryanodine receptor/calcium-release channel genes in human tissues including the hippocampus and cerebellum. Neuroscience *85*, 205–216. https://doi.org/10.1016/S0306-4522(97)00612-X.

62. Lanner, J.T., Georgiou, D.K., Joshi, A.D., and Hamilton, S.L. (2010). Ryanodine receptors: structure, expression, molecular details, and function in calcium release. Cold Spring Harb. Perspect. Biol. *2*, a003996.

63. Tu, J.C., Xiao, B., Naisbitt, S., Yuan, J.P., Petralia, R.S., Brakeman, P., Doan, A., Aakalu, V.K., Lanahan, A.A., Sheng, M., and Worley, P.F. (1999). Coupling of mGluR/Homer and PSD-95 Complexes by the Shank family of postsynaptic density proteins. Neuron *23*, 583–592. https://doi.org/10.1016/S0896-6273(00)80810-7.

64. Greenwood, T.A., Lazzeroni, L.C., Murray, S.S., Cadenhead, K.S., Calkins, M.E., Dobie, D.J., Green, M.F., Gur, R.E., Gur, R.C., Hardiman, G., et al. (2011). Analysis of 94 candidate genes and 12 Endophenotypes for schizophrenia from the Consortium on the genetics of schizophrenia. Am. J. Psychiatr. *168*, 930–946. https://doi.org/10.1176/appi.ajp.2011.10050723.

65. Sweeney, C., Lai, C., Riese, D.J., Diamonti, A.J., II, Cantley, L.C., and Carraway, K.L., III (2000). Ligand discrimination in signaling through an ErbB4 receptor Homodimer. J. Biol. Chem. *275*, 19803–19807. https://doi.org/10.1074/jbc.C901015199.

66. Howard, D.M., Adams, M.J., Clarke, T.-K., Hafferty, J.D., Gibson, J., Shirali, M., Coleman, J.R.I., Hagenaars, S.P., Ward, J., Wigmore, E.M., et al. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat. Neurosci. *22*, 343–352. https://doi.org/10.1038/s41593-018-0326-7.

67. Bi, L.-L., Sun, X.-D., Zhang, J., Lu, Y.-S., Chen, Y.-H., Wang, J., Geng, F., Liu, F., Zhang, M., Liu, J.-H., et al. (2015). Amygdala NRG1–ErbB4 is Critical for the Modulation of anxiety-like behaviors. Neuropsychopharmacology *40*, 974–986. https://doi.org/10.1038/npp.2014.274.

68. Gliemann, J., Hermey, G., NykjÆR, A., Petersen, C.M., Jacobsen, C., and Andreasen, P.A. (2004). The mosaic receptor sorLA/LR11 binds components of the plasminogen-activating system and platelet-derived growth factor-BB similarly to LRP1 (low-density lipoprotein receptor-related protein), but mediates slow internalization of bound ligand. Biochem. J. *381*, 203–212. https://doi.org/10.1042/BJ20040149.

69. Marchianò, S., Catapano, A.L., Corsini, A., and Ferri, N. (2017). PCSK9 modulates phenotype, proliferation and migration of smooth muscle cells in response to PDGF-BB. Nutr. Metabol. Cardiovasc. Dis. *27*, e28. https://doi.org/10.1016/j.numecd.2016.11.076.

70. Gustafsen, C., Kjolby, M., Nyegaard, M., Mattheisen, M., Lundhede, J., Buttenschøn, H., Mors, O., Bentzon, J.F., Madsen, P., Nykjaer, A., and Glerup, S. (2014). The Hypercholesterolemia-risk gene SORT1 facilitates PCSK9 Secretion. Cell Metabol. *19*, 310–318. https://doi.org/10.1016/j.cmet.2013.12.006.

71. Belinky, F., Nativ, N., Stelzer, G., Zimmerman, S., Iny Stein, T., Safran, M., and Lancet, D. (2015). PathCards: multi-source consolidation of human biological pathways. Database *2015*, bav006. https://doi.org/10.1093/database/bav006.

72. Helkkula, P., Kiiskinen, T., Havulinna, A.S., Karjalainen, J., Koskinen, S., Salomaa, V., Daly, M.J., Palotie, A., Surakka, I., Ripatti, S., and FinnGen. (2021). ANGPTL8 protein-truncating variant associated with lower serum triglycerides and risk of coronary disease. PLoS Genet. *17*, e1009501. https://doi.org/10.1371/journal.pgen.1009501.

73. Alavi Naini, S.M., and Soussi-Yanicostas, N. (2018). Heparan sulfate as a therapeutic target in Tauopathies: insights from Zebrafish. Front. Cell Dev. Biol. *6*. https://doi.org/10.3389/fcell.2018.00163.

74. Clarke, H.E. (2017). Altered Heparan Sulfate in Ageing and Dementia: A Potential Axis for the Dysregulation of BACE-1 in Alzheimer's Disease (The University of Liverpool).

75. Rong, J., Habuchi, H., Kimata, K., Lindahl, U., and Kusche-Gullberg, M. (2001). Substrate specificity of the heparan sulfate Hexuronic acid 2-O-sulfotransferase. Biochemistry *40*, 5548–5555. https://doi.org/10.1021/bi002926p.

76. Thacker, B.E., Xu, D., Lawrence, R., and Esko, J.D. (2014). Heparan sulfate 3-O-sulfation: a rare modification in search of a function. Matrix Biol. *35*, 60–72. https://doi.org/10.1016/j.matbio.2013.12.001.

77. Thacker, B.E., Seamen, E., Lawrence, R., Parker, M.W., Xu, Y., Liu, J., Vander Kooi, C.W., and Esko, J.D. (2016). Expanding the 3-O-sulfate proteome—enhanced binding of Neuropilin-1 to 3-O-sulfated heparan sulfate modulates its activity. ACS Chem. Biol. *11*, 971–980. https://doi.org/10.1021/acschembio.5b00897.

78. Kantor, D.B., Chivatakarn, O., Peer, K.L., Oster, S.F., Inatani, M., Hansen, M.J., Flanagan, J.G., Yamaguchi, Y., Sretavan, D.W., Giger, R.J., and Kolodkin, A.L. (2004). Semaphorin 5A is a bifunctional axon guidance Cue regulated by heparan and Chondroitin sulfate proteoglycans. Neuron *44*, 961–975. https://doi.org/10.1016/j.neuron.2004.12.002.

79. Pérez, Y., Bonet, R., Corredor, M., Domingo, C., Moure, A., Messeguer, À., Bujons, J., and Alfonso, I. (2021). Semaphorin 3A—glycosaminoglycans interaction as therapeutic target for axonal regeneration. Pharmaceuticals *14*. https://doi.org/10.3390/ph14090906.

80. Choi, B.Y. (2020). Targeting Wnt/β-catenin pathway for developing therapies for hair loss. Int. J. Mol. Sci. *21*. https://doi.org/10.3390/ijms21144915.

81. Liu, Q., Shi, X., Zhang, Y., Huang, Y., Yang, K., Tang, Y., Ma, Y., Zhang, Y., Wang, J.a., Zhang, L., et al. (2021). Increased expression of Zyxin and its potential function in androgenetic alopecia. Front. Cell Dev. Biol. *8*. https://doi.org/10.3389/fcell.2020.582282.

82. Zhang, P., and Dressler, G.R. (2013). The Groucho protein Grg4 suppresses Smad7 to activate BMP signaling. Biochem. Biophys. Res. Commun. *440*, 454–459. https://doi.org/10.1016/j.bbrc.2013.09.128.

83. Li, J., Zhou, L., Ouyang, X., and He, P. (2021). Transcription factor-7-like-2 (TCF7L2) in atherosclerosis: a potential biomarker and therapeutic target. Front. Cardiovasc. Med. *8*. https://doi.org/10.3389/fcvm.2021.701279.

84. Nakano, N., Itoh, S., Watanabe, Y., Maeyama, K., Itoh, F., and Kato, M. (2010). Requirement of TCF7L2 for TGF-β-dependent transcriptional activation of the TMEPAI gene. J. Biol. Chem. *285*, 38023–38033. https://doi.org/10.1074/jbc.M110.132209.

85. Zhang, S., Wang, Y., Zhu, X., Song, L., Zhan, X., Ma, E., McDonough, J., Fu, H., Cambi, F., Grinspan, J., and Guo, F. (2021). The Wnt effector TCF7l2 promotes oligodendroglial differentiation by repressing autocrine BMP4-Mediated signaling. J. Neurosci. *41*, 1650. https://doi.org/10.1523/JNEUROSCI.2386-20.2021.

86. Consortium, B., Anttila, V., Bulik-Sullivan, B., Finucane, H.K., Walters, R.K., Bras, J., Duncan, L., Escott-Price, V., Falcone, G.J., Gormley, P., et al. (2018). Analysis of shared heritability in common disorders of the brain. Science *360*, eaap8757. https://doi.org/10.1126/science.aap8757.

87. Peyrot, W.J., and Price, A.L. (2021). Identifying loci with different allele frequencies among cases of eight psychiatric disorders using CC-GWAS. Nat. Genet. *53*, 445–454. https://doi.org/10.1038/s41588-021-00787-1.

88. Bulik-Sullivan, B., Finucane, H.K., Anttila, V., Gusev, A., Day, F.R., Loh, P.-R., Duncan, L., Perry, J.R.B., Patterson, N., Robinson, E.B., et al. (2015). An atlas of genetic correlations across human diseases and traits. Nat. Genet. *47*, 1236–1241. https://doi.org/10.1038/ng.3406.

89. Berkowicz, S.R., Featherby, T.J., Qu, Z., Giousoh, A., Borg, N.A., Heng, J.I., Whisstock, J.C., and Bird, P.I. (2016). Brinp1−/−mice exhibit autism-like behaviour, altered memory, hyperactivity and increased parvalbumin-positive cortical interneuron density. Mol. Autism. *7*, 22. https://doi.org/10.1186/s13229-016-0079-7.

90. Flati, T., Gioiosa, S., Chillemi, G., Mele, A., Oliverio, A., Mannironi, C., Rinaldi, A., and Castrignanò, T. (2020). A gene expression atlas for different kinds of stress in the mouse brain. Sci. Data *7*, 437. https://doi.org/10.1038/s41597-020-00772-z.

91. Schreiber, E., Tobler, A., Malipiero, U., Schaffner, W., and Fontana, A. (1993). cDNA cloning of human N-Oct 3, a nervous-system specific POU domain transcription factor binding to the octamer DNA motif. Nucleic Acids Res. *21*, 253–258. https://doi.org/10.1093/nar/21.2.253.

92. Chen, C., Meng, Q., Xia, Y., Ding, C., Wang, L., Dai, R., Cheng, L., Gunaratne, P., Gibbs, R.A., Min, S., et al. (2018). The transcription factor POU3F2 regulates a gene coexpression network in brain tissue from patients with psychiatric disorders. Sci. Transl. Med. *10*, eaat8178. https://doi.org/10.1126/scitranslmed.aat8178.

93. Gutman, G.A., Chandy, K.G., Grissmer, S., Lazdunski, M., McKinnon, D., Pardo, L.A., Robertson, G.A., Rudy, B., Sanguinetti, M.C., Stühmer, W., and Wang, X. (2005). International union of pharmacology. LIII. Nomenclature and molecular relationships of voltage-gated potassium channels. Pharmacol. Rev. *57*, 473. https://doi.org/10.1124/pr.57.4.10.

94. Schanze, I., Bunt, J., Lim, J.W.C., Schanze, D., Dean, R.J., Alders, M., Blanchet, P., Attié-Bitach, T., Berland, S., Boogert, S., et al. (2018). NFIB Haploinsufficiency is associated with Intellectual Disability and Macrocephaly. Am. J. Hum. Genet. *103*, 752–768. https://doi.org/10.1016/j.ajhg.2018.10.006.

95. Menche, J., Sharma, A., Kitsak, M., Ghiassian, S.D., Vidal, M., Loscalzo, J., and Barabási, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. Science *347*, 1257601. https://doi.org/10.1126/science.1257601.

96. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. *33*, D514–D517. https://doi.org/10.1093/nar/gki033.

97. Ramos, E.M., Hoffman, D., Junkins, H.A., Maglott, D., Phan, L., Sherry, S.T., Feolo, M., and Hindorff, L.A. (2014). Phenotype–Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. Eur. J. Hum. Genet. *22*, 144–147. https://doi.org/10.1038/ejhg.2013.96.

98. Akram, P., and Liao, L. (2019). Prediction of comorbid diseases using weighted geometric embedding of human interactome. BMC Med. Genom. *12*, 161. https://doi.org/10.1186/s12920-019-0605-5.

Patterns, Volume *4*

Supplemental information

Self-supervised graph representation learning integrates multiple molecular networks and decodes gene-disease relationships

Yi Wang, Zijun Sun, Qiushun He, Jiwei Li, Ming Ni, and Meng Yang

Supplementary Figures

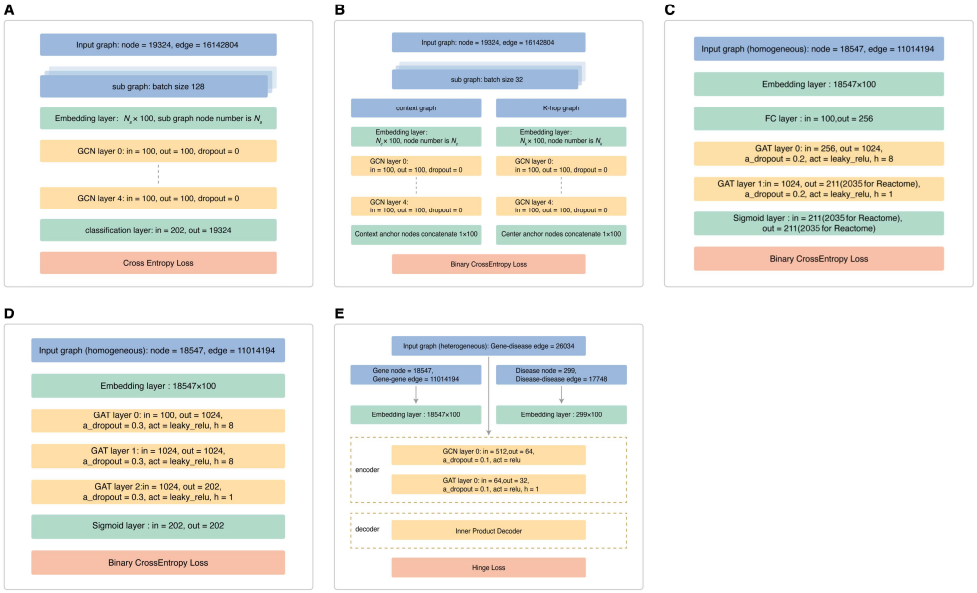


Figure S1. Schematic diagram of Graphene model architecture. (A) Pre-training for masked node recovery (B) Pre-training for context prediction (C) Re-training for gene set member identification task (D) Re-training for disease gene reprioritization task (E) Bipartite architecture for comorbidity prediction task.

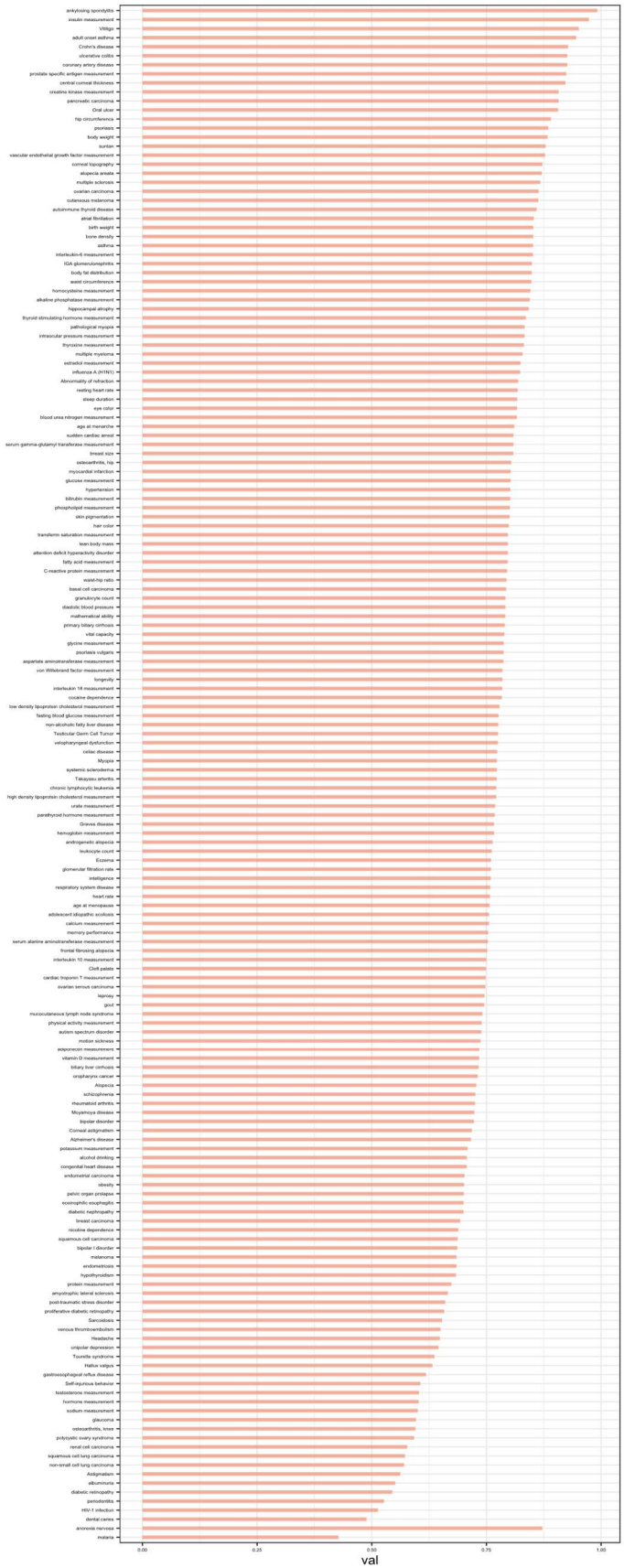


Figure S2. AUROC results for all 171 DisGeNET diseases. Diseases are represented by UMLS CUI ids.

Figure S3. Tissue specificity heatmap of original GWAS Catalog hits.

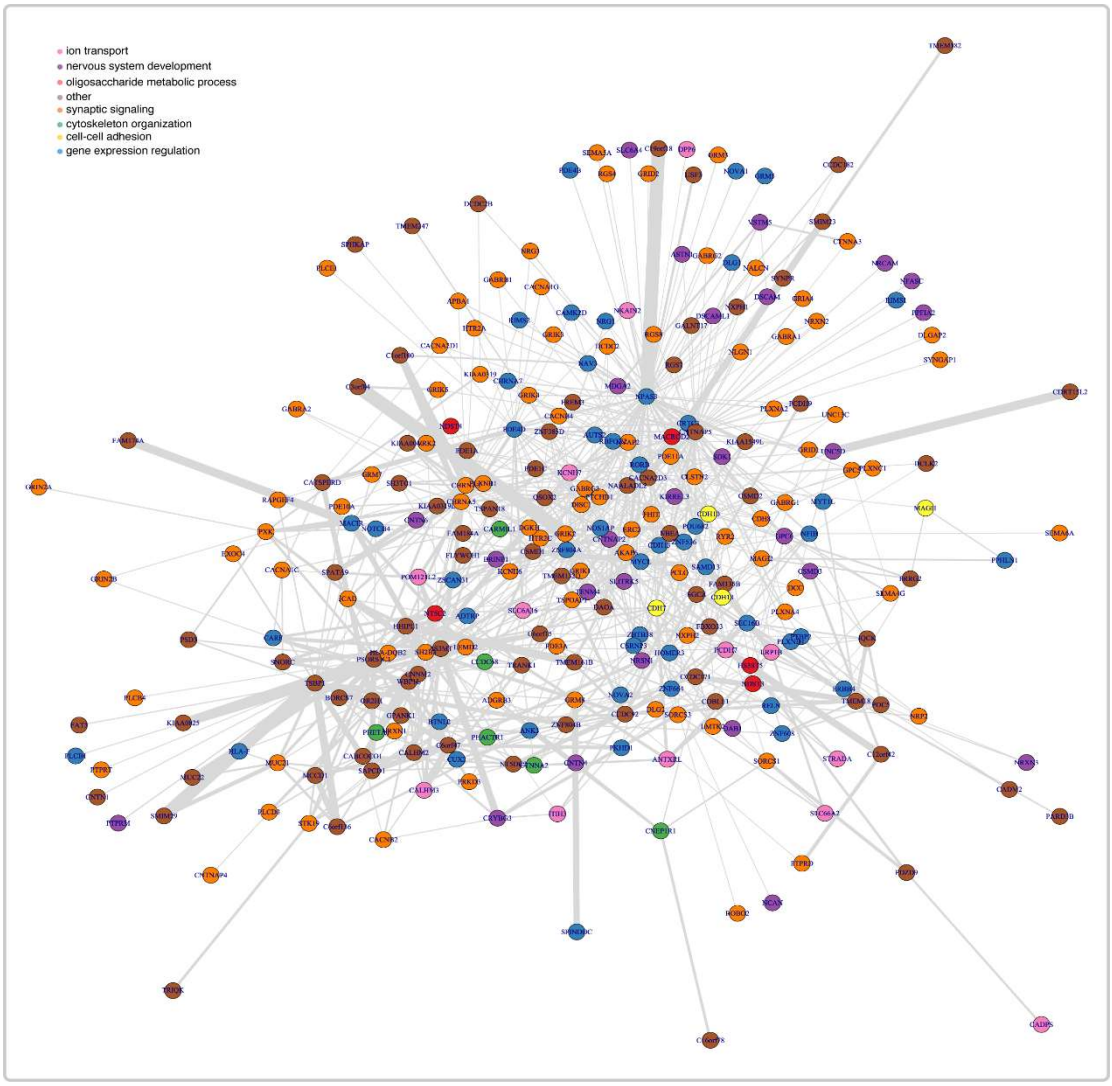


Figure S4. Attention weights showcase interaction patterns of 300 Graphene TPGs for SCZ.

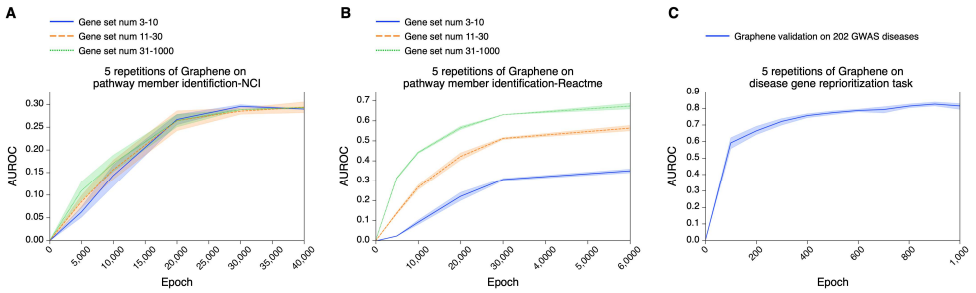


Figure S5. Evaluation with 5 repetitions of Graphene on pathway member identification task and disease gene reprioritization task.

Supplementary Tables

Table S1. Datasets used in this study

Dataset	Download link
Humanbase	https://hb.flatironinstitute.org/download
HuRI	http://www.interactome-atlas.org/download
STRING9606	https://string-db.org
PCNet	http://www.ndexbio.org
NCI(PID)	http://www.ndexbio.org
Reactome	https://reactome.org/download-data
GWAS Catalog	https://www.ebi.ac.uk/gwas/docs/file-downloads
DisGeNET	https://www.disgenet.org/downloads
UK BioBank	https://pan.ukbb.broadinstitute.org/downloads

Table S2. Runtime required to obtain the results of Figure 2C, 2D in downstream re-training of Graphene and competing methods, evaluated on DisGeNET cross-dataset test (171 diseases) and UK Biobank cross-dataset test (81 diseases) by a single Quadro RTX 6000 GPU.

	GenePanda	Graphene	GWAS P-value	N2V	NAGA	Set2Gaussian
UK Biobank cross-dataset test	2,840 seconds	295 seconds	243 seconds	>2 days	2,358 seconds	1,615 seconds
DisGeNET cross-dataset test	5,960 seconds	306 seconds	513 seconds	>4 days	4,980 seconds	1,620 seconds

Table S3. Methods comparison for pathway gene set recovery

Dataset	Method	AUPRC		
		3-10	**11-30**	**31-1000**
NCI	Mean	0.30	0.22	0.19
	Set2Gaussian	0.31	0.27	0.26
	Graphene (random input embedding)	0.24	0.27	0.29
	Graphene (pre-trained input embedding)	0.29	0.31	0.29
Reactome	Mean	0.40	0.47	0.49
	Set2Gaussian	0.43	0.56	0.63
	Graphene (random input embedding)	0.29	0.52	0.67
	Graphene (pre-trained input embedding)	0.42	0.58	0.69

Table S4. Normalized Jaccard Index of GWAS hits for eight diseases

Disease1	Disease2	value
unipolar depression	post-traumatic stress disorder	0.0792
unipolar depression	attention deficit hyperactivity disorder	0.368
unipolar depression	bipolar disorder	0.324
unipolar depression	schizophrenia	0.863
unipolar depression	autism spectrum disorder	0.154
unipolar depression	Tourette syndrome	0.102
unipolar depression	anorexia nervosa	0.121
post-traumatic stress disorder	attention deficit hyperactivity disorder	0.122
post-traumatic stress disorder	bipolar disorder	0.0298
post-traumatic stress disorder	schizophrenia	0.0847
post-traumatic stress disorder	autism spectrum disorder	0
post-traumatic stress disorder	Tourette syndrome	0
post-traumatic stress disorder	anorexia nervosa	0.133
attention deficit hyperactivity disorder	bipolar disorder	0.141
attention deficit hyperactivity disorder	schizophrenia	0.267
attention deficit hyperactivity disorder	autism spectrum disorder	0.388
attention deficit hyperactivity disorder	Tourette syndrome	0.0428
attention deficit hyperactivity disorder	anorexia nervosa	0.126
bipolar disorder	schizophrenia	1
bipolar disorder	autism spectrum disorder	0.114
bipolar disorder	Tourette syndrome	0.0934
bipolar disorder	anorexia nervosa	0.185
schizophrenia	autism spectrum disorder	0.179
schizophrenia	Tourette syndrome	0.0490
schizophrenia	anorexia nervosa	0.0486
autism spectrum disorder	Tourette syndrome	0
autism spectrum disorder	anorexia nervosa	0.106
Tourette syndrome	anorexia nervosa	0.495

Table S5. Normalized Jaccard Index of DisGeNET genes for eight diseases

Disease1	Disease2	value
unipolar depression	post-traumatic stress disorder	0.464
unipolar depression	attention deficit hyperactivity disorder	0.115
unipolar depression	bipolar disorder	0.775
unipolar depression	schizophrenia	0.560
unipolar depression	autism spectrum disorder	0.0562
unipolar depression	Tourette syndrome	0.341
unipolar depression	anorexia nervosa	0.343
post-traumatic stress disorder	attention deficit hyperactivity disorder	0.0562
post-traumatic stress disorder	bipolar disorder	0.507
post-traumatic stress disorder	schizophrenia	0.446
post-traumatic stress disorder	autism spectrum disorder	0.0137
post-traumatic stress disorder	Tourette syndrome	0.297
post-traumatic stress disorder	anorexia nervosa	0.487
attention deficit hyperactivity disorder	bipolar disorder	0.0542
attention deficit hyperactivity disorder	schizophrenia	0.0258
attention deficit hyperactivity disorder	autism spectrum disorder	0
attention deficit hyperactivity disorder	Tourette syndrome	0.157
attention deficit hyperactivity disorder	anorexia nervosa	0.111
bipolar disorder	schizophrenia	1
bipolar disorder	autism spectrum disorder	0.0613
bipolar disorder	Tourette syndrome	0.307
bipolar disorder	anorexia nervosa	0.293
schizophrenia	autism spectrum disorder	0.116
schizophrenia	Tourette syndrome	0.258
schizophrenia	anorexia nervosa	0.200
autism spectrum disorder	Tourette syndrome	0.108
autism spectrum disorder	anorexia nervosa	0.0413
Tourette syndrome	anorexia nervosa	0.329

Table S6. Normalized Jaccard Index of Graphene TPGs for eight diseases

Disease1	Disease2	value
unipolar depression	post-traumatic stress disorder	0.455
unipolar depression	attention deficit hyperactivity disorder	0.479
unipolar depression	bipolar disorder	0.938
unipolar depression	schizophrenia	1
unipolar depression	autism spectrum disorder	0.536
unipolar depression	Tourette syndrome	0.0203
unipolar depression	anorexia nervosa	0.346
post-traumatic stress disorder	attention deficit hyperactivity disorder	0.407
post-traumatic stress disorder	bipolar disorder	0.596
post-traumatic stress disorder	schizophrenia	0.452
post-traumatic stress disorder	autism spectrum disorder	0.937
post-traumatic stress disorder	Tourette syndrome	0.363
post-traumatic stress disorder	anorexia nervosa	0.315
attention deficit hyperactivity disorder	bipolar disorder	0.455
attention deficit hyperactivity disorder	schizophrenia	0.458
attention deficit hyperactivity disorder	autism spectrum disorder	0.406
attention deficit hyperactivity disorder	Tourette syndrome	0.0662
attention deficit hyperactivity disorder	anorexia nervosa	0.504
bipolar disorder	schizophrenia	0.955
bipolar disorder	autism spectrum disorder	0.405
bipolar disorder	Tourette syndrome	0
bipolar disorder	anorexia nervosa	0.372
schizophrenia	autism spectrum disorder	0.343
schizophrenia	Tourette syndrome	0.0271
schizophrenia	anorexia nervosa	0.309
autism spectrum disorder	Tourette syndrome	0.404
autism spectrum disorder	anorexia nervosa	0.439
Tourette syndrome	anorexia nervosa	0.0324

Table S7. Normalized Jaccard Index of NAGA rankings for eight diseases

Disease1	Disease2	value
unipolar depression	post-traumatic stress disorder	0.09
unipolar depression	attention deficit hyperactivity disorder	0.10
unipolar depression	bipolar disorder	0.36
unipolar depression	schizophrenia	1.00
unipolar depression	autism spectrum disorder	0.26
unipolar depression	Tourette syndrome	0.00
unipolar depression	anorexia nervosa	0.15
post-traumatic stress disorder	attention deficit hyperactivity disorder	0.01
post-traumatic stress disorder	bipolar disorder	0.05
post-traumatic stress disorder	schizophrenia	0.05
post-traumatic stress disorder	autism spectrum disorder	0.09
post-traumatic stress disorder	Tourette syndrome	0.07
post-traumatic stress disorder	anorexia nervosa	0.07
attention deficit hyperactivity disorder	bipolar disorder	0.06
attention deficit hyperactivity disorder	schizophrenia	0.06
attention deficit hyperactivity disorder	autism spectrum disorder	0.14
attention deficit hyperactivity disorder	Tourette syndrome	0.04
attention deficit hyperactivity disorder	anorexia nervosa	0.01
bipolar disorder	schizophrenia	0.95
bipolar disorder	autism spectrum disorder	0.03
bipolar disorder	Tourette syndrome	0.03
bipolar disorder	anorexia nervosa	0.05
schizophrenia	autism spectrum disorder	0.11
schizophrenia	Tourette syndrome	0.04
schizophrenia	anorexia nervosa	0.05
autism spectrum disorder	Tourette syndrome	0.03
autism spectrum disorder	anorexia nervosa	0.52
Tourette syndrome	anorexia nervosa	0.17

Table S8. cross-trait LD score for eight diseases

Disease1	Disease2	value
unipolar depression	attention deficit hyperactivity disorder	0.629
unipolar depression	anorexia nervosa	0.400
unipolar depression	autism spectrum disorder	0.486
unipolar depression	Tourette syndrome	0.329
post-traumatic stress disorder	unipolar depression	0
post-traumatic stress disorder	attention deficit hyperactivity disorder	0
post-traumatic stress disorder	bipolar disorder	0
post-traumatic stress disorder	schizophrenia	0
post-traumatic stress disorder	autism spectrum disorder	0
post-traumatic stress disorder	Tourette syndrome	0
post-traumatic stress disorder	anorexia nervosa	0
attention deficit hyperactivity disorder	anorexia nervosa	0.0143
attention deficit hyperactivity disorder	autism spectrum disorder	0.529
attention deficit hyperactivity disorder	Tourette syndrome	0.271
bipolar disorder	unipolar depression	0.471
bipolar disorder	attention deficit hyperactivity disorder	0.257
bipolar disorder	anorexia nervosa	0.143
bipolar disorder	autism spectrum disorder	0.243
bipolar disorder	Tourette syndrome	0.114
schizophrenia	bipolar disorder	1
schizophrenia	unipolar depression	0.443
schizophrenia	attention deficit hyperactivity disorder	0.229
schizophrenia	anorexia nervosa	0.371
schizophrenia	autism spectrum disorder	0.357
schizophrenia	Tourette syndrome	0.157
autism spectrum disorder	Tourette syndrome	0.229
Tourette syndrome	anorexia nervosa	0.114
anorexia nervosa	autism spectrum disorder	0.157

Table S9. Spearman correlation of the four methods (Graphene, DisGeNet, NAGA, GWAS) against ct-LDSC.

Methods	Graphene	DisGeNet	NAGA	GWAS
Spearman correlation	0.767	0.596	0.770	0.731

Table S10. Overlapping TPGs identified by Graphene for 4 disease pairs.

Disease Pair	Overlapping genes
BIP vs MDD	PDE1A,GRID1,GRIA4,GRID2,KCNC2,PLCH1,MDGA2,GABRB2,FAM104A,CARMIL1,GRM3,RAPGEF4,NRXN1,GRM5,DGKH,NCAN,ANTXRL,CTNNA2,TMPRSS7,GABRG3,GABRB3,CNTNAP2,CNTN6,SCN8A,RIMS1,PDE1C,GRM1,NYAP2,MAGI2,BTNL2,ZNF804A,PDE11A,FHIT,CBLN4,RBFOX1,DAOA,DISC1,GRIK2,GRIK4,CACNA1E,GRM8,GRIA1,KCNA4,SGCZ,NXPH1,NBEA,SAMD13,NBPF3,CSRNP3,TMEM247,KCNB2,MTUS2,MACROD2,CNTNAP5,CLSTN2,TMEM108,CSMD1,DAB1,KCNQ3,CPNE4,FAM135B,DLG2,RAPGEF5,GABRG2,GRM7,CDH8,RGS7,STPG3,AKAP6,KCND2,CACNB4,ERC2,GABRA2,CADM2,CACNA1C,CACNA2D3,CSMD3,TMEM161B,KCNV1,KCNH7,RIMS2,LRP1B,GRIK1,RESF1,AUTS2,SDK1,CACNA2D1,NKAIN2,PTPRT,CNTNAP4,OPCML,PCLO,CNTN1,NAV3,TRANK1,SPINDOC,PTPRM,THSD7B,SMIM22,CFAP54,NRG3,KCNH6,NALCN,PTPRD,NXPH2,SCN1A,SCN2A,KCND3,KCNC1,ZNF664,PCDH7,GABRB1,NPAS3,CDH13,CSMD2,VSTM5,GRIN2A,NAALADL2,KCNB1,RGS8,NRXN3,NELL1
SCZ vs MDD	PDE1A,CADPS,GRID1,GRIA2,MUC21,GRIA4,GRID2,FREM3,PLCH1,SMIM29,HLA-G,NOVA2,MDGA2,GABRB2,SNAP91,CARMIL1,GRM3,NRXN1,MAGI1,RAPGEF4,GRM5,DGKH,NCAN,ANTXRL,CTNNA2,SORCS3,FAM174A,FAT3,GABRB3,GABRG3,CNTNAP2,PTPRK,SCN8A,NETO1,CNTN6,TSPAN18,CTNND2,POU6F2,RIMS1,PDE1C,GPC6,GRM1,NYAP2,MAGI2,UNC13C,BTNL2,CDH9,ROBO2,ZNF804A,ASIC2,ZNF536,RAB3C,PDE11A,FHIT,RBFOX1,GRIN2B,DAOA,DISC1,GRIK2,GRIK4,GRIA1,CACNA1E,GRM8,CNTN4,ANKS1B,SGCZ,PPFIA2,CTNNA3,NXPH1,NBEA,SAMD13,UNC5D,TMEM247,MTUS2,CSRNP3,MACROD2,NRG1,CNTNAP5,JCAD,RELN,CLSTN2,SAPCD1,TRIM26,SYNPR,TMEM108,LMTK2,PCDH9,CSMD1,DAB1,C6orf47,CPNE4,FAM135B,DLG2,RAPGEF5,GABRG2,GRM7,CDH8,ASTN1,ZNF385D,STUM,PLXNC1,RGS7,DSCAML1,SEMA6D,DLG1,STPG3,RBFOX2,AKAP6,NPIPB5,NDST4,CACNB4,ERC2,GABRA2,CADM2,SNORC,TMEM161B,CSMD3,CACNA1C,PCDH17,CACNA2D3,ROBO1,KCNH7,NOVA1,RIMS2,LRP1B,GRIK1,CNTN5,RESF1,SDK1,C3orf84,AUTS2,CACNA2D1,CNEP1R1,NKAIN2,PTPRT,OPCML,CNTNAP4,PCLO,DSCAM,CNTN1,GPC5,NAV3,PRRG2,TRANK1,CDRT15L2,DPP6,NRCAM,SPINDOC,PTPRM,NRXN2,SMIM22,SEMA5A,GALNT17,KCNQ5,NRG3,CBLN2,KCNH6,NALCN,PTPRD,NXPH2,CDH7,SCN1A,SCN2A,DCC,QSOX2,CDH18,HOMER3,ZNF664,PCDH7,GABRB1,NLGN1,NFASC,NRXN3,NPAS3,CDH13,CSMD2,GRIK3,VSTM5,GRIN2A,NAALADL2,RGS8,CDH10,C6orf136,UNC5C,NELL1,NCAM2
MDD vs ADHD	FHIT,CADPS,RBFOX1,DAOA,GRM7,SPINDOC,CDH8,PTPRM,ZNF385D,GRID2,THSD7B,SMIM22,CFAP54,SEMA5A,NRG3,SGCZ,SEMA6D,PCDHGA1,AKAP6,NRXN1,NXPH1,CTNNA3,NPIPB5,SAMD13,ERC2,UNC5D,NALCN,PTPRD,CADM2,CTNNA2,MACROD2,CNTNAP5,RELN,DCC,CLSTN2,CACNA2D3,GABRG3,CNTNAP2,PCDH7,NLGN1,ROBO1,CTNND2,NPAS3,GPC6,LRP1B,CDH13,PCDH9,CSMD1,GRIK1,CSMD2,NYAP2,SDK1,C3orf84,CPNE4,ROBO2,CDH9,CACNA2D1,NAALADL2,FAM135B,NRXN3,UNC5C,GPC5,NELL1,CDH10,NAV3
PTSD vs ASD	CBLN4,CSMD1,NCAN,FAM135B,CNTNAP2,SGCZ,SAMD13,TSBP1,NRXN2,LRP1B,SMIM22,GPC5,CNTNAP5,NRXN3,GRID1,MACROD2,GALNT17,HS3ST3A1,EXTL2,GPC6,HS3ST1,DAOA,HS2ST1,HS3ST3B1,NRXN1,HS3ST5,NDST4,RBFOX1,NDST2,CADPS,NKAIN2,PCDH9,GLCE,NDST3,NLGN1,HS3ST2,CNTNAP4,BRINP1,ARSK,GRID2,ERC2,FAM20B,UST,CADM2,GPC2,CACNA2D3,MACIR

Table S11. Non-overlapping CC-GWAS genes identified by Graphene for 9 disease pairs.

Disease Pair	Graphene prediction of CC-GWAS Gene(symbol)
ANO vs TS	POU3F2
BIP vs ADHD	TRANK1, SCN2A
SCZ vs ADHD	TRANK1
SCZ vs MDD	ZNF536, GRM3, PTBP2, DGKI, TRANK1, AS3MT, SATB2, ZNF804A, SDCCAG8
SCZ vs ANO	SPHKAP, IMMP2L, SEMA6D(not in training set), GRIA1, DLG2, CHRNA3, CNNM2
SCZ vs ASD	CCDC68, NEGR1, TRANK1, AS3MT, SATB2, DRD2
SCZ vs BIP	MAF (not in training set)
SCZ vs MDD	KCNV1(not in training set), IMMP2L, HS3ST5(not in training set), DGKI, CACNB2, PRKD1, CNNM2, SATB2, SNX19, PSD3, PCDHA7(not in training set), CHRNA3, ERBB4
SCZ vs TS	NFIB (not in training set), ZNF536, POU3F2(not in training set)