Supplementary materials for

Zero-shot drug repurposing with geometric deep learning and clinician centered design

Kexin Huang[1,†,*], Payal Chandak[2,*], Qianwen Wang[1], Shreyas Havaldar[3], Akhil Vaid[3,4], Jure Leskovec[5], Girish Nadkarni[4], Benjamin S. Glicksberg[3,4], Nils Gehlenborg[1], and Marinka Zitnik[1,6,7,8,‡]

[1]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115

[2]Harvard-MIT Program in Health Sciences and Technology, Cambridge, MA 02139

[3]Hasso Plattner Institute for Digital Health, Icahn School of Medicine at Mount Sinai, NY 10029

[4]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, NY 10029

[5]Department of Computer Science, Stanford University, Stanford, CA 94305

[6]Broad Institute of MIT and Harvard, Cambridge, MA 02142

[7]Harvard Data Science Initiative, Cambridge, MA 02138

[8]Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Allston, MA 02134

† Present address: Department of Computer Science, Stanford University

* Equal contribution

‡ Corresponding author: marinka@hms.harvard.edu

Supplementary Figures

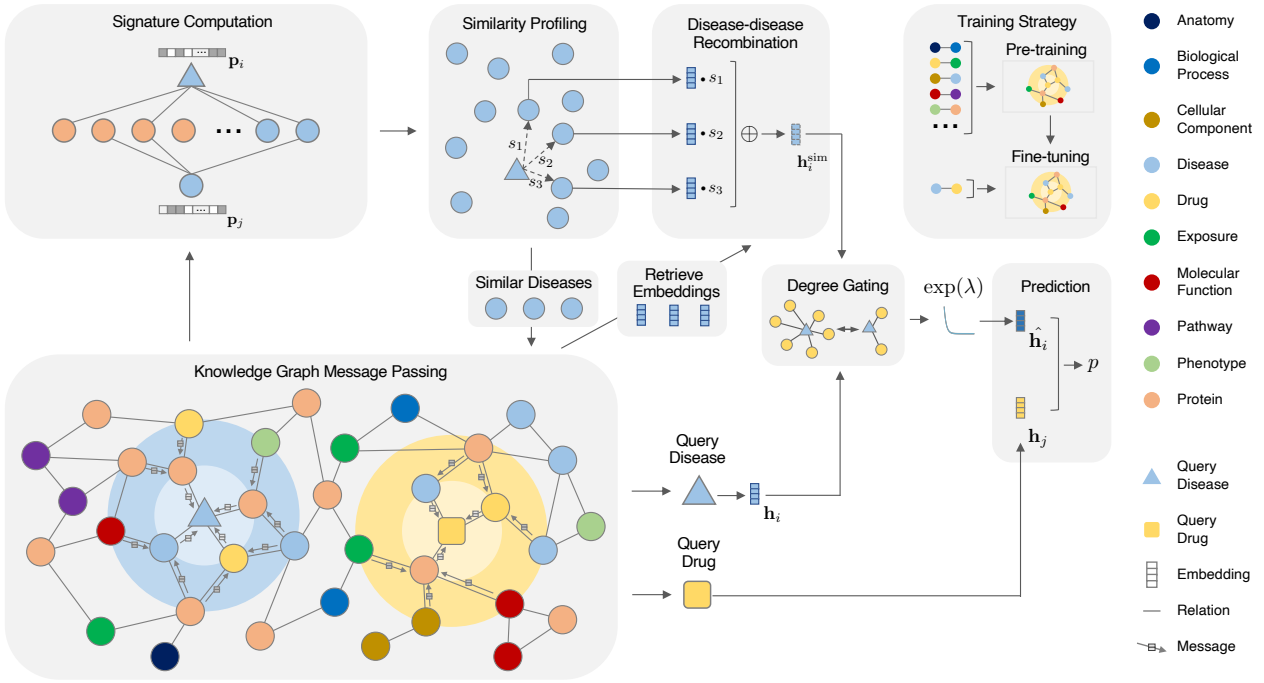

Figure S1: Detailed TxGNN predictor illustration. (1) TxGNN projects biological concepts into meaningful representations through knowledge graph neural network message passing on the KG. (2) It then designs a similarity disease search component to enrich molecularly uncharacterized diseases and it has three modules (2.1) It computes a signature vector for each disease that captures the disease similarity. (2.2) Based on the signature vector distance, it profiles a set of similar diseases and retrieves their latent embeddings. (2.3) It then aggregates the different similar diseases into a powerful auxiliary embedding. (2.4) A gating mechanism is designed to control the effect between the original disease embedding and the auxiliary disease embedding since many well-characterized diseases have sufficient embeddings and do not need subsidies. (3) A decoder then maps the query drug and disease representation to predict the outcome. A pretext learning stage is devised to allow TxGNN to learn an initialized embedding that captures complex biological knowledge.

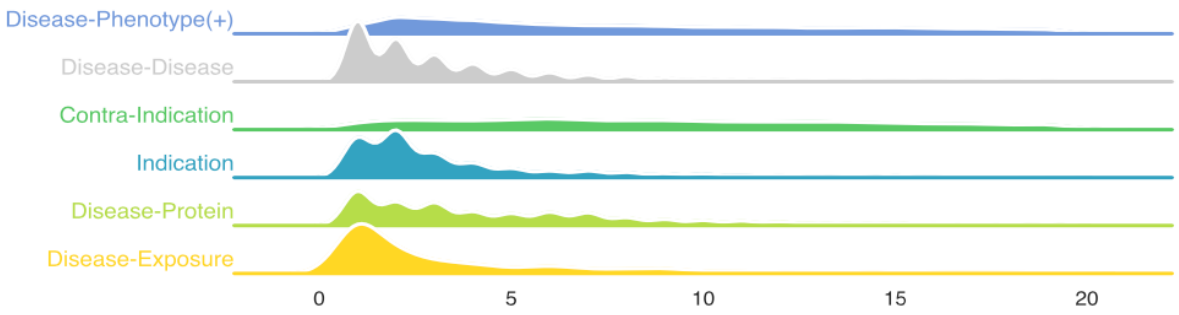

Figure S2: Distribution of the number of disease neighbor types. Our KG has rich information about diseases. On median, for diseases that have known indications, a disease node is connected to 5 proteins, 14 phenotypes, 3 other diseases, and 2 exposures in the network.

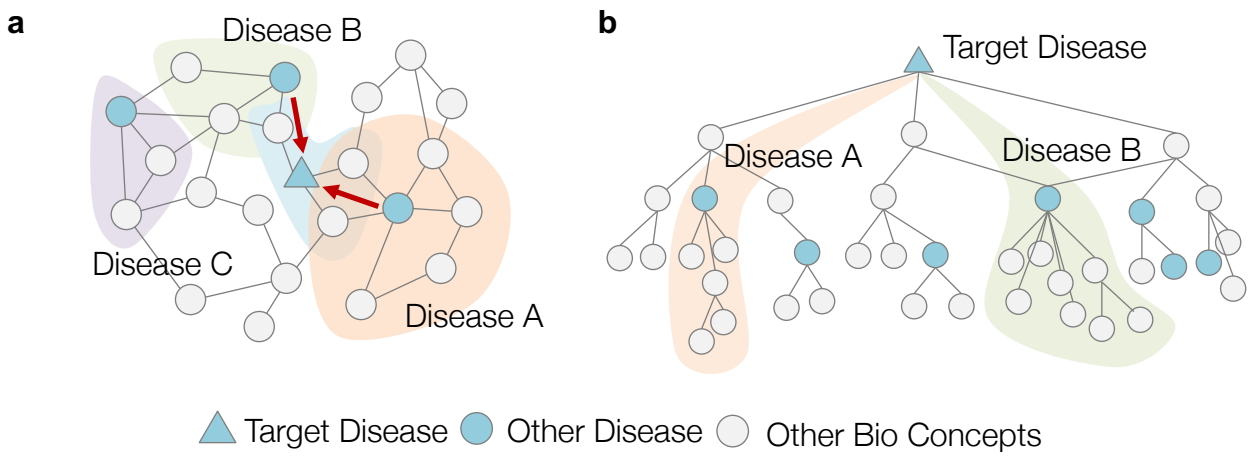

Figure S3: Illustration of TxGNN graph rewiring perspective. a. For a target disease, we use a disease signature to select similar disease nodes in the entire network (in the figure, target disease selects disease A and B, not C). We then want to aggregate the disease module information of these similar disease nodes into the target disease, where this information would not be available with the classic GNN approach. TxGNN fuses this information into target disease embedding by aggregating these remote disease embeddings to the target disease embeddings. After fusing, the disease module information is available in the target disease node. **b.** This fusing step happens in the latent embedding space, which is equivalent to adding a network edge between the target disease node and selected similar diseases. We see that through this network rewiring perspective, TxGNN conducts a long-range selective aggregation guided by domain prior.

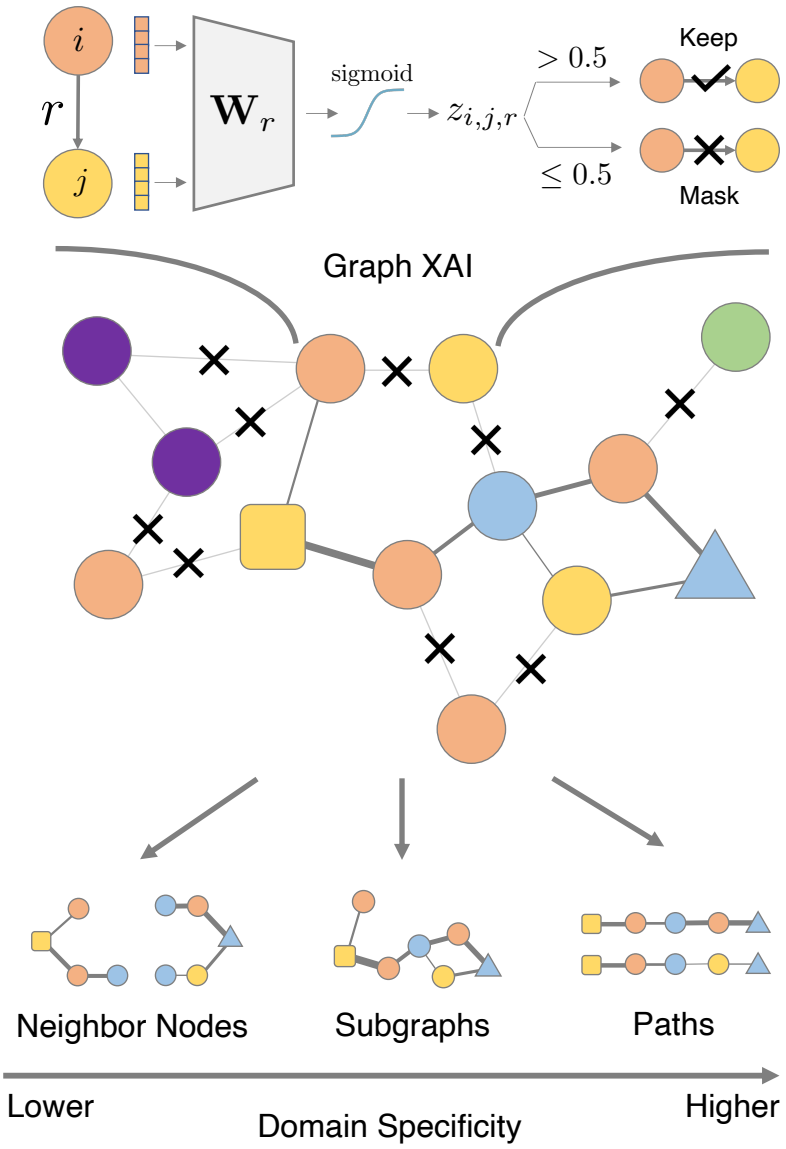

Figure S4: Detailed illustration of TxGNN explainer. For each edge between node i and j with relation r, a weight matrix $\mathbf{W}_r$ takes in the node embedding and produces a score $z_{i,j,r}$. If the score measures the importance of this edge to the prediction. If it is larger than some threshold, *e.g.*, 0.5, then the edge is kept and vice versus. Applying this to every edge, we can obtain a sparse subgraph that depicts the essential connections for predictions. This same subgraph explanation is then converted to various forms of visualizations such as neighbor nodes, subgraphs, and paths. We study and find that path-based visualization aligns the best with domain scientists.

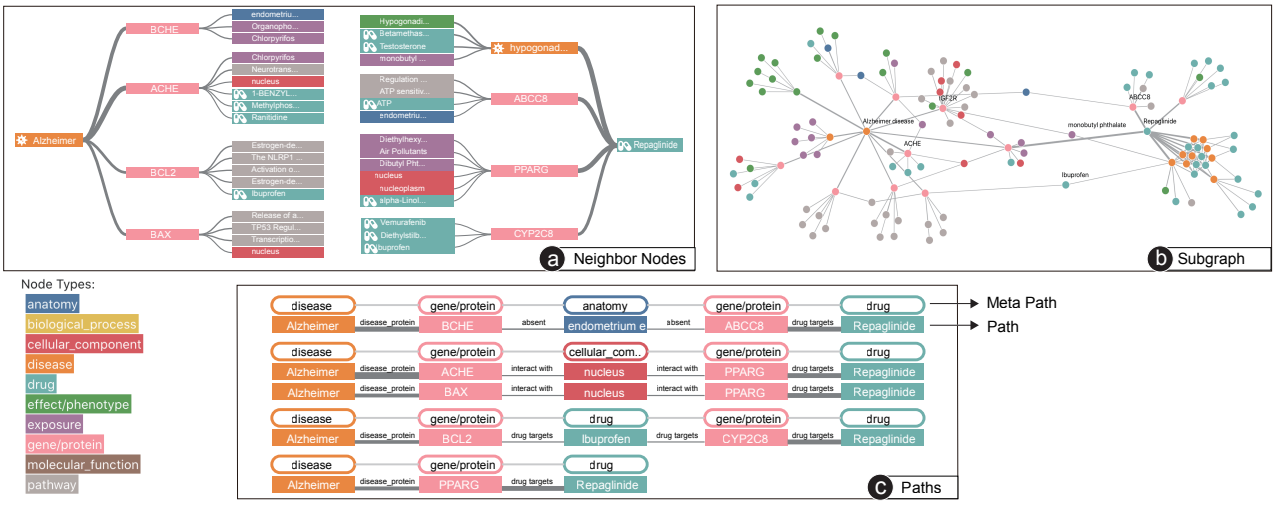

Figure S5: Three types of visual explanations for TxGNN predictions. We categorize instance explanations for GNNs into three main groups, neighbor nodes (a), subgraphs (b), and paths (c). We compared the three visual explanations and selected path-based explanations due to their similarity to the clinicians' reasoning about indications.

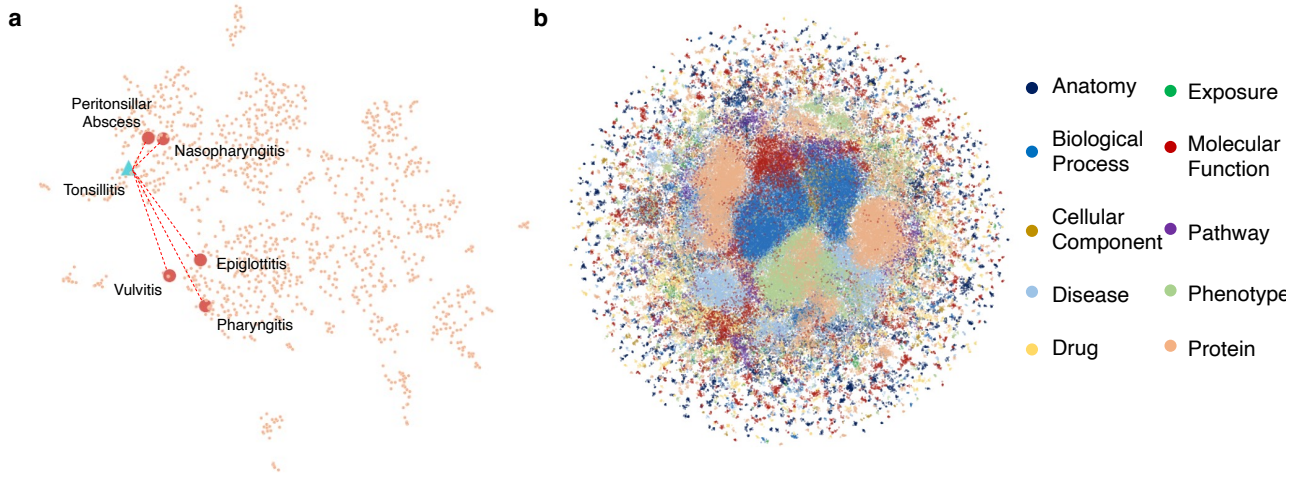

Figure S6: Visualize TxGNN latent representations. a. To further understand the performance gain of the metric learning module from a machine learning standpoint, we explore the example of tonsillitis. Diseases similar to tonsillitis (epiglottitis, peritonsillar abscess, nasopharyngitis, pharyngitis, vulvitis) are initially distant in the embedding space. Thus, by fusing distant disease embeddings, TxGNN establishes a long-range skip connection to the disease module of these similar diseases and provides complementary information missing from the local neighborhood around the target disease. This is especially beneficial in predicting therapeutic use for conditions with few or no treatments and limited molecular understanding. TxGNN uses disease signatures as a learnable disease look-up catalog to identify the appropriate distant disease information that can be transferred to the underpowered target disease. **b.** Visualization of embeddings for all nodes in the KG.

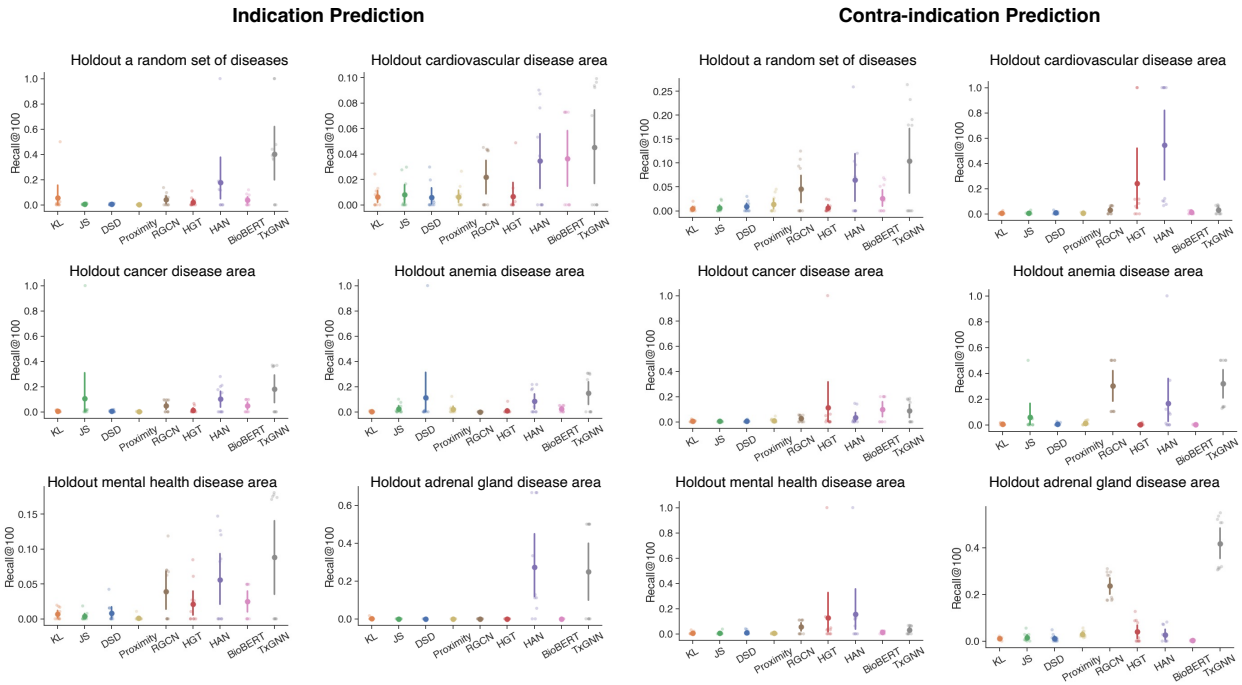

Figure S7: Prioritizing indications and contra-indications with high accuracy. We generate a list of the top-100 predicted drugs for each disease and calculate the fraction of correct hits in the list (Recall@100). TxGNN performs consistently better across hold-out sets.

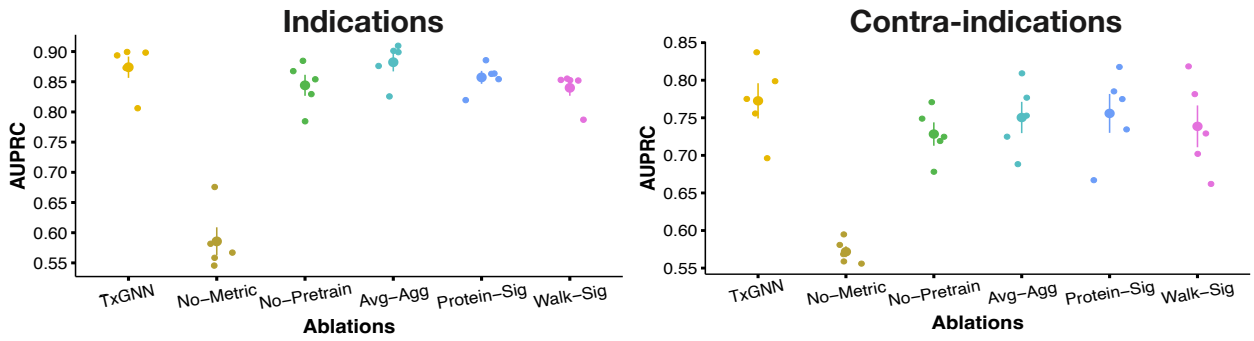

Figure S8: Ablation study. We conduct a systematic study by modifying individual components of TxGNN to test their utility on the systematic disease split. First, we remove the entire metric learning procedure, and it degrades to regular GNN ('No-Metric'). We find TxGNN has a 0.2884 AUPRC increase over the ablation for indication and 0.2008 AUPRC increase for contraindication. Then, we keep the metric learning procedure but remove pretraining ('No-Pretrain'). The ablation has 0.030 decrease in AUPRC and retrieves 7.5% fewer hits in the top 100 predictions for indication. We observe similar behaviors for predicting contraindicated use, where 'No-Pretrain' leads to a 0.044 decrease in AUPRC and recalls 7.7% fewer hits, showing that the biomedical knowledge-grounded pretraining strategy is valuable and leads to positive knowledge transfer. To test the utility of degree-based aggregation, we use a simple alternative by taking the average between the auxiliary and original disease embeddings ('Avg-Agg'). We find TxGNN has relatively similar performances in indication prediction but improves contraindication prediction by 0.022 in AUPRC and retrieves 1.8% more hits, showing the usefulness of this component. Lastly, we experiment with two alternative strategies to calculate the disease signature, one is only using protein nodes to calculate disease similarity ('Protein-Sig'), and another is a diffusion-based random walk signature ('Walk-Sig'). We find TxGNN retrieves 8.4%/5.4% more hits than 'Protein-Sig' and 9.6%/6.4% more hits than 'Walk-Sig' in indication/contraindication prediction, respectively, suggesting the importance of signature selection to characterize the similarity among diseases. We find that each component is indispensable in the success of TxGNN. The deep metric learning module is the key factor that drives TxGNN performance, corroborating our hypothesis on disease similarity.

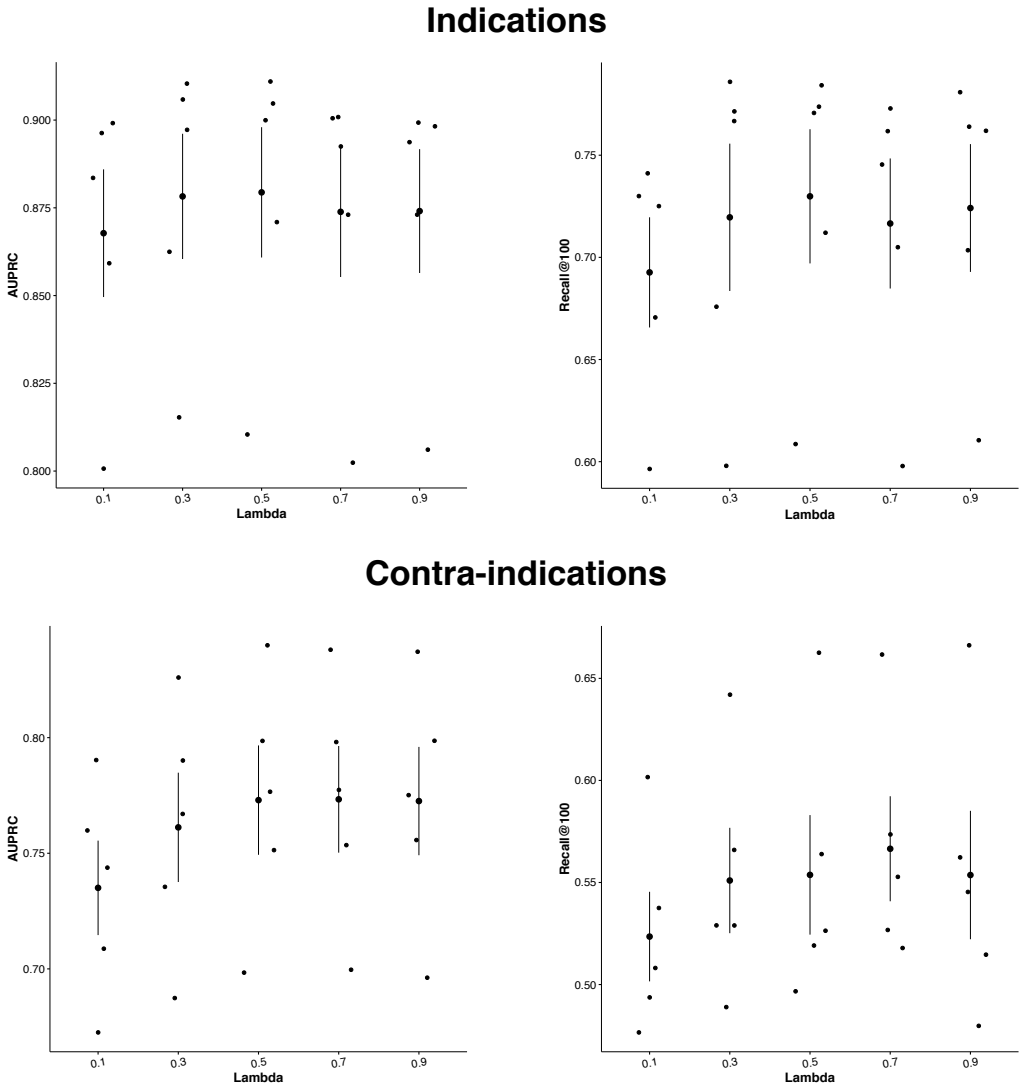

Figure S9: Sensitivity analysis of λ in the embedding gating. When updating the original disease embedding with disease-disease metric learning embedding, we want to give higher weight to the original disease embedding when the disease node degree is large and vice versus. We use an exponential function on the node degree to approximate this effect. Here, we show that the model performance is not sensitive to the exponential function parameter λ where larger λ gives a more steep increase of weight as the node degree becomes smaller. The y-axis on the left side panel is AUPRC and y-axis on the right side panel is Recall@100. The metric is calculated using systematic split with five independent data splits.

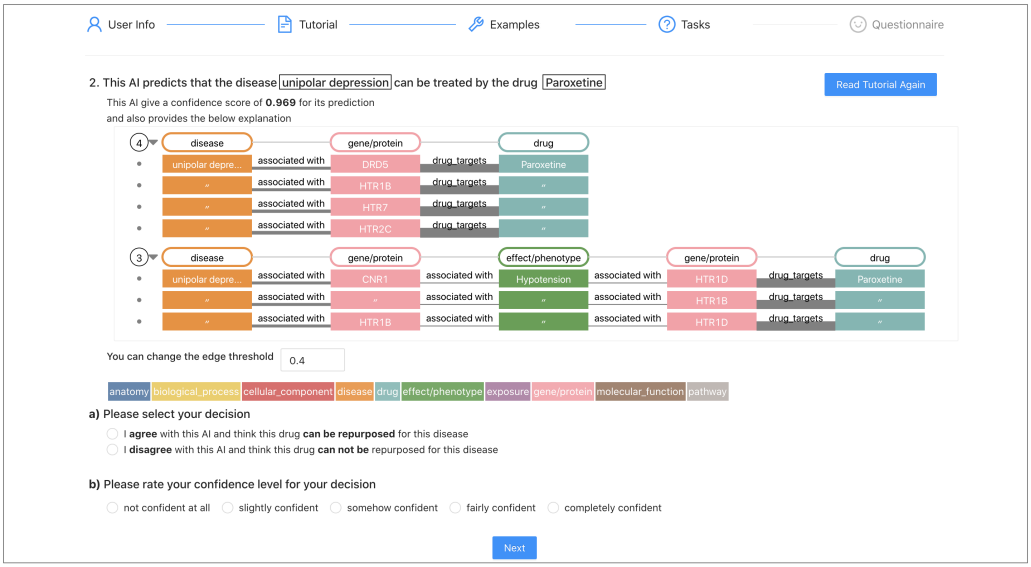

Figure S10: Interface used in the usability study of TxGNN. We compare path-based explanations with a non-explanation baseline. For each prediction, participants decided whether the predicted drug can be used for treating a certain disease and reported their confidence levels using a 5-point Likert scale (1=not confident at all, 5=completely confident).

Supplementary Tables

Drug name	Active ingredient	Disease	Approval	FDA Number	Orphan	TxGNN	Percentile
Vabysmo	Faricimab	Macular degeneration	01/28/2022	BLA761235	No	0.938	2.25%
Welireg	Belzutifan	von Hippel-Lindau disease	08/13/2021	NDA215383	Yes	0.720	4.11%
Mounjaro	Tirzepatide	Type 2 diabetes mellitus	05/13/2022	NDA215866	No	0.286	12.50%
Ztalmy	Ganaxolone	CDKL5 disorder	03/18/2022	NDA215904	Yes	0.335	18.73%
Leqvio	Inclisiran sodium	Familial hypercholesterolemia	12/22/2021	NDA214012	No	0.301	19.32%
Tezspire	Tezepelumab-ekko	Asthma	12/17/2021	BLA761224	No	0.233	32.41%
Vtama	Tapinarof	Psoriasis	05/23/2022	NDA215272	No	0.261	32.70%
Adbry	Tralokinumab	Atopic dermatitis	12/27/2021	BLA761180	No	0.040	50.37%
Vonjo	Pacritinib citrate	Myelofibrosis	02/28/2022	NDA208712	Yes	0.011	63.14%
Livtencity	Maribavir	Cytomegalovirus infection	11/23/2021	NDA215596	Yes	0.033	66.37%

Table S1: Evaluation of TxGNN on recently approved therapies. To demonstrate that TxGNN was not driven by confirmation bias from indications and contraindications already present in the knowledge graph, we considered ten therapies that were approved by the FDA after TxGNN's dataset and model development were completed (June 2021). None of these therapies had direct relationships between their drug-disease nodes in the TxGNN dataset. We then asked TxGNN to make predictions for those diseases without any prompting for the recently approved drug and observed that TxGNN consistently ranked newly introduced drugs highly, with the recently approved drugs found in the first third (30.19%) of the full-length prediction list on average. Occasionally, TxGNN ranked the approved drug in the top 5% of therapeutic candidates, such as faricimab to treat macular degeneration (top 2.25%) and belzutifan to treat von Hippel-Lindau disease (top 4.11%). In one case, TxGNN ranked maribavir in the bottom two-thirds of the prediction list for treating cytomegalovirus infections in patients post-transplantation, likely because TxGNN's knowledge graph did not contain information about the host-pathogen interactions that inform the treatment pathway.

Node Type	Count	Percent (%)
Biological process	28,642	22.1
Protein	27,671	21.4
Disease	17,080	13.2
Phenotype	15,311	11.8
Anatomy	14,035	10.8
Molecular function	11,169	8.6
Drug	7,957	6.2
Cellular component	4,176	3.2
Pathway	2,516	1.9
Exposure	818	0.6
Total number of nodes	129,375	100.0

Table S2: Statistics on nodes in the therapeutics-centered knowledge graph.

Relation	Count	Percent (%)
Anatomy – Protein (present)	3,036,406	37.5
Drug – Drug	2,672,628	33.0
Protein – Protein	642,150	7.9
Disease – Phenotype (positive)	300,634	3.7
Biological process – Protein	289,610	3.6
Cellular component – Protein	166,804	2.1
Disease – Protein	160,822	2.0
Molecular function – Protein	139,060	1.7
Drug – Phenotype	129,568	1.6
Biological process – Biological process	105,772	1.3
Pathway – Protein	85,292	1.1
Disease – Disease	64,388	0.8
Drug – Disease (contraindication)	61,350	0.8
Drug – Protein	51,306	0.6
Anatomy – Protein (absent)	39,774	0.5
Phenotype – Phenotype	37,472	0.5
Anatomy – Anatomy	28,064	0.3
Molecular function – Molecular function	27,148	0.3
Drug – Disease (indication)	18,776	0.2
Cellular component – Cellular component	9,690	0.1
Phenotype – Protein	6,660	0.1
Drug – Disease (off-label use)	5,136	0.1
Pathway – Pathway	5,070	0.1
Exposure – Disease	4,608	0.1
Exposure – Exposure	4,140	0.1
Exposure – Biological process	3,250	$<$0.1
Exposure – Protein	2,424	$<$0.1
Disease – Phenotype (negative)	2,386	$<$0.1
Exposure – Molecular function	90	$<$0.1
Exposure – Cellular component	20	$<$0.1
Total number of edges	8,100,498	100.0

Table S3: Statistics on edges in the therapeutics-centered knowledge graph.