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, MA 02134

† Present address: Department of Computer Science, Stanford University
⋆ Equal contribution
‡ Corresponding author: marinka@hms.harvard.edu

Historically, drug repurposing – identifying new therapeutic uses for approved drugs – has been attributed to serendipity. While recent advances have leveraged knowledge graphs and deep learning to identify potential therapeutic candidates, their clinical utility remains limited because they focus on diseases with available existing treatments and rich molecular knowledge. Here, we introduce TxGNN, a geometric deep learning approach designed for "zero-shot" drug repurposing, identifying therapeutic candidates even for diseases with no existing medicines. Trained on a medical knowledge graph, TxGNN utilizes a graph neural network and metric-learning module to rank therapeutic candidates as potential indications and contraindications across 17,080 diseases. When benchmarked against eight methods, TxGNN significantly improves prediction accuracy for indications by 49.2% and contraindications by 35.1% under stringent zero-shot evaluation. To facilitate interpretation and analysis of the model's predictions, TxGNN's Explainer module offers transparent insights into the multi-hop paths that form TxGNN's predictive rationale. Our pilot human evaluation of TxGNN's Explainer showed that TxGNN's novel predictions and explanations perform encouragingly on multiple axes of model performance beyond accuracy. Many of TxGNN's novel predictions are aligned with off-label prescriptions made by clinicians within a large healthcare system, affirming their potential clinical utility. TxGNN provides drug repurposing predictions that are more accurate than existing methods, are consistent with off-label prescription decisions made by clinicians, and can be investigated by human experts through multi-hop interpretable explanations.

Introduction

There is a pressing need to develop therapies for many diseases that currently lack treatments[1,2]. Of over 7,000 rare diseases worldwide, only 5-7% of rare diseases have FDA-approved drugs[3]. Leveraging existing therapies and expanding their use by identifying new therapeutic indications via drug repurposing can alleviate the global disease burden. By using safety and efficacy data for existing drugs, drug repurposing can expedite translation to the clinic and lower development costs than designing drugs from scratch[4] (Figure 1a). The fundamental premise behind repurposing is that drugs can have pleiotropic effects beyond the mechanism of action of their direct targets[5]. Approximately 30% of FDA-approved drugs are issued at least one post-approval new indication, and many drugs have accrued over ten indications over the years[6]. However, most repurposed drugs are the result of serendipity[7,8] – either observed through off-label prescriptions written by clinicians, as with gabapentin and bupropion[8] or discovered through patient experience, as with sildenafil[6]. The relationships between drug candidates and their potential new applications have not been studied systematically because the underlying mechanism 'connecting' them is often intricate and dispersed through the biomedical literature[7].

Owing to technological advances, the effects of drugs can now be prospectively matched to new indications by systematically analyzing medical knowledge graphs[5,9]. The new strategies rely on identifying therapeutic candidates based on their effects on cell signalling, gene expression, and disease phenotypes[5,10–12]. Machine learning has been used to analyze high-throughput molecular interactomes to unravel genetic architecture perturbed in disease[12,13] and help design therapies to target them[14]. To provide therapeutic predictions, geometric deep learning models optimized on large medical knowledge graphs[15] can match disease signatures to therapeutic candidates based on networks perturbed in disease[15–19].

Although computational approaches have identified promising repurposing candidates for complex diseases[16,20,21], there remain two key challenges that could significantly enhance the clinical relevance of repurposing predictions made by machine learning models. (1) First, existing methods assume that diseases for which we would like to make therapeutic predictions are well-understood and likely to have existing therapies. While this is the case for more widespread diseases[9], a long tail of diseases does not satisfy this assumption – 92% of 17,080 diseases examined in our study have no indications. Moreover, around 95% of rare diseases have no FDA-approved drugs, and up to 85% of rare diseases do not have even one drug developed that would show promise in rare disease treatment, diagnosis, or prevention[22]. This long tail of diseases with

few or no therapies and limited molecular understanding presents a clinically fruitful challenge for drug repurposing models to prioritize. (2) Second, a repurposed indication for a therapeutic candidate can be unrelated to the indication for which the drug was initially studied. Originally proposed to help with morning sickness during pregnancy, Thalidomide was repurposed in 1964 for an autoimmune complication of leprosy and again in 2006 for multiple myeloma[8]. Collectively, we refer to these challenges as the zero-shot drug repurposing problem (Figure 1b). To be clinically useful, machine learning models must make "zero-shot" predictions; that is, they need to extend therapeutic predictions to diseases whose understanding is incomplete and, further, to diseases with no approved drugs. Unfortunately, the ability of existing machine learning models to identify therapeutic candidates for diseases with incomplete, sparse data and zero known therapies drops drastically[16,23] (as we demonstrate across eight benchmarks in Figures 2c and 2d).

Here, we introduce TxGNN, a geometric deep learning approach for zero-shot drug repurposing that can prioritize therapeutic candidates for diseases with no therapies (Figure 1c). Foundation models like TxGNN are transforming deep learning: instead of training disease-specific models for every disease, TxGNN is a single pretrained model that adapts across many diseases. TxGNN is trained on a medical knowledge graph that collates decades of biological research across 17,080 diseases (Figure 1d). TxGNN uses a graph neural network model to embed therapeutic candidates and diseases into a latent representation space and is optimized to reflect the geometry of TxGNN's medical knowledge graph. To make therapeutic predictions under zero-shot settings, TxGNN implements a metric learning module to learn similarities between diseases with indications and diseases without indications to transfer knowledge between these diseases and make zero-shot predictions. Once trained, TxGNN performs zero-shot inference on new diseases without additional parameters or fine-tuning. To facilitate interpretation and analysis of the therapeutic candidates that TxGNN ranks highly, we develop a TxGNN Explainer module that offers transparent insights into the multi-hop pathways that form TxGNN's predictive rationale. TxGNN's predictions and explanations are available at http://txgnn.org. Our pilot human evaluation of TxGNN's Explainer showed that TxGNN's explanations perform encouragingly on multiple axes of model performance such as accuracy, trust, usefulness, and time efficiency (Figure 4). Moreover, many of TxGNN's novel predictions have shown alignment with off-label prescriptions made by clinicians within a large healthcare system and TxGNN's explanatory rationales have demonstrated consistency with medical reasoning in selected case studies, encouraging the potential real-world clinical utility of TxGNN.

Results

Overview of TxGNN zero-shot drug repurposing model. A problem not previously considered in biomedical deep learning research, zero-shot drug repurposing involves predicting therapeutic candidates for diseases that do not have any existing indications (Figure 1b). Mathematically, the model takes a query drug-disease pair as input and provides the likelihood of the drug acting on the disease as output. The gold standard labels for evaluating such a model come from our previously curated and validated a large-scale medical knowledge graph[9] (Figure 1d, Tables S2 and S3) that consists of 9,388 indications and 30,675 contraindications[24]. The knowledge graph covers a vast range of 17,080 diseases where 92% have no FDA-approved drugs, including rare diseases and less-understood complex diseases. The knowledge graph also comprises 7,957 potential candidates for drug repurposing, ranging from FDA-approved drugs to experimental drugs investigated in ongoing clinical trials. Our model for zero-shot drug repurposing, TxGNN operates on the principle that effective drugs can target disease-perturbed and disease-associated networks of biomolecules, and it has two modules: (1) the TxGNN *Predictor* module enables the accurate prediction of indications and contraindications in the zero-shot setting and (2) the TxGNN *Explainer* module provides interpretable multi-hop pathways that connect the drug to the disease (Figure 1c).

TxGNN Predictor The Predictor module consists of a graph neural network (GNN) optimized on the relationships within the biomedical knowledge graph (Methods 2.2). Through large-scale self-supervised pre-training, the GNN produces biologically meaningful representations for any entity in this knowledge graph. Then, this GNN is finetuned to predict relationships between therapeutic candidates and diseases. TxGNN leverages metric learning for zero-shot prediction. TxGNN capitalizes on the insight that diseases are intrinsically related[10,14] by leveraging molecular mechanisms of well-annotated diseases to enhance predictions on diseases with limited annotations (Figure 2a, Figure S1). This is achieved by creating a disease signature vector for each disease based on its neighbors in the knowledge graph. The similarity between a pair of diseases is measured by the normalized dot product of their signature vectors. Since most disease pairs do not share underlying pathologies, they have low similarity scores. In contrast, a relatively high similarity score (>0.2) between diseases suggests similar mechanisms (Figure 2b). A detailed description of the model and its architecture can be found in Methods 2 and Figure S2.

When querying a specific disease, TxGNN retrieves similar diseases, generates embeddings for them, and then adaptively aggregates them based on their similarity to the queried disease. The aggregated output embedding summarizes knowledge borrowed from similar diseases fused with

the query disease embedding. This step can also be interpreted as a graph rewiring technique in the geometric machine learning literature (Figure S3). TxGNN processes different downstream therapeutic tasks, such as indication and contraindication prediction, in a unified manner using shared drug and disease embeddings (Methods 2.3). Given a query disease, TxGNN ranks drugs based on their predicted likelihood scores, offering a prioritized list of therapeutic candidates with potential for repurposing.

TxGNN Explainer While TxGNN Predictor provides likelihood scores for therapeutic candidates, these scores alone are insufficient for trustworthy model deployment. Clinicians and scientists seek to understand the reasoning behind these predictions to validate the model's hypotheses and better understand disease pathology. To this end, TxGNN Explainer delves into the knowledge graph to pinpoint and succinctly present relevant biological pathways for the drug-disease pair of interest (Figure 4a). This conceptual subgraph mirrors the analytical process clinical researchers use to examine relationships between therapeutic candidates and disease and how the drug perturbs local biological networks to produce a therapeutic effect on disease.

TxGNN uses a self-explaining approach called GraphMask[25] (Methods 2.6). For a particular therapeutic use prediction, GraphMask generates a sparse yet sufficient subgraph of biological entities considered critical by TxGNN for making the prediction. Particularly, it yields an importance score between 0 and 1 for every edge in the subgraph between the drug and disease, with 1 indicating the edge is vital for prediction and 0 suggesting it is irrelevant. TxGNN Explainer combines the drug-disease subgraph and edge importance scores to produce multi-hop explanations connecting the disease to the predicted therapeutic candidate. Unlike widely recognized explainability techniques such as SHAP[26] that generate feature attribution maps, TxGNN Explainer offers granular and straightforward explanations that are, as we show in a pilot human study, aligned with clinician/scientist's intuition.

We developed a human-centered graphical user interface that presents these subgraph explanations proposed by TxGNN Explainer (Figure 4b). Amongst a range of designs, as shown in Figures S4 and S5, we focused on visual path-based reasoning because our pilot human study demonstrated that this design choice enhanced clinician comprehension and satisfaction[27]. This interface with TxGNN's predictions and explanations is openly accessible at http://txgnn.org.

Comparative assessment of TxGNN against existing methods. We evaluated model performance in drug repurposing across various hold-out datasets. We generated a hold-out dataset by sampling diseases from the knowledge graph. These diseases were deliberately omitted during the

training phase and later served as test cases to gauge the model's ability to generalize its insights to previously unseen diseases. These held-out diseases were either chosen randomly, following a standard evaluation strategy, or specifically selected to evaluate zero-shot prediction. In our study, we used both types of hold-out datasets to thoroughly evaluate methods. We compared TxGNN to eight established methods in predicting therapeutic use. They included network medicine statistical techniques, including KL and JS divergence[16], graph-theoretic network proximity approach[20], and diffusion state distance (DSD)[28], state-of-the-art graph neural network methods, including relational graph convolutional networks (RGCN)[19,29], heterogeneous graph transformer (HGT)[30], and heterogeneous attention networks (HAN)[31], and a natural language processing model, BioBERT[32] (Supplementary Note S4).

Initially, we followed the standard evaluation strategy where drug-disease pairs were randomly shuffled, and a subset of these pairs was set aside as a hold-out set (testing set; Figure 2c). Under this strategy, the diseases being evaluated as hold-outs may already have had indications and contraindication relationships with drugs in the training set. Therefore, the learning objective was to identify additional therapeutic candidates for well-studied diseases. This evaluation method aligns with the approach predominantly used in literature[19]. We use the area under the precision-recall curve (AUPRC) as the evaluation metric as it measures the recall and precision tradeoff of a model at different thresholds. Our experimental results in this setting concur, with 3 of 8 existing methods achieving AUPRC greater than 0.80, and HAN as the best at 0.873 AUPRC. TxGNN also had a comparable performance as established methods. In predicting indications, TxGNN achieved a 4.3% increase in AUPRC (0.913) over the strongest baseline, HAN.

As shown by the above experiments, machine learning methods can help identify repurposing opportunities for diseases that already have some FDA-approved drugs[12–16,20,21]. However, Duran et al.[33] reason that many methods simply retrieve additional therapeutic candidates that are similar to existing ones across biological levels. This suggests the standard evaluation strategy is unsuitable for evaluating diseases that have no FDA-approved drugs (Figure 1b). Given this limitation, we evaluate models under zero-shot drug repurposing. We began by holding out a random set of diseases and then moved all their associated drugs to the hold-out set (Figure 2d). From a biological standpoint, the model was required to predict therapeutic candidates for diseases that lacked treatments, meaning it had to operate without any available data on drug similarities. In this scenario, TxGNN outperformed all existing methods by a large margin. TxGNN significantly improves over the next best baseline in predicting both indications (19.0% AUPRC gain) and con-

traindications (23.9% AUPRC gain). While established methods achieved satisfactory results in conventional drug repurposing evaluations, they often fell short on more challenging zero-shot drug repurposing scenarios. TxGNN was the only method that achieved consistent performance in both settings.

TxGNN's zero-shot drug repurposing performance across disease areas. Diseases with biological similarities often share therapeutic candidates[10]. For instance, beta-blockers are effective in treating a multitude of cardiovascular issues, including heart failure, cardiac arrest, and hypertension. Likewise, selective serotonin reuptake inhibitors (SSRIs) can address various psychiatric conditions such as major depressive disorder, anxiety disorder, and obsessive-compulsive disorder. If, during training, a model learns that an SSRI is indicated for major depressive disorder, it does not take a large leap to suggest that the same SSRI could be effective for obsessive-compulsive disorder during testing[23]. This phenomenon is known as shortcut learning[34,35] and underlies many of deep learning's failures[36,37]. Shortcut decision rules tend to perform well on standard benchmarks but typically fail to transfer to challenging testing conditions[38], such as the real-world scenario of predicting therapeutic candidates for rare or neglected diseases.

To evaluate drug repurposing models for these challenging diseases, we curated a stringent hold-out dataset that contained a group of biologically related diseases that we refer to as a disease area. Given the diseases in a specific disease area, all their indications and contraindications were removed from the training dataset. Further, a fraction of the connections from medical entities to these diseases were excluded from the training dataset. For diseases in the chosen area, these conditions simulated limited molecular characterization and lack of existing treatments (Figure 3a). Under this setup, we observe that diseases in the hold-out evaluation set have a significantly smaller number of neighbors compared to the training set (Figure S6). In this study, we considered nine disease area hold-out datasets characterized in Table 1 and listed here in order of increasing disease area size: (1) diabetes-related diseases such as Gestational diabetes and Lipoatrophic diabetes; (2) 'adrenal gland' diseases like Addison and ectopic crushing syndrome; (3) 'autoimmune' diseases like Celiac disease and Graves disease; (4) 'anemia' with conditions such as thalassemia and hemoglobin C disease; (5) 'neurodegenerative' diseases include pick disease and Neuroferritinopathy; (6) 'mental health' disorders like anorexia nervosa and depressive disorder; (7) 'metabolic disorder' such as Macroglobulinemia and Gilbert syndrome; (8) 'cardiovascular' diseases, including long QT syndrome and mitral valve stenosis; (9) 'cancerous' diseases such as neurofibroma and Leydig cell tumors. These cover a wide range of diverse disease areas.

We benchmarked the performance of TxGNN and all methods above on these rigorous hold-out datasets in Figure 3b-f and S7. We found that TxGNN consistently improved predictive performance over existing methods. For indications, TxGNN had 26.1%, 59.3%, 32.2%, 42.3%, 13.6%, 36.2%, 11.1%, 10.2%, 0.5% relative gain in AUPRC over the next best baseline across diabetes, adrenal glands, autoimmune, anemia, neurodegenerative, mental health, metabolic disorder, cancer, and cardiovascular disease hold-outs respectively. For contraindications, TxGNN robustly improved over the next best baseline, with relative gains ranging from 11.8% to 35.6%. For indication prediction, the natural language processing method, BioBERT, had the best performance (in 7/9 disease area hold-outs) amongst the group of established methods. For contraindication prediction, the graph-based method, RGCN, was the best baseline across 8 of 9 hold-out datasets, and BioBERT's performance gain observed for indication prediction disappeared. TxGNN was consistently the best-performing method across all nine disease area hold-outs for both indication and contraindication prediction tasks. These rigorous benchmarks demonstrate that TxGNN was broadly generalizable and produced accurate predictions in zero-shot drug repurposing settings.

TxGNN demonstrated higher performance in eight of nine disease area hold-outs; however, its performance was equivalent to existing methods in the cardiovascular hold-out. This equivalence may be due to an absence of related disease knowledge in the training dataset when entire disease areas are excluded. Visualization of the latent representations of TxGNN Predictor revealed that it supports knowledge transfer from unrelated diseases to those with limited information (Figure S8). Additional evaluation metrics, including AUROC and recall, are detailed in Figures S9, S10, and S11. Ablation analyses confirmed that each component of TxGNN Predictor is critical for the model's predictive performance (Figure S12). Additional data splits were conducted to stress test the model, including evaluations on diseases with minimal connections to the knowledge graph (Figure S13), evaluations with certain percentages of disease local neighborhood masked (Figure S14), and evaluations on various knowledge graph configurations (Figure S15). These evaluations showed that TxGNN maintains robust and strong predictive performance.

TxGNN's multi-hop explanations reflect model's predictive rationale. TxGNN's Explainer extracts multi-hop explanations as sequences of associations between predicted drugs and diseases in the knowledge graph to substantiate TxGNN's predictions. This tool identifies maximally predictive subgraphs within the knowledge graph, connecting the query drug to the query disease through multiple hops, following relationships in the graph. The performance of these subgraphs is nearly equivalent to that of the entire knowledge graph. To assess the quality of ex-

planations, we first compared the AUPRC of TxGNN's predictions using the entire knowledge graph against the AUPRC derived from only the predictive subgraphs. A strong correlation indicates that TxGNN's Explainer effectively identifies key associations[39] and that explanations accurately reflect TxGNN's internal reasoning[40]. Focusing on the most predictive relationships (i.e., edges with importance scores above 0.5, representing an average of 14.9% of edges from the knowledge graph), the model's performance showed a slight reduction from AUPRC=0.890 (STD: 0.006) to AUPRC=0.886 (STD: 0.005). Conversely, when excluding edges deemed predictive by TxGNN and considering the remaining irrelevant relationships (i.e., edges with importance scores below 0.5, accounting for an average of 85.1% of edges), the predictive performance significantly dropped from AUPRC=0.890 (STD: 0.006) to AUPRC=0.628 (STD: 0.026).

To assess the quality of TxGNN's explanations, we employed three established metrics: (1) insertion, which measures predictive performance using only the top K% of edges ranked highest by explanation weight; (2) deletion, which assesses performance after removing the top K% of edges considered most explainable; (3) stability, which evaluates the consistency of explanation weights through Pearson's correlation before and after introducing random perturbations to the knowledge graph. We included experiments with three graph explainability methods: GNNExplainer[41], Integrated Gradients[42], and Information Bottleneck[43]. As shown in Figure S16, the top-ranked explainable edges are crucial, significantly impacting performance when either removed from or inserted into a graph. The performance remained consistent across all insertion and deletion percentages. Additionally, TxGNN Explainer demonstrated the most stable explanation weights under various levels of knowledge graph perturbation. These analyses confirm that TxGNN's multi-hop explanations capture elements of the knowledge graph most critical for making accurate predictions.

TxGNN Explainer supports the human-centric evaluation of therapeutic candidates. To examine the utility of TxGNN's multi-hop interpretable explanations for human expert evaluations, we conducted a pilot human study with clinicians and scientists (see Figure S17 for the study interface). The study participants included five clinicians, five clinical researchers, and two pharmacists (7 males, 5 females, mean age=34.3, Figure 4c). The user study took around 65 minutes in average, including the assessment of drug-disease indication predictions from TxGNN, a usability questionnaire, and a semi-structured interview. For assessing drug-disease indication predictions, these participants were asked to assess 16 predictions from TxGNN, 12 of which were accurate. For each prediction, we recorded participants' assessment accuracy, exploration time, and confidence

scores, totaling 192 trials (16 predictions × 12 participants).

In evaluating the drug repurposing candidates, participants reported a significant improvement in both accuracy (+46%, $p = 0.0443 < 0.05$) and confidence (+49%, $p = 0.0041 < 0.05$) when provided with explanations. Participants took more time to think ($p = 0.0014$) to contextualize TxGNN's explanations with their domain expertise, which led to more confident decisions (confidence +49%, $p = 0.0041 < 0.05$). When using TxGNN Explainer, participants are more accurate in evaluating the correctness of drug repurposing predictions than using TxGNN predictions alone (accuracy +46%, $p = 0.0443 < 0.05$; Tables S6 and S7).

In the post-task questionnaires and interviews, participants reported greater satisfaction when using TxGNN Explainer compared to the baseline (Figure 4e), with 11/12 (91.6%) agreeing or strongly agreeing that the predictions and explanations provided by TxGNN were valuable. In contrast, without explanations, 8/12 (75.0%) disagreed or strongly disagreed with relying on TxGNN's predictions. Participants expressed significantly more confidence in correct predictions made by TxGNN when the TxGNN Explainer was included ($t(11) = 3.64, p < 0.01$, using a two-sided Tukey's honestly significant difference test[44]). Some participants indicated that multi-hop interpretable explanations were helpful when examining molecular target interactions identified by TxGNN Explainer and guiding evaluations of potential adverse drug events.

Alignment between TxGNN's drug repurposing predictions and medical evidence. For three rare diseases, we investigated whether predicted drugs and their multi-hop explanations align with medical reasoning. The evaluation protocol was structured into three stages (Figure 5a). Initially, a human expert queried TxGNN Predictor to identify drugs potentially repurposable for a specific disease. The TxGNN Predictor provided a candidate drug, specifying the confidence in the prediction and its comparative ranking against other candidates. Subsequently, the TxGNN Explainer was queried to elucidate why the selected drug was considered for repurposing. This model revealed its rationale through multi-hop interpretable paths linking the disease to the drug via intermediate biological interactions. In the final stage, independent medical evidence was collected and analyzed to verify the model's predictions and explanations.

First, we examined TxGNN's predictions for Kleefstra syndrome, a disease with a prevalence of less than one in a million. The condition is attributed to mutations in the EHMT1 gene, leading to pronounced speech development delays, autism spectrum disorder, and childhood hypotonia. Kleefstra syndrome often features underdeveloped brains with many dormant neuronal pathways. On querying TxGNN Predictor, it recommended Zolpidem as the number one drug

repurposing candidate (Figure 5b). At first, this seemed like it would worsen the underdeveloped brains since Zolpidem is commonly used as a sedative and has an inhibitory effect on GABA-A receptors (gene GABRG2) in the brain. TxGNN Explainer's pathways proposed that Zolpidem's action on GABRG2 could reduce autism susceptibility and enhance prefrontal cortex functioning. Surprisingly, we found that Zolpidem has also demonstrated unexpected stimulative effects in various neurological conditions. For various neurodevelopmental disorders, Zolpidem has been observed to temporarily awaken underactive neurons, offering a potential therapeutic avenue[45]. This paradoxical improvement in neuronal activity can lead to enhancements in speech, motor skills, and alertness in individuals with severe brain injuries or neurodevelopmental disorders, as supported by anecdotal evidence and a handful of clinical studies[46,47]. TxGNN 's prediction and explanatory rationale are both aligned with medical evidence about the paradoxical mechanism of action for Zolpidem, despite none of these clinical cases being directly encountered by the model during training.

Next, we explored TxGNN's prediction of Tretinoin for Ehlers-Danlos syndrome, a rare connective tissue disorder that affects 1-9 individuals per 100,000. This disorder arises from mutations in collagen-coding genes (such as COL1A1 and COL1A2) and is marked by impaired wound healing and the development of atypical scars. TxGNN Predictor ranks Tretinoin as the number one drug repurposing candidate for Ehlers-Danlos syndrome. Tretinoin, a vitamin A derivative commonly used for acne treatment, is transported by albumin (ALB) and targets ALDH1A2 to mitigate collagen loss and inflammation. Both of these members of Tretinoin's mechanism of action occur in TxGNN 's predictive rationale for this prediction (seen in Figure 5c), indicating that TxGNN 's predictive rationale is aligned with medical reasoning. Tretinoin may help in Ehlers-Danlos syndrome by potentially enhancing wound healing and improving the appearance of scars due to its ability to stimulate collagen production in the skin. Further, some subtypes of Ehlers-Danlos syndrome have been associated with a pathogenic mutation in the ALB gene in Landrum et al.[48] and weakly linked to ALDH1AI in Javed et al.[49]. In this case, TxGNN Explainer's reasoning about the pathways that connect Tretinoin to Ehlers-Danlos syndrome was congruent with contemporary clinical evidence.

In the final example, we looked at a rare condition, nephrogenic syndrome of inappropriate antidiuresis (NSIAD). This disease is characterized by water and sodium imbalance caused by a mutation in the AVPR2 gene. Patients with congestive heart failure face similar fluid retention challenges, and congestive heart failure has been strongly associated with both AVPR2 and NPR1

genes[50–52]. TxGNN Predictor identified Amyl Nitrite among the top 5 therapeutic candidates (Figure 5d). TxGNN Explainer proposed that the relationship between NSIAD and Amyl Nitrite passes through AVPR2, congestive heart failure, and NPR1. As per medical literature, the AVPR2 and NPR1 genes play pivotal roles in regulating fluid and electrolyte balance via complementary but distinct pathways. AVPR2 contributes to water retention and urine concentration, whereas NPR1 facilitates vasodilation, lowers blood pressure, and enhances water excretion[53]. Enhancing NPR1 activity could counteract the excessive water reabsorption caused by the malfunctioning AVPR2 receptors in NSIAD patients. Amyl Nitrite, which targets the NPR1 gene, emerges as a potential therapeutic option for NSIAD, confirming consistency of TxGNN's explanations with medical evidence. We share TxGNN drug repurposing predictions and explanations for 17,080 diseases at http://txgnn.org.

Evaluation of TxGNN's predictions using medical records from a large healthcare system. TxGNN's remarkable performance in previous evaluations suggests that its novel predictions—*i.e.*, therapies not yet FDA-approved for a disease but ranked highly by TxGNN —may hold significant clinical value. As these therapies have not yet been approved for treatment, there is no established gold standard against which to validate them. Recognizing the longstanding clinical practice of off-label drug prescription, we used the enrichment of disease-drug pair co-occurrence in a health system's electronic health records as a proxy measure of being a potential indication. From the Mount Sinai Health System medical records, we curated a cohort of 1,272,085 adults with at least one drug prescription and one diagnosis each (Figure 6a). This cohort was 40.1 percent male, and the average age was 48.6 years (STD: 18.6 years). The demographic breakdown is in Figure 6b-c. Diseases were included if at least one patient was diagnosed with it, and drugs were included if prescribed to a minimum of ten patients (Table 2 and Methods 4), resulting in a broad spectrum of 480 diseases and 1,290 drugs as illustrated in Figure 6d.

Across these medical records, we measured disease-drug co-occurrence enrichment as the ratio of the odds of using a specific drug for a disease to the odds of using it for other diseases. We derived 619,200 log-odds ratios (log-ORs) for each drug-disease pair. We found that FDA-approved drug-disease pairs exhibited significantly higher log-ORs than other pairs (Figure 6e). Contraindications represented a potential confounding factor in this analysis because adverse drug events could increase the co-occurrence between drug-disease pairs. However, in our study of contraindications, we found no significant enrichment in the co-occurrence of drug-disease pairs, which suggested that adverse drug effects were not a major confounding factor.

For each disease in the electronic health records, TxGNN produced a ranked list of potential therapeutic candidates. We omitted drugs already linked to the disease, categorized the remaining novel candidates into top-1, top-5, top-5%, and bottom-50%, and calculated their respective mean log-ORs (Figure 6f). We found that the top-1 novel TxGNN prediction had, on average, a 107% higher log-OR than the mean log-OR of the bottom-50% predictions. This suggested that TxGNN's top candidate had much higher enrichment in the medical records and, thereby, had a greater likelihood of being an appropriate indication. In addition, the log-OR increased as we broadened the fraction of retrieved candidates, suggesting that TxGNN's prediction scores were meaningful in capturing the likelihood of indication. Although the average log-OR stands at 1.09, the top-1 therapeutic candidate predicted by TxGNN had a log-OR of 2.26, approaching the average log-OR of 2.92 for FDA-approved indications, indicating the enrichment of off-label drug prescriptions among TxGNN's top-ranked predictions.

Examining TxGNN's predicted drugs for Wilson's disease, a rare disease causing excessive copper accumulation that frequently instigates liver cirrhosis in children (Figure 3g), we observed that TxGNN predicts likelihoods close to zero for most drugs, with only a select few drugs highly likely to be indications. TxGNN ranked Deferasirox as the most promising candidate for Wilson's disease. Wilson's disease and Deferasirox had a log-OR of 5.26 in the medical records, and literature indicates that Deferasirox may effectively eliminate hepatic iron[54]. In a separate analysis, we evaluated TxGNN on ten recent FDA approvals introduced after the knowledge cutoff date (Table S1). TxGNN consistently ranked newly introduced drugs favorably and, in two instances, placed the newly approved drugs within the top 5% of predicted drugs.

Discussion

Drug repurposing has been embraced as a drug discovery approach to address the major productivity issues of cost, time to market, and the inherent risks of developing entirely new drugs. While the conventional 'one disease–one model' approach has been utilized in drug repurposing efforts to enhance success rates, the majority of successful drug repurposing cases have resulted from unexpected findings in clinical and preclinical in vivo settings. We propose that a comprehensive way to reposition drugs is to find new indications through multi-disease predictive models. Yet, existing predictive models are based on the assumption that, for a disease, some drugs already exist for it or that drugs already exist for closely related diseases. This overlooks the vast array of diseases—92% of the 17,080 diseases we analyzed—lacking such pre-existing indications and

known molecular target interactions. Addressing the needs of these diseases, many of which are complex, neglected, or rare, is a top clinical priority[55–57]. We define this challenge as zero-shot drug repurposing.

We introduce TxGNN, a geometric deep learning model that addresses this problem head-on, specifically targeting diseases with limited molecular understanding and no treatment avenues. TxGNN achieves state-of-the-art performance in drug repurposing by leveraging a network medicine principle that focuses on disease-treatment mechanisms[15]. When asked to suggest therapeutic candidates for a disease, TxGNN identifies diseases with shared pathways, phenotypes, and pathologies, extracts relevant knowledge, and fuses it back into the disease of interest. By effectively capturing these latent relationships between diseases, TxGNN can generalize to diseases with few treatment options and perform zero-shot inference for unseen diseases. The design behind TxGNN that enables effective zero-shot drug repurposing can be adapted to a wide range of problems, such as disease-target identification and phenotype modeling.

TxGNN Predictor is a unified model for indication and contraindication prediction across 17,080 diseases. It satisfies an early drug repurposing approach as a high-capacity model that is not limited to a single therapeutic area. Our findings suggest that evaluating a large number of approved or development-stage drugs through multi-disease predictive models should yield a larger number of repositioned drug candidates than approaches limited to a single therapeutic area that can produce infrequent hits. It was found that predicted drug candidates are consistent with off-label prescription rates in a large healthcare system. In the limited evaluation using clinical prescription data and human expert assessment, it was found that predicted drugs were aligned with scientific and clinical consensus. While these estimates suggest beneficial therapeutic potential for existing drugs, predicted drugs would need to undergo extensive screening to establish safety and efficacy as well as determine other drug parameters, such as drug dosage and the sequence and timing of treatments.

TxGNN Explainer generates multi-hop interpretable explanations, offering rationales for predicted drugs. These rationales can be analyzed to assess if predicted drugs might elicit additional biological responses, considering the original indication or molecular target interactions identified by TxGNN Explainer. A pilot human evaluation showed that experts could examine predicted drugs and identify failure points more effectively with multi-hop explanations compared to alternative explanation visualizations. These findings confirm the importance of considering clinical needs and explainability when integrating machine learning models into discovery workflows[58].

While TxGNN demonstrates promising performance for zero-shot drug repurposing, its capabilities depend on the quality of medical knowledge graphs. These graphs may lack comprehensive data on host-pathogen interactions, essential for predicting drug repurposing in infectious diseases (Table S1), and information on the pathogenicity of genetic variants, crucial for identifying repurposing opportunities for genetic diseases[59]. Additionally, challenges such as data biases and the potential for outdated information within the knowledge graph must be addressed. Strategies for overcoming these issues include using techniques for continual learning and model editing[60], and utilizing easily updatable knowledge graphs, as the one used in this study[9]. Another fruitful future direction is using uncertainty quantification techniques to evaluate the reliability of model predictions[61]. We also envision integrating patient information with medical knowledge graphs to provide personalized drug repurposing predictions. Our pilot human evaluation engaged a small sample size (N=12) of clinicians and scientists, prioritizing an in-depth analysis with a smaller, more qualified group over a broader study with a larger, potentially less specialized participant pool. While the results were statistically significant and this participant number is considered a common practice for evaluating highly specialized tools[62,63], a larger study could incorporate a greater diversity of user expertise. Despite the promising performance of TxGNN's predictions on tests using medical records, confounders might have biased the enrichment scores measured. We conducted a comprehensive evaluation across multiple axes of model performance beyond accuracy, including evaluation across diverse hold-out datasets, a pilot evaluation with human experts, and a large-scale enrichment analysis using medical records.

TxGNN zero-shot drug repurposing model predicts drugs for diseases without FDA-approved treatments and with minimal available knowledge. TxGNN's Explainer enhances the transparency of TxGNN's predictions, fostering trust and aiding human expert evaluations. TxGNN streamlines drug repurposing prediction, especially when the limited availability of disease-specific datasets hinders drug development. In the quest for cost-effective therapeutic innovations, models like TxGNN highlight the computational potential for novel therapeutic avenues.

Data availability. TxGNN's website is at https://zitniklab.hms.harvard.edu/projects/TxGNN. The knowledge graph dataset is available at Harvard Dataverse under a persistent identifier https://doi.org/10.7910/DVN/IXA7BM. All clinical and electronic medical record data were deidentified, and the Institutional Review Board at Mount Sinai, New York City, U.S., approved the study.

Code availability. Python implementation of the methodology developed and used in the study is available via the project website at https://zitniklab.hms.harvard.edu/projects/TxGNN. The code to reproduce results, documentation, and usage examples are at https://github.com/mims-harvard/TxGNN. To facilitate the usage of the algorithm, we developed a TxGNN Explainer, a web-based app available at http://txgnn.org to access TxGNN's predictions.

Acknowledgements. K.H., P.C., and M.Z. gratefully acknowledge the support of NIH R01-HD108794, NSF CAREER 2339524, US DoD FA8702-15-D-0001, awards from Harvard Data Science Initiative, Amazon Faculty Research, Google Research Scholar Program, AstraZeneca Research, Roche Alliance with Distinguished Scientists, Sanofi iDEA-iTECH Award, Pfizer Research, Chan Zuckerberg Initiative, John and Virginia Kaneb Fellowship award at Harvard Medical School, Aligning Science Across Parkinson's (ASAP) Initiative, Biswas Computational Biology Initiative in partnership with the Milken Institute, and Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. P.C. was supported, in part, by the Harvard Summer Institute in Biomedical Informatics. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Authors contribution. P.C. retrieved, processed, and analyzed the knowledge graph. K.H. and P.C. developed and implemented new machine learning methods, benchmarked machine learning models, and analyzed model behavior, all together with M.Z. Q.W. and N.G. implemented the clinician-centered visual explorer of model predictions and performed a user study to evaluate its usability. S.H., A.V., G.N. and B.S.G. performed a validation study examining new predictions of therapeutic use through the electronic health record system. K.H., P.C., Q.W., S.H., A.V., J.L., G.N., B.S.G., N.G., and M.Z. contributed new analytic tools and wrote the manuscript. All authors discussed the results and contributed to the final manuscript. M.Z. designed the study.

Competing interests. The authors declare no competing interests.

Inclusion and ethics statement in global research. We have complied with all relevant ethical regulations. Our research team represents a diverse group of collaborators. Roles and responsibilities were clearly defined and agreed upon among collaborators before the start of the research. All researchers were included in the study design, study implementation, data ownership, intellectual property, and authorship of publications. Our research did not face severe restrictions or prohibitions in the setting of the local researchers, and no specific exceptions were granted for this study in agreement with local stakeholders. Animal welfare regulations, environmental protection and risk-related regulations, transfer of biological materials, cultural artifacts, or associated traditional knowledge out of the country do not apply to our research. Our research does not result in stigmatization, incrimination, discrimination, or personal risk to participants. Appropriate provisions were taken to ensure the safety and well-being of all participants involved. Our team was committed to promoting equitable access to resources and benefits resulting from the research.

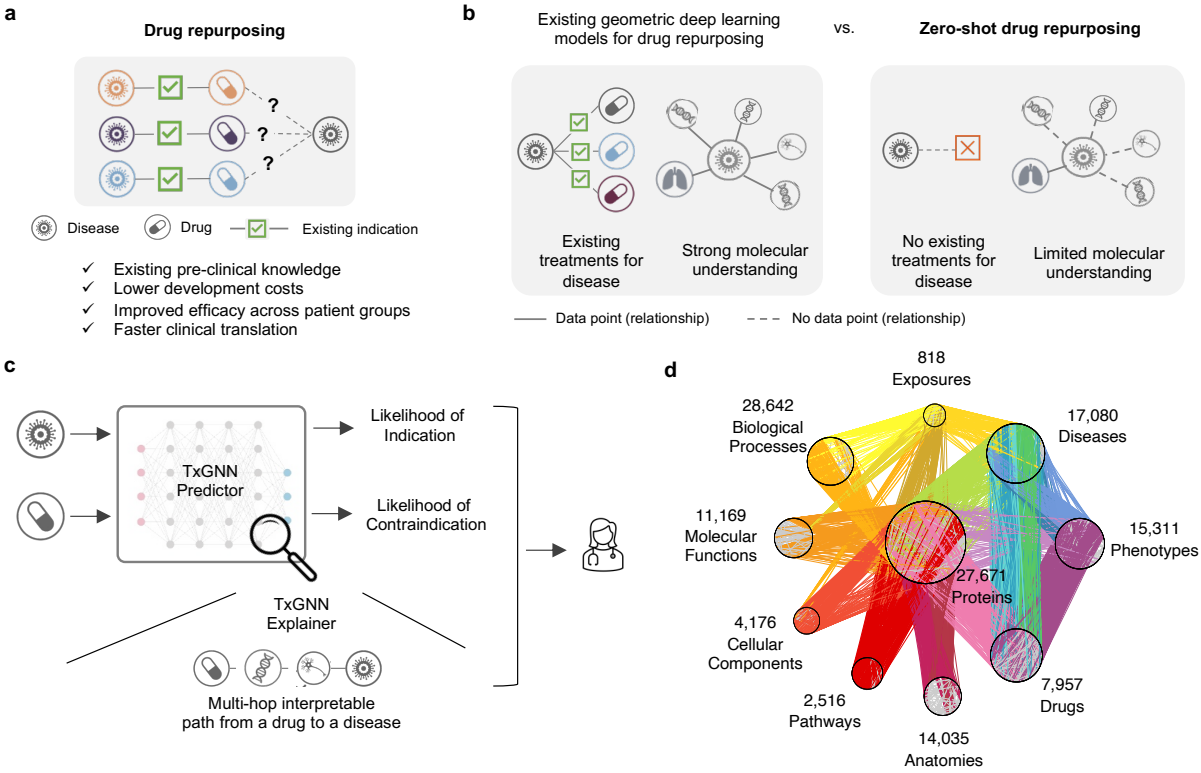

Figure 1: TxGNN is a geometric deep learning approach for drug repurposing across challenging diseases with no known treatments and limited molecular understanding. a. Drug repurposing involves exploring new therapeutic applications for existing drugs to treat different diseases. By capitalizing on abundant pre-existing safety and efficacy data, it can dramatically cut down the cost and time to deliver life-saving therapeutics. **b.** Although AI-based drug repurposing has shown promise, its success has been primarily evaluated on diseases with approved treatments and well-understood molecular mechanisms. However, many diseases of critical pharmaceutical interest lack any available treatments (i.e., zero-shot) and exhibit unclear disease mechanisms. These inherent constraints pose challenges to existing AI methods. In this work, we tackle this problem head-on by formulating it as a zero-shot drug repurposing challenge. **c.** TxGNN presents a novel AI framework that generates actionable predictions for zero-shot drug repurposing. TxGNN geometric deep learning model incorporates a vast and comprehensive biological knowledge graph to accurately predict the likelihood of indication or contraindication for any given disease-drug pair. Additionally, TxGNN generates explainable multi-hop paths, facilitating a scientist-friendly understanding of how the prediction is grounded in biological mechanisms in the KG. The combined power of rich predictions and path-based explanations empowers practitioners to prioritize the most promising drug repurposing candidates. **d.** To support our drug repurposing efforts, we develop a large-scale therapeutics-driven knowledge graph that integrates 17 primary data sources. This knowledge graph paints a comprehensive landscape of biological mechanisms across 17,080 diseases and 7,957 repurposable drugs, compiling scientific knowledge for zero-shot drug repurposing endeavors.

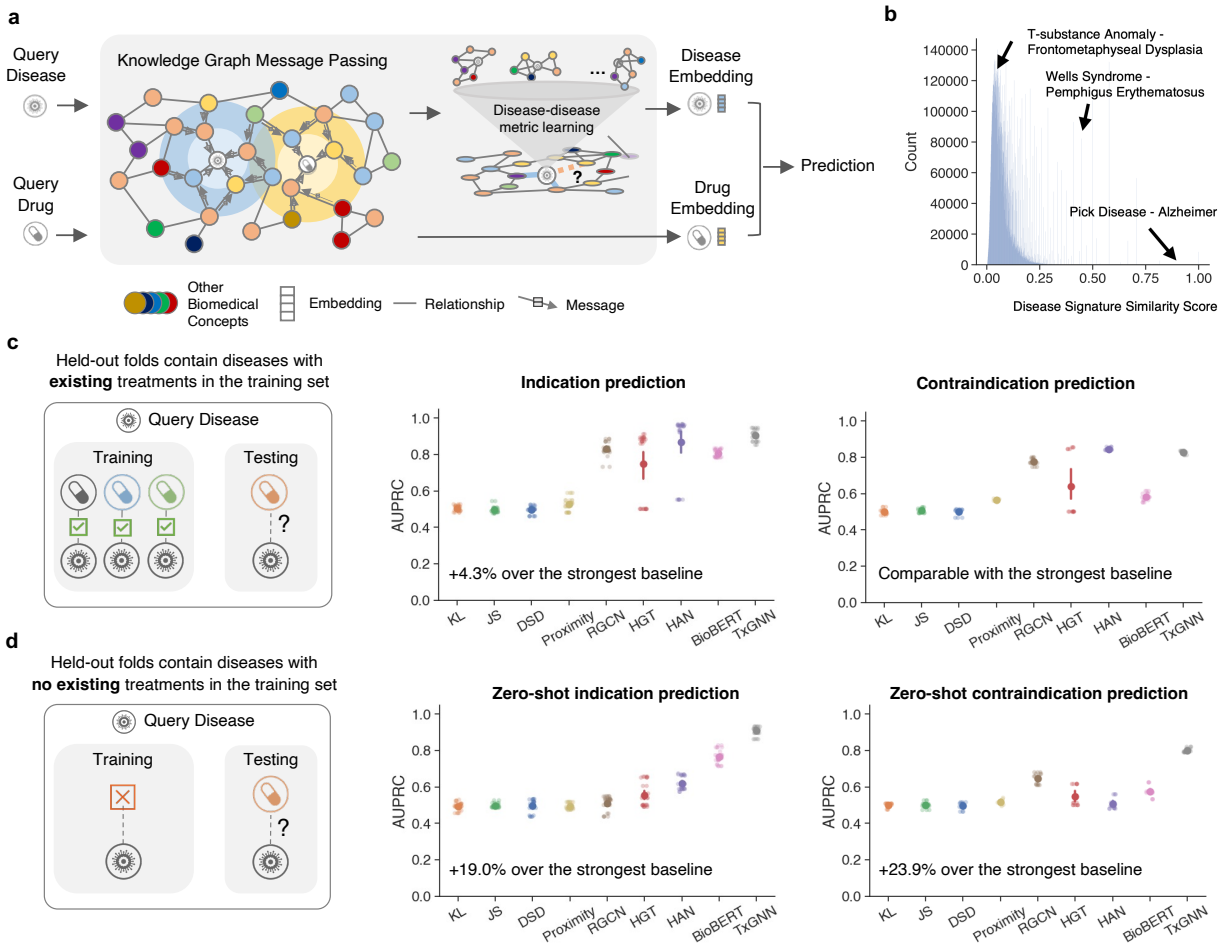

Figure 2: TxGNN predicts indications and contraindications for diseases of no known treatments with high precision. a. TxGNN is a deep learning model that learns to reason over large-scale knowledge graph on predicting the relationship between drug and disease. In zero-shot repurposing, there is limited indication and mechanism information available for the query disease. Our key insight revolves around the interconnectedness of biological systems. We recognize that diseases, despite their distinctiveness, can exhibit partial similarities and share multiple underlying mechanisms. Based on this motivation, we have developed a specialized module known as disease pooling, which harnesses the power of network medicine principles. This module identifies mechanistically similar diseases and employs them to enhance the information available for the query disease. The disease pooling module has demonstrated significant improvements in the prioritization of repurposing candidates within zero-shot settings. **b.** The TxGNN disease similarity score provides a nuanced and meaningful measure of the relationship between diseases. For instance, disease pairs with low similarity scores, such as T-substance anomaly and frontometaphyseal dysplasia (score: 0.084), indicate a lack of shared mechanisms. Conversely, significant similarity is observed when two diseases receive relatively high scores (>0.2). For instance, Wells syndrome and pemphigus erythematosus exhibit a similarity score of 0.433. Both diseases are skin disorders caused by autoimmune dysregulation, although they differ in phenotypic manifestations, with Wells syndrome characterized by redness and swelling and pemphigus erythematosus characterized by blisters. Moreover, certain disease pairs display exceptionally high similarity scores, such as Pick's disease and Alzheimer's (similarity: 0.909), due to their shared neurological causes. This metric empowers TxGNN to discover similar diseases that can inform and enrich the understanding of query diseases lacking treatment and mechanistic information. **c.** The conventional AI-based repurposing evaluates indication predictions on diseases where the model may have seen other approved drugs during training. In this scenario, we show that TxGNN achieves good performance along with existing methods. **d.** To provide a more realistic evaluation, we introduce a novel setup for assessing zero-shot repurposing, where the model is evaluated on diseases that have no approved drugs available during training. In this challenging setting, we observe a significant degradation in performance for baseline methods. In contrast, TxGNN consistently exhibits robust performance, surpassing the best baseline by up to 19% for indications and 23.9% for contraindications. These results highlight the advanced reasoning capabilities of TxGNN when confronted with query diseases lacking treatment options. The evaluation utilizes the area under the precision-recall curve (AUPRC) and is conducted with five random data splits. The mean performance is highlighted, while the 95% confidence intervals are represented by error bars.

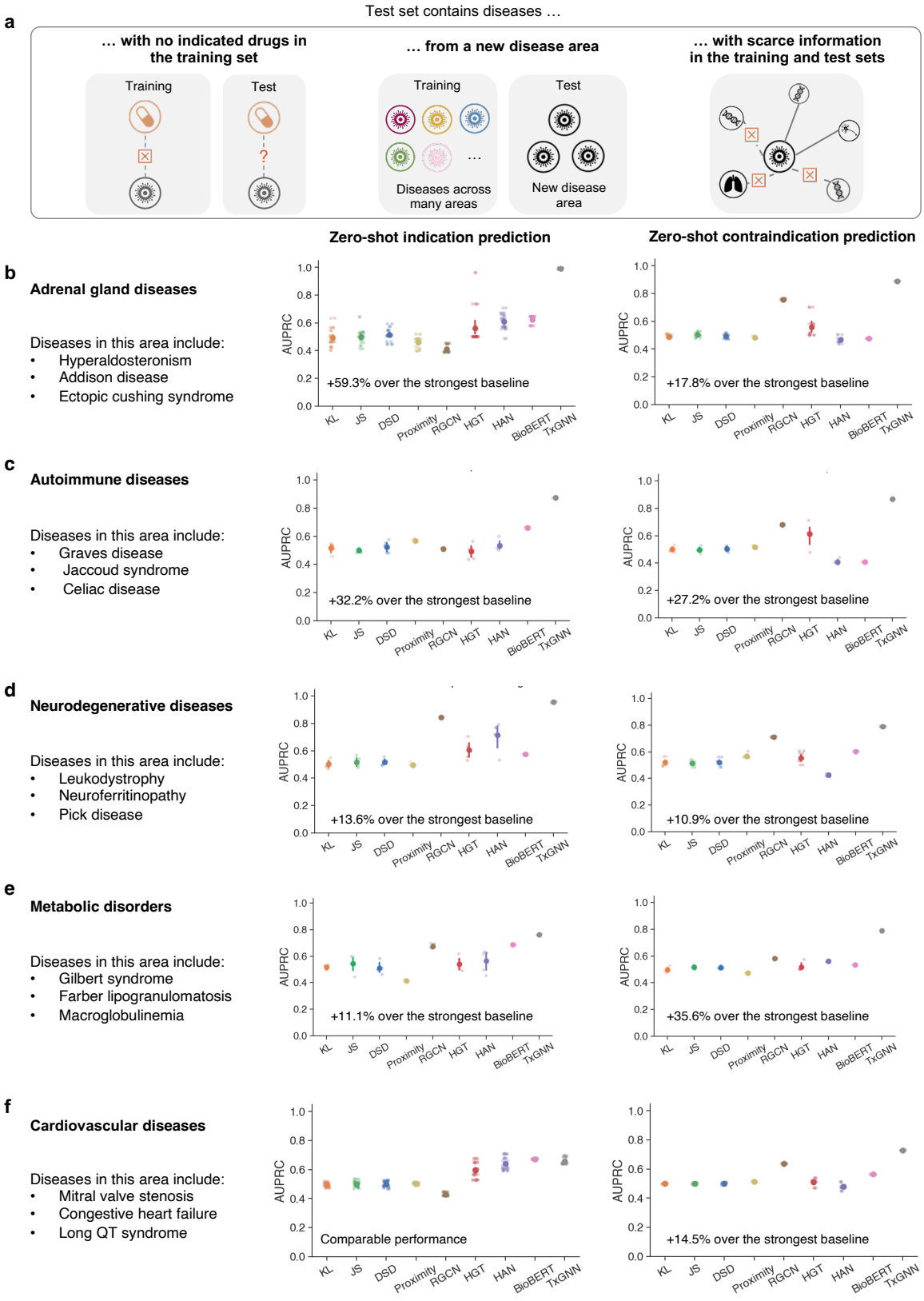

Figure 3: TxGNN accurately predicts therapeutics indications and contraindications across challenging disease areas with limited mechanism understanding. a. Zero-shot drug repurposing addresses diseases without any existing treatments and with a dearth of prior biomedical knowledge. We construct a set of 'disease area' splits to simulate these conditions. The diseases in the holdout set have (1) no approved drugs in training, (2) limited overlap with the training disease set because we exclude similar diseases, and (3) lack molecular data because we deliberately remove their biological neighbors from the training set. These data splits constitute challenging but realistic evaluation scenarios that mimic zero-shot drug repurposing settings. **b-f.** Holdout folds evaluate diseases related to adrenal glands, autoimmune diseases, neurodegenerative diseases, metabolic disorders, and cardiovascular diseases. Additional four disease areas in anemia, diabetes, cancer, and mental health are provided in Figure S7. Raw scores are provided in Tables S4 and S5.TxGNN shows up to 59.3% improvement over the next best baseline in ranking therapeutic candidates, measured by area under the precision-recall curve.

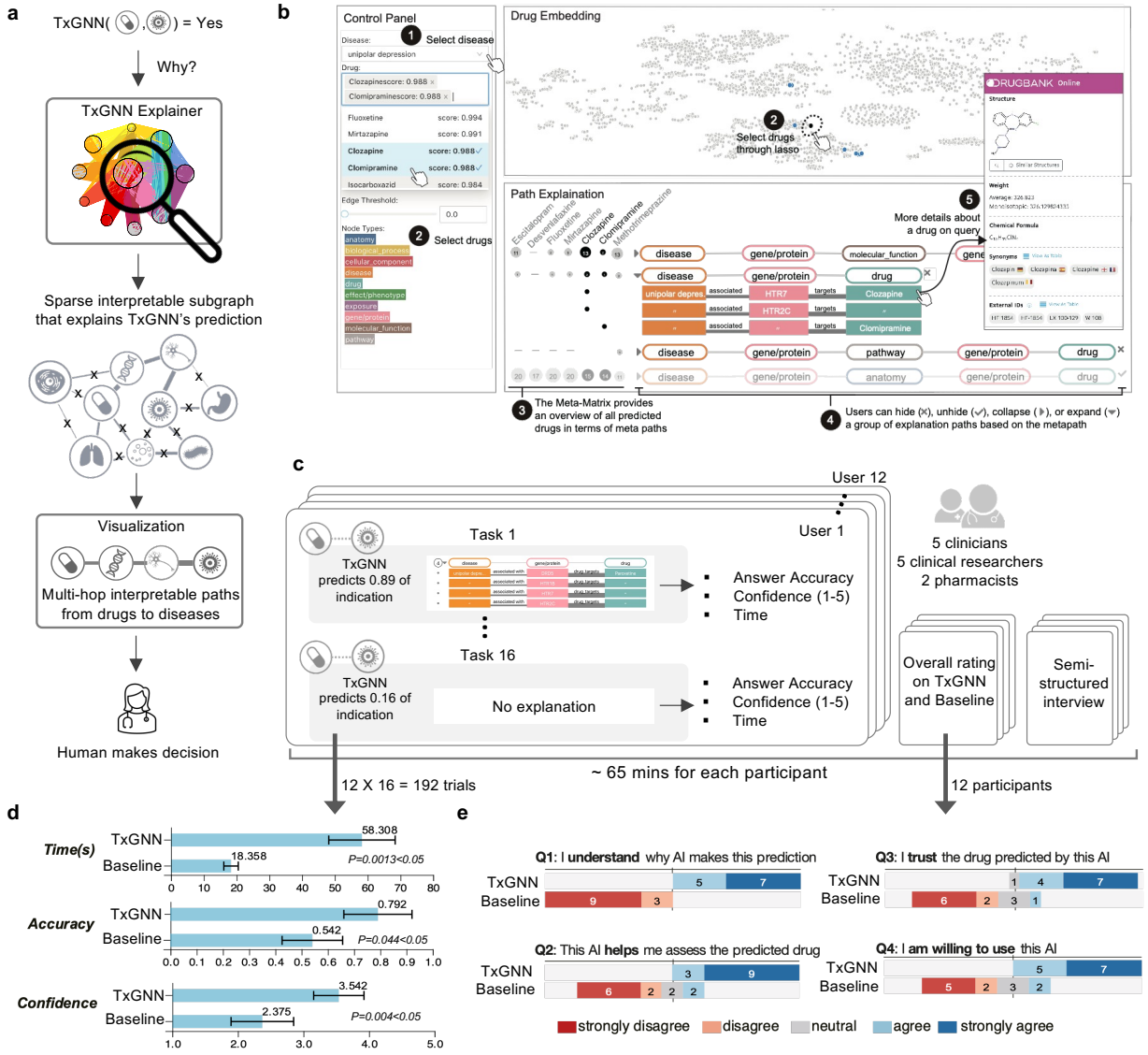

Figure 4: Development, visualization, and evaluation of explanations provided by TxGNN. a. Since prediction scores alone are often insufficient for trustworthy deployment of machine learning models, we develop TxGNN Explainer to facilitate adoption by clinicians and scientists. TxGNN Explainer uses state-of-the-art graph explainability techniques to identify a sparse interpretable subgraph that underlies the model's predictions. For each therapeutic candidate, TxGNN Explainer generates a multi-hop pathway composed of various biomedical entities that connect the query disease to the proposed therapeutic candidate. We develop a visualization module that transforms the identified subgraph into these multi-hop paths in a manner that aligns with the cognitive processes of clinicians and scientists. **b.** We design a web-based graphical user interface to support clinicians and scientists in exploring and analyzing the predictions and explanations generated by TxGNN. The 'Control Panel' allows users to select the disease of interest and view the top-ranked TxGNN predictions for the query disease. The 'edge threshold' module enables users to modify the sparsity of the explanation and thereby control the density of the multi-hop paths displayed. The 'Drug Embedding' panel allows users to compare the position of a selected drug relative to the entire repurposing candidate library. The 'Path Explanation' panel displays the biological relations that have been identified as crucial for TxGNN's predictions regarding therapeutic use. **c.** To evaluate the usefulness of TxGNN explanations, we conducted a user study involving 5 clinicians, 5 clinical researchers, and 2 pharmacists. These participants were shown 16 drug-disease combinations with TxGNN's predictions, where 12 predictions were accurate. For each pairing, participants indicated whether they agreed or disagreed with TxGNN's predictions using the explanations provided. **d.** We compared the performance of TxGNN Explainer with a no-explanation baseline in terms of user answer accuracy, task completion time, and user confidence. The results are aggregated on 192 trials (12 participants $\times$ 16 tasks) and reveal a significant improvement in accuracy (+46%) and confidence (+49%) when explanations were provided. Error bars represent 95% confidence intervals. **e.** At the conclusion of the user study, participants were asked qualitative usability questions. Clinicians and scientists agreed that the explanations provided by TxGNN were helpful in assessing the predicted drug-disease relationships and instilled greater trust in the TxGNN's predictions.

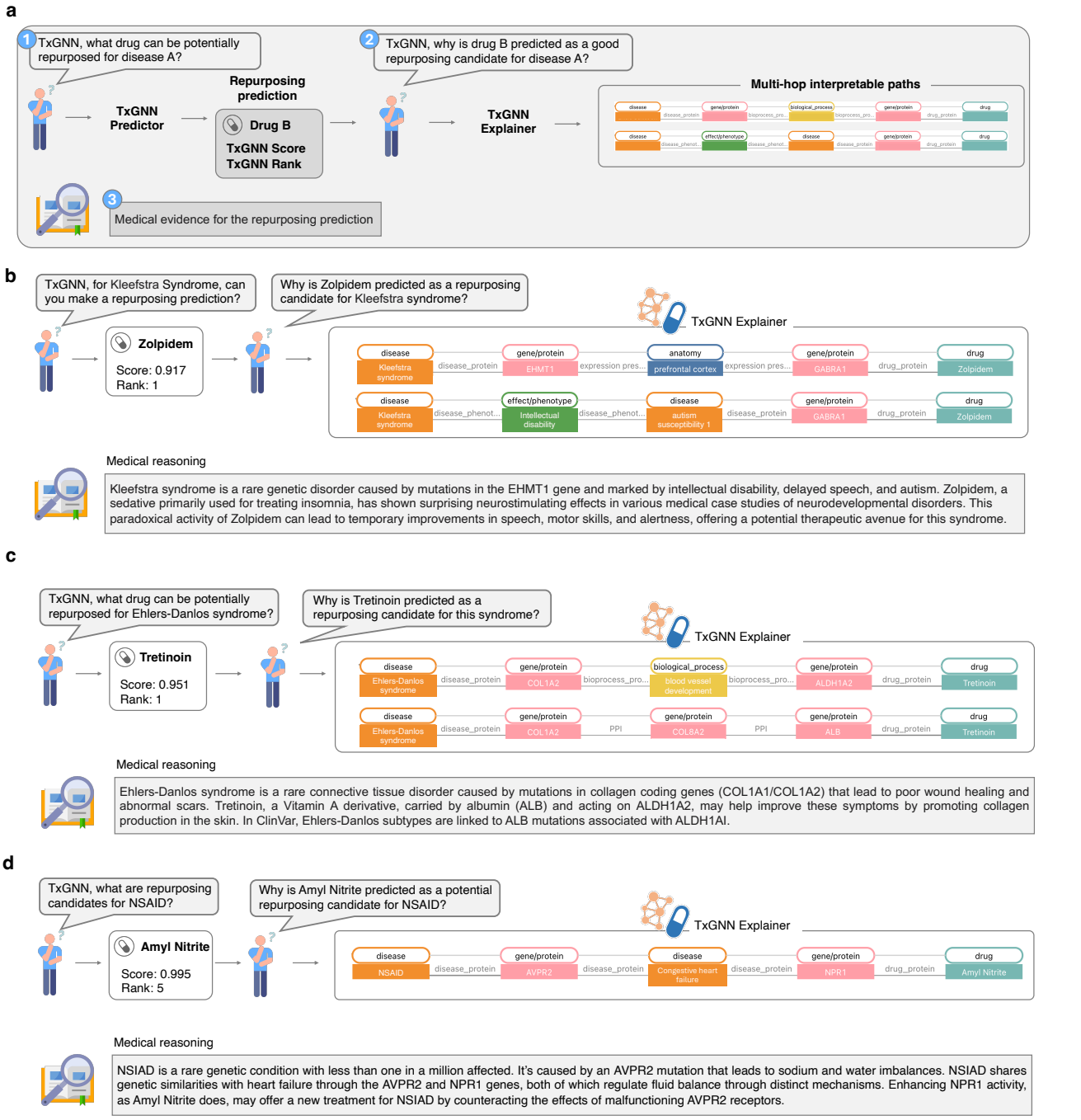

Figure 5: Highlighted cases where interpretable paths produced by TxGNN Explainer align with clinical evidence a. We assess the alignment of drug repurposing candidates identified by TxGNN with established medical reasoning across three rare diseases. The process begins with the TxGNN Predictor, which selects potential drugs for repurposing based on a disease query, and continues with the TxGNN Explorer, which provides interpretable paths explaining the selection. Our case studies conclude with an independent verification of the TxGNN 's predictions against clinical knowledge, showcasing the congruence between the TxGNN's recommendations and medical insights. **b.** TxGNN predicts Zolpidem, typically used as a sedative, as a repurposing candidate for Kleefstra syndrome, characterized by developmental delays and neurological symptoms. Despite Zolpidem's conventional inhibitory effects on the brain, TxGNN Explainer suggests its potential to enhance prefrontal cortex activity and improve cognitive functions in those with Kleefstra syndrome. TxGNN's counterintuitive recommendation aligns with emerging clinical evidence of Zolpidem's ability to "awaken" dormant neurons, thereby potentially aiding in speech, motor skills, and alertness in individuals with neurodevelopmental disorders. **c.** TxGNN identifies Tretinoin as the top candidate for treating Ehlers-Danlos syndrome. TxGNN's predictive rationale is rooted in the drug's interactions with albumin (ALB) and ALDH1A2, which aligns with medical insights about Ehlers-Danlos syndrome regarding collagen loss and inflammation mitigation. **d.** TxGNN identifies Amyl Nitrite as a therapeutic option for nephrogenic syndrome of inappropriate antidiuresis (NSIAD). In NSIAD, an AVPR2 mutation leads to water and sodium imbalances. TxGNN Explorer points out the connection between NSIAD and Amyl Nitrite through congestive heart failure, a condition with similar fluid retention issues, by exploring gene interactions (AVPR2 and NPR1) that regulate electrolyte balance.

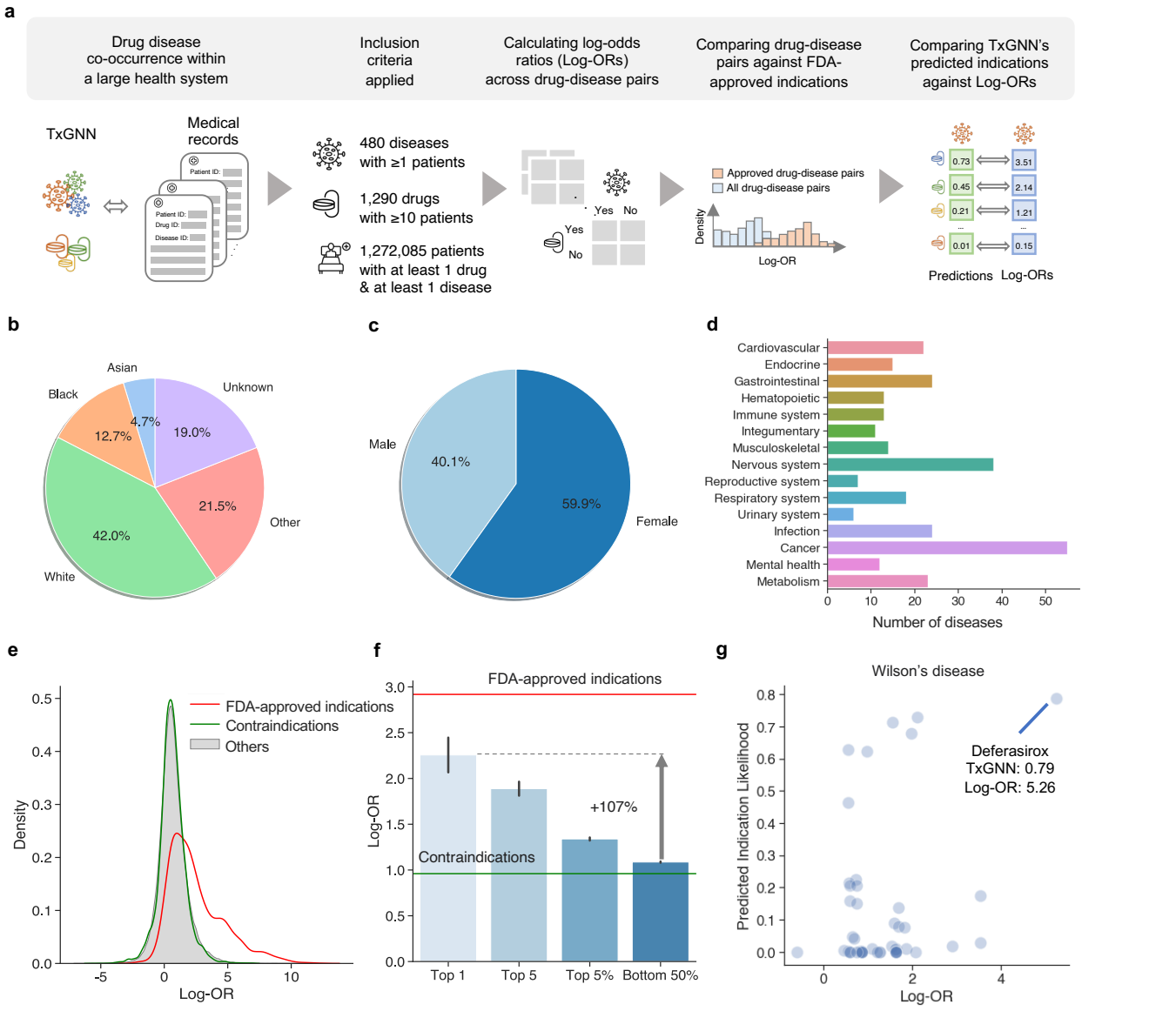

Figure 6: Evaluating TxGNN's novel predictions in a large healthcare system. a. We illustrate the steps taken to evaluate TxGNN's novel indications predictions in Mount Sinai's electronic health record (EHR) system. First, we matched the drugs and diseases in the TxGNN knowledge graph to the EHR database, resulting in a curated cohort of 1.27 million patients spanning 480 diseases and 1,290 drugs. Next, we calculated the log-odds ratio (log-OR) for each drug-disease pair, which served as an indicator of the usage of a particular drug for a specific disease. We then validated the log-OR metric as a proxy for clinical usage by comparing drug-disease pairs against FDA-approved indications. Finally, we evaluated TxGNN's novel predictions to determine if their Log-ORs exhibited enrichment within the medical records. **b.** The racial diversity within the patient cohort. **c.** The sex distribution of the patient cohort. **d.** The medical records encompassed a diverse range of diseases spanning major disease areas, ensuring comprehensive coverage and representation. **e.** In validating log-ORs as a proxy metric for clinical prescription, we observed that while the majority of drug-disease pairs exhibited low log-OR values, there was a significant enrichment of log-OR values for FDA-approved indications. Additionally, we noted that contraindications displayed similar log-OR values to the general non-indicated drug-disease pairs, minimizing potential confounders such as adverse drug effects. **f.** We evaluated Log-ORs for the novel indications proposed by TxGNN. The y-axis represents the Log-OR of the disease-drug pairs, serving as a proxy for clinical usage. For each disease, we ranked TxGNN's predictions and extracted the average Log-OR values for the top 1, top 5, top 5%, and bottom 50% of novel drug candidates. The red horizontal line represents the average Log-OR for FDA-approved indications, while the green horizontal line represents the average Log-OR for contraindications. We observed a remarkable enrichment in the clinical usage of TxGNN's novel predictions. The error bar is 95% confidence interval. **g.** We provide a case study of TxGNN's predicted scores plotted against the Log-OR for Wilson's disease. Each point on the plot represents a therapeutic candidate. The top 1 most probable candidate suggested by TxGNN is highlighted, indicating its associated TxGNN score and Log-OR.

Online Methods

The Methods are structured as follows: 1) curation of knowledge graph dataset (Section 1), 2) description of machine learning approach (Section 2), 3) pilot human evaluation and usability study (Section 3), and 4) evaluation of novel predictions against medical records within a large healthcare system (Section 4).

1 Training dataset

The knowledge graph is heterogeneous, with 10 types of nodes and 29 types of undirected edges. It contains 123,527 nodes and 8,063,026 edges. Tables S2 and S3 show a breakdown of nodes by node type and edges by edge type, respectively. The knowledge graph and all auxiliary data files are available via Harvard Dataverse at https://doi.org/10.7910/DVN/IXA7BM. Supplementary Note S1 provides detailed information about datasets and curation of the knowledge graph.

2 Geometric deep learning approach

Notation. We are given a heterogeneous knowledge graph (KG) $G = (\mathcal{V}, \mathcal{E}, \mathcal{T}_R)$ with nodes in the vertex set $v_i \in \mathcal{V}$, edges $e_{i,j} = (v_i, r, v_j)$ in the edge set $\mathcal{E}$, where $r \in \mathcal{T}_R$ indicates the relation type, v_i is called the head/source node and v_j is called the tail/target node. Each node also belongs to a node type set $\mathcal{T}_V$. Each node also has an initial embedding, which we denote as $\mathbf{h}_i^{(0)}$.

Problem definition. Given a disease i and drug j, we want to predict the likelihood of the drug being (1) indicated for the disease or (2) contraindicated for the disease. Our approach is to induce inductive priors in the model by incorporating factual knowledge from the KG into the model. This process enhances the model's reasoning capabilities to form hypotheses and make predictions about disease treatments.

Experimental setup. We describe detailed experimental protocols, including data split curation, negative sampling scheme, hyperparameter tuning, and implementation details in Supplementary Note S4.

2.1 Overview of TxGNN approach

TxGNN is a deep learning approach for mechanistic predictions in drug discovery based on molecular networks perturbed in disease and targeted by therapeutics. TxGNN is composed of four modules: (1) a heterogeneous graph neural network-based encoder to obtain biologically meaningful network representation for each biomedical entity; (2) a disease similarity-based metric

learning decoder to leverage auxiliary information to enrich the representation of diseases that lack molecular characterization; (3) an all-relation stochastic pre-training followed by a drug-disease centric full-graph fine-tuning strategy; (4) a graph explainability module to retain a sparse set of edges that are crucial for prediction as a post-training step. Next, we expand each module in detail.

2.2 Heterogeneous graph neural network encoder

Our objective is to learn a general encoder of a biomedical knowledge graph by learning a numerical vector (embedding) for each node, encapsulating the biomedical knowledge contained within its neighboring relational structures. This involves transforming initial node embeddings using a sequence of local graph-based non-linear function transformations to refine embeddings[29,64]. These transformations are subject to iterative optimization, guided by a loss function aimed at minimizing the error in therapeutic use predictions. Through this process, the system converges to an optimized set of node embeddings.

Step 1: Initializing latent representations. We denote the input node embedding $\mathbf{X}_i$ for each node i, which is initialized using Xavier uniform initialization[65]. For every layer l of message-passing, there are the following three stages:

Step 2: Propagating relation-specific neural messages. For every relation type, first calculates a transformation of node embedding from the previous layer $\mathbf{h}^{(l-1)}$, where the first layer $\mathbf{h}^{(0)} = \mathbf{X}$. This is achieved via applying a relation-specific weight matrix $\mathbf{W}_{r,M}^{(l)}$ on the previous layer embedding:

$$\mathbf{m}_{r,i}^{(l)} = \mathbf{W}_{r,M}^{(l)} \mathbf{h}_i^{(l-1)}. \tag{1}$$

Step 3: Aggregating local network neighborhoods. For each node v_i, we aggregate on the incoming messages $\{\mathbf{m}_{r,j}^{(l)} | j \in \mathcal{N}_r(i)\}$ from neighboring nodes of each relation r denoted as $\mathcal{N}_r(i)$ by taking the average of these messages:

$$\widetilde{\mathbf{m}^{(l)}}_{r,i} = \frac{1}{|\mathcal{N}_r(i)|} \sum_{j \in \mathcal{N}_r(i)} \mathbf{m}_{r,j}^{(l)}. \tag{2}$$

Step 4: Updating latent representations. We then combine the node embedding from the last

layer and the aggregated messages from all relations to obtain the new node embedding:

$$\mathbf{h}_i^{(l)} = \mathbf{h}_i^{(l-1)} + \sum_{r \in \mathcal{T}_R} \widetilde{\mathbf{m}^{(l)}}_{r,i}. \tag{3}$$

After L layers of propagation, we arrive at our encoded node embeddings $\mathbf{h}_i$ for each node i.

2.3 Predicting drug-disease relationships

TxGNN employs disease and drug embeddings to predict indications, contra-indications, and off-label use for each disease-drug pair. Considering the three relation types that need prediction, a trainable weight vector $\mathbf{w}_r$ is assigned to each type. The interaction likelihood for a specific relation is then determined using the DistMult approach[66]. Formally, for a disease i, drug j, and relation r, the predicted likelihood p is calculated as follows:

$$p_{i,j,r} = \frac{1}{1 + \exp(-\text{sum}(\mathbf{h}_i \cdot \mathbf{w}_r \cdot \mathbf{h}_j))}. \tag{4}$$

2.4 Embedding-based disease similarity search

Research on diseases varies widely based on factors such as their prevalence and complexity. For instance, the molecular basis of many rare diseases remains poorly understood[67,68]. Despite this, rare diseases often offer significant opportunities for therapeutic advancements[69]. The limited knowledge surrounding these diseases has heightened the importance of machine learning predictions. This shortage of research is evident in the biological knowledge graph, where rare diseases are characterized by a lack of relevant nodes and edges, leading to lower-quality graph embeddings. For example, diseases without any connections in the knowledge graph are assigned a random initialization for their embedding. Empirical evidence indicates that GNN models exhibit substantially reduced predictive performance on disease-centric splits designed to reflect the sparse nature of knowledge on these diseases, as opposed to random splits (Figure 1g).

We posit that the network embeddings generated for these diseases lack significance due to the sparse prior information in the KG. Consequently, there is a necessity for a model to enhance and supplement the network embeddings for these diseases. The underlying principle is that human physiology represents an interconnected system wherein diseases exhibit similarities across various dimensions—e.g., lung cancer and brain cancer are analogous within the cancer disease dimension, while lung cancer and asthma are comparable within the lung disease dimension. Leveraging this

concept by utilizing a model to extract relevant information from a group of similar but better-characterized diseases in the KG, it is possible to enrich the embedding of a target disease, thereby improving its predictive accuracy.

To achieve this, TxGNN employs a three-step procedure: (1) it constructs a disease signature vector to capture the complex similarities among diseases; (2) it utilizes an aggregation mechanism to combine the embeddings of similar diseases into a comprehensive auxiliary embedding, which supplements the original disease embedding; (3) it introduces a gating mechanism to modulate the influence between the original disease embedding and the auxiliary disease embedding, acknowledging that many well-characterized diseases possess adequate embeddings and do not require supplementation. Each of these steps is elaborated upon in the sections that follow.

Disease signature vectors. The primary objective of this module is to derive a signature vector $\mathbf{p}_i$ for each disease i. Given the insufficiency of disease representations produced solely by graph neural networks in fully capturing the nuances of diseases, these representations are not ideal for direct similarity computations. Instead, we employ graph theoretical methods[14] to calculate disease similarities. Additionally, variations of signature vectors are detailed in Supplementary Note S2. Specifically, we generate a vector that encapsulates the local neighborhoods surrounding a disease. For disease i, the signature vector is formally defined as follows:

$$\mathbf{p}_i^{\mathrm{AT}} = [\ \mathrm{p}_1 \cdots \mathrm{p}_{|\mathcal{V}_{\mathrm{P}}|}\ \mathrm{ep}_1 \cdots \mathrm{ep}_{|\mathcal{V}_{\mathrm{EP}}|}\ \mathrm{ex}_1 \cdots \mathrm{ex}_{|\mathcal{V}_{\mathrm{EX}}|}\ \mathrm{ep}_1 \cdots \mathrm{ep}_{|\mathcal{V}_{\mathrm{EP}}|}\ \mathrm{d}_1 \cdots \mathrm{d}_{|\mathcal{V}_{\mathrm{D}}|}\ ], \tag{5}$$

where

$$\mathrm{p}_j = \begin{cases} 1 & \text{if } j \in \mathcal{N}_i^{\mathrm{P}} \\ 0 & \text{otherwise} \end{cases}, \mathrm{ep}_j = \begin{cases} 1 & \text{if } j \in \mathcal{N}_i^{\mathrm{EP}} \\ 0 & \text{otherwise} \end{cases}, \mathrm{ex}_j = \begin{cases} 1 & \text{if } j \in \mathcal{N}_i^{\mathrm{EX}} \\ 0 & \text{otherwise} \end{cases} \mathrm{d}_j = \begin{cases} 1 & \text{if } j \in \mathcal{N}_i^{\mathrm{D}} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

and $\mathcal{N}_i^{\mathrm{P}}, \mathcal{N}_i^{\mathrm{EP}}, \mathcal{N}_i^{\mathrm{EX}}, \mathcal{N}_i^{\mathrm{D}}$ is the set of gene/protein, effect/phenotype, exposure, diseases nodes lie in the 1-hop neighborhood of disease i. We also adopt the dot product as the similarity measure, which means the similarity is the sum of all shared nodes across the four node types:

$$\mathrm{sim}^{\mathrm{AT}}(i,j) = \mathbf{p}_i^{\mathrm{AT}} \cdot \mathbf{p}_j^{\mathrm{AT}} = |\mathcal{N}_i^{\mathrm{P}} \cap \mathcal{N}_j^{\mathrm{P}}| + |\mathcal{N}_i^{\mathrm{EP}} \cap \mathcal{N}_j^{\mathrm{EP}}| + |\mathcal{N}_i^{\mathrm{EX}} \cap \mathcal{N}_j^{\mathrm{EX}}| + |\mathcal{N}_i^{\mathrm{D}} \cap \mathcal{N}_j^{\mathrm{D}}|. \tag{7}$$

Given the selected signature for diseases and calculated similarities among the diseases, for

a query disease, we can then obtain k most similar diseases for a query disease i:

$$\mathcal{D}_{\text{sim},i} = \text{argmax}^{k}_{j \in \mathcal{V}_{\text{D}}} \text{sim}(i, j). \tag{8}$$

Disease metric learning. Given a set of similar diseases, TxGNN generates disease embeddings that integrate various measures of disease similarity into a unified embedding, capable of augmenting the representation of a query disease that may be sparsely annotated. To achieve this, we adopt a weighted scheme, wherein each disease is weighted according to its similarity score, as follows:

$$\mathbf{h}_i^{sim} = \sum_{j \in \mathcal{D}_{\text{sim}}} \frac{\text{sim}(i, j)}{\sum_{k \in \mathcal{D}_{\text{sim}}} \text{sim}(i, k)} * \mathbf{h}_j. \tag{9}$$

Gating disease embeddings. The final stage involves updating the original disease embedding $\mathbf{h}_i$ with the disease-disease metric learning embedding $\mathbf{h}_i^{sim}$ via a gating mechanism. This mechanism employs a scalar $c \in [0, 1]$ to modulate the influence between these two embeddings. Special consideration is needed here because, for diseases that are well-documented in the knowledge graph, the disease-disease metric learning embedding might not be necessary and could potentially skew the final embedding. Conversely, for diseases lacking characterization, the disease-disease metric learning embedding is invaluable due to the original embedding's inadequacy in representing molecular mechanisms. The use of a learnable attention mechanism for deciding whether to prioritize the original or augmented embedding is not effective, as it tends to overvalue the original embeddings for well-characterized diseases, thereby neglecting the supplementary embedding. Alternatively, we introduce a heuristic algorithm that determines weighting based on the degree of node connectivity $|\mathcal{N}_i^r|$ within the drug-disease relationship being analyzed. A higher degree indicates a well-characterized disease, suggesting a reduced reliance on the disease-disease metric learning embedding and vice versa. The scalar's value is designed to be significantly high for minimal node degrees (0 or 1) and to decrease rapidly with increasing node degrees. To achieve this gradient, we use an inflated exponential distribution density function with $\lambda = 0.7$:

$$c_i = 0.7 * \exp(-0.7 * |\mathcal{N}_i^r|) + 0.2. \tag{10}$$

We observe the result is not sensitive to λ (Figure S6). Finally, we use parameter search and find

optimal $\lambda = 0.7$. Then, we can finally obtain an augmented disease embedding:

$$\widehat{\mathbf{h}}_i = c_i * \mathbf{h}_i^{sim} + (1 - c_i) * \mathbf{h}_i. \tag{11}$$

Finally, TxGNN uses augmented disease embeddings as input to the latent decoder described in Section 2.3 to produce drug repurposing predictions.

2.5 Training TxGNN deep graph models

Objective function. The objective of the training process is to predict the presence of a relation between two entities within a knowledge graph, which can be viewed as a binary classification task for each relation type. The dataset for positive samples, denoted as $\mathcal{D}_+$, comprises all pairs (i, j) across various relation types $r \in \mathcal{T}_R$, with the label $y_{i,r,j} = 1$ indicating the presence of a relation. To generate the dataset for negative samples, $\mathcal{D}_-$, we use a sampling technique detailed in Supplementary Note S4.3, creating counterparts for each positive pair. For a given pair i, j and relation type r, the model estimates the probability $p_{i,r,j}$ of a relation's existence. The training loss is then calculated using the binary cross-entropy loss formula:

$$\mathcal{L} = \sum_{(i,r,j) \in \mathcal{D}_+ \cup \mathcal{D}_-} y_{i,r,j} * \log\left(p_{i,r,j}\right) + \left(1 - y_{i,r,j}\right) * \log\left(1 - p_{i,r,j}\right). \tag{12}$$

Previous research has emphasized knowledge graph completion, optimizing models across the entire spectrum of relations within a knowledge graph[70]. This approach, however, may dilute the model's capacity to capture specific knowledge, particularly when the interest lies solely in drug-disease relations. Given that drug-disease interactions are governed by complex biological mechanisms, the extensive range of biomedical relations in a knowledge graph can offer a comprehensive view of biological systems. The primary challenge lies in optimizing performance on a select group of relations while beneficially leveraging the broader set of relations for knowledge transfer, avoiding catastrophic forgetting of general knowledge.

To address this challenge, TxGNN adopts a pre-training strategy. Initially, during pre-training, TxGNN learns to predict relations across the entire KG using stochastic mini-batching, which helps to encapsulate biomedical knowledge within enriched node embeddings. Subsequently, in the fine-tuning phase, TxGNN focuses specifically on drug-disease relations. This targeted training sharpens the model's ability to generate drug-disease-specific embeddings, thereby

optimizing the quality of drug repurposing predictions.

Pre-training. TXGNN initially undergoes pre-training on millions of biomedical entity pairs spanning the entire array of relations. Given the extensive number of edges, training on the full graph is not computationally viable. Therefore, stochastic mini-batching is employed, allowing for the training on a subset of pairs at each step. This process ensures that each epoch covers all data pairs within the training knowledge graph. During this phase, degree-adjusted disease augmentation is deactivated and all relation types are treated equally. The weights from the pre-trained encoder are subsequently utilized to initialize the encoder model for the fine-tuning phase. It is important to note that the weights in the decoder, specifically for DistMult $\mathbf{w}_r$, are reinitialized prior to fine-tuning to mitigate the risk of negative knowledge transfer.

Fine-tuning. After the pre-training phase, the model initialization encapsulates a broad spectrum of biological knowledge. The subsequent phase concentrates on refining the prediction of drug-disease relations. This refinement is achieved by exclusively considering samples of drug-disease pairs (i, j), where the relation types r fall within the set {indication, contraindication, rev_indication, rev_contraindication}. Other relation types, while not directly included in the training objective, remain part of the knowledge graph to facilitate information flow between drug and disease nodes. During the fine-tuning phase, the model activates the degree-adjusted inter-disease embedding feature. The TXGNN model undergoes both pre-training and fine-tuning in an end-to-end process. The variant that exhibits the highest performance on the validation set is selected for evaluation on the test set and is used for downstream analyses.

2.6 Generating multi-hop interpretable explanations

In a trained drug repurposing prediction model, consider a target node j and a neighboring source node i connected by an edge $e_{i,j}$ at layer l. For each relation r, intermediate messages $\mathbf{m}_{r,i}^{(l)}$ and $\mathbf{m}_{r,j}^{(l)}$ are computed. These embeddings are concatenated and input into a relation-specific, single-layer neural network parameterized by $\mathbf{W}_{g,r}^{(l)}$. This network predicts the probability of masking the message from source node i during the computation of the target node j's embedding. The output is processed through a gate, which includes a sigmoid layer to constrain the probability to the range [0, 1], followed by an indicator function that determines whether the edge should be dropped:

$$z_{i,j,r}^{(l)} = \mathbf{1}_{[\mathbb{R}>0.5]}\left(\text{sigmoid}\left(\mathbf{W}_{g,r}^{(l)}\left(\mathbf{m}_{r,i}^{(l)} \| \mathbf{m}_{r,i}^{(l)}\right)\right)\right), \tag{13}$$

such that $z_{i,j,r}^{(l)} \in 0, 1$. In practice, a location bias of 3 is added to the $\mathrm{sigmoid}$ function during initialization to ensure that its outputs are initially close to 1. This means that at the start, the gates remain open, allowing the model to adaptively close the gates and mask edges within the subgraph as needed. This approach is essential because starting with random initialization, which drops edges randomly, creates a significant discrepancy between the original and updated predictions. Consequently, the model's primary focus shifts towards minimizing this discrepancy rather than balancing the two objectives. To refine this mechanism, when a gate outputs 0, the corresponding message is not simply removed. Instead, it is substituted with a learnable baseline vector $\mathbf{b}_r^{(l)}$ for each relation r and layer l. Therefore, the revised message from source node i to target node j is represented as follows:

$$\hat{\mathbf{m}}_{i,r}^{(l)} = z_{i,j,r}^{(l)} \cdot \mathbf{m}_{i,r}^{(l)} + (1 - z_{i,j,r}^{(l)}) \cdot \mathbf{b}_r^{(l)}. \tag{14}$$

Following the modification of messages with the learnable baseline vector, the process continues with the standard steps of message aggregation and node embedding updates as described in Section 2.2. This updated node embedding is then utilized in inter-disease augmentation (Section 2.4) and to generate the updated predictions $\hat{p}$ for the interaction between a drug and a disease (Section 2.3). The optimization of the GraphMask gate weights is guided by two objectives. The first, faithfulness, aims to ensure that the updated predictions, after applying the mask, align closely with the initial prediction outcomes. The second objective encourages the model to apply as extensive a masking as feasible. These objectives inherently entail a trade-off: increasing the extent of masking tends to enlarge the discrepancy between the updated and original predictions. This scenario is addressed through constrained optimization, employing Lagrange relaxation to balance the objectives. Specifically, the optimization seeks to maximize the Lagrange multiplier λ to enforce the constraint while simultaneously minimizing the primary objective. The loss function employed for this purpose is formulated as follows:

$$\max_{\lambda} \min_{\mathbf{W}_g} \sum_{k=1}^{L} \sum_{(i,j,r) \in \mathcal{D}_+ \cup \mathcal{D}_-} \mathbf{1}_{[\mathbb{R} \neq 0]} z_{i,j,r}^{(k)} + \lambda \left(\|\hat{p}_{i,j,r} - p_{i,j,r}\|_2^2 - \beta \right), \tag{15}$$

where β is the margin between the updated and original prediction. After the training process is complete, edges (i, j, r) for which $z_{i,j,r}^{(k)} = 0$ can be eliminated. The remaining edges serve as explanations for the model's predictions. Additionally, the value computed prior to the application of the indicator function can be employed to quantify each edge's contribution to the prediction. This

facilitates the adjustment of granular differences in the contributions. More detailed adaptations of the GraphMask approach are discussed in Supplementary Note S3.

3 Pilot usability evaluation of TxGNN with medical experts

The TxGNN Explorer was developed following a user-centric design study process, as outlined in a prior study[27]. This process involved comparing three visual presentations of GNN explanations from the user's perspective. The findings from this comparison motivated the adoption of path-based explanations, which were preferred based on user feedback. The usability of the TxGNN Explorer was assessed through a comparison with a baseline that only displayed drug predictions and their associated confidence scores.

For this usability study, twelve medical experts (7 males and 5 females, average age 34.25, referred to as P1-12) were recruited through personal contacts, Slack channels, and email lists from collaborating institutions, with all participants providing informed consent. The group comprised five clinical researchers (P1-3, P11-12) and five practicing physicians (P4, P7-10), all holding M.D. degrees, and two medical school students with prior experience as pharmacists (P5, P6). Each participant had at least five years of experience in various medical specialties.

The study was conducted remotely via Zoom in compliance with COVID-19-related restrictions. Participants accessed the study system (as shown in Figure S5) using their own computers and shared their screens with the interviewer. The sequence in which predictions were presented, along with the conditions (TxGNN Explorer or the baseline approach), was randomized and counterbalanced across participants and tasks.

In the drug assessment tasks, participants' accuracy, confidence levels, and task completion times were evaluated across 192 trials (16 tasks × 12 participants). Specifically, participants were tasked with 1) determining the correctness of a drug prediction (i.e., if the drug could potentially be used to treat the disease) and 2) rating their confidence in their decision on a 5-point Likert scale (1=not confident at all, 5=completely confident). The system automatically logged the time taken to evaluate each prediction.

Upon completing all predictions, participants provided subjective ratings for both tasks regarding *Trust*, *Helpfulness*, *Understandability*, and *Willingness to Use*, using a 5-point Likert scale (1=strongly disagree, 5=strongly agree). Subsequent semi-structured interviews yielded insights and feedback on the tool's predictions, explanations, and overall user experience. Each session of the user study lasted approximately 65 minutes.

4 Analysis of medical records from a large healthcare system

Patient data from the Mount Sinai Health System's electronic health records (EHR) in New York City, U.S., were utilized to examine patterns from predictions in clinical practice. The Mount Sinai Institutional Review Board approved the study, ensuring all clinical data were de-identified. The initial cohort included over 10 million patients, refined to those over 18 years of age with at least one drug and one diagnosis on record, resulting in 1,272,085 patients. This refined cohort comprised 40.1% males, with an average age of 48.6 years (SD: 18.6 years). The racial composition of the dataset is detailed in Table 2.

Disease and medication data were structured according to the Observational Medical Outcomes Partnership (OMOP) standard data model[71,72]. Predictions were generated for 1,363 diseases, identified by training a knowledge graph on 5% of randomly selected drug-disease pairs, serving as a validation set for early stopping. This methodology does not extend to zero-shot performance evaluation across all 17,080 diseases, focusing instead on conditions with established indications. Disease names in the prediction dataset were aligned with SNOMED or ICD-10 codes and then mapped to OMOP concepts within the Mount Sinai data system. The analysis was restricted to diseases diagnosed in at least one patient, narrowing the focus to 480 conditions. Similarly, medication names were matched to DrugBank IDs, then to RxNorm IDs and OMOP concepts, limiting the scope to medications prescribed to at least one patient, resulting in 1,290 medications. Drug-disease pairs were further refined to those with at least one recorded instance of a patient being prescribed the drug for the disease, leading to a final count of 1,236 drugs and 470 diseases. Contingency tables were created for each drug-disease pair, and the Fisher exact function from the SciPy library[73] was employed to calculate 2-sided odds ratios and p-values for each pair. A two-sided Bonferroni correction was applied to the p-values using the statsmodels Python library's multi-test function[74], identifying statistically significant drug-disease pairs as those with $p < 0.005$.

Disease area	Number of diseases	Number of indications	Number of Contraindications
Diseases of cell proliferation	183	854	1007
Mental health diseases	56	213	1038
Cardiovascular diseases	104	300	3131
Diseases of anemia	15	55	545
Adrenal gland diseases	6	33	303
Autoimmune diseases	18	75	319
Metabolic disorders	54	68	523
Diabetes	3	102	364
Neurodegenerative diseases	16	123	134

Table 1: Statistics on disease-area-based dataset splits used to evaluate the zero-shot prediction of therapeutic use. Given all diseases in a given disease area, all indications and contraindications were removed from the dataset used to train machine learning models. Additionally, a fraction (5%) of the connections between biomedical entities to these diseases were removed from the therapeutics-centered knowledge graph. Disease-area splits were curated to evaluate model performance on diseases with limited molecular understanding and no existing treatments.

Racial group	Number of patients	Percent (%)
Asian	60,041	4.7
Black	162,102	12.7
White	534,305	42.0
Unknown	241,998	19.0
Other	273,639	21.5
Total number of patients	1,272,085	100.0

Table 2: Demographics of the electronic health record dataset at Mount Sinai Health System in New York City used to validate TxGNN's hypotheses on therapeutic use prediction.

References

1. Feigin, V. L. *et al.* Burden of neurological disorders across the us from 1990-2017: a global burden of disease study. *JAMA neurology* **78**, 165–176 (2021).

2. Vetter, N. Editor's choice. *British Medical Bulletin* **93**, 1–5 (2010).

3. Food, U. & Administration, D. Rare Disease Day 2021. https://www.fda.gov/news-events/fda-voices/rare-disease-day-2021-fda-shows-sustained-support-rare-disease-product-development-during-public (2023). [Online; accessed 19-September-2023].

4. Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. *Nature Reviews Drug Discovery* **18**, 41–58 (2019).

5. Abdelsayed, M., Kort, E. J., Jovinge, S. & Mercola, M. Repurposing drugs to treat cardiovascular disease in the era of precision medicine. *Nature Reviews Cardiology* **19**, 751–764 (2022).

6. Sahragardjoonegani, B., Beall, R. F., Kesselheim, A. S. & Hollis, A. Repurposing existing drugs for new uses: a cohort study of the frequency of FDA-granted new indication exclusivities since 1997. *Journal of Pharmaceutical Policy and Practice* **14** (2021).

7. Sardana, D. *et al.* Drug repositioning for orphan diseases. *Briefings in Bioinformatics* **12**, 346–356 (2011).

8. Jourdan, J.-P., Bureau, R., Rochais, C. & Dallemagne, P. Drug repositioning: a brief overview. *Journal of Pharmacy and Pharmacology* **72**, 1145–1151 (2020).

9. Chandak, P., Huang, K. & Zitnik, M. Building a knowledge graph to enable precision medicine. *Scientific Data* **10**, 67 (2023).

10. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347** (2015).

11. Zitnik, M., Feldman, M. W., Leskovec, J. *et al.* Evolution of resilience in protein interactomes across the tree of life. *Proceedings of the National Academy of Sciences* **116**, 4426–4433 (2019).

12. Ruiz, C., Zitnik, M. & Leskovec, J. Identification of disease treatment mechanisms through the multiscale interactome. *Nature Communications* **12**, 1–15 (2021).

13. Goh, K.-I. *et al.* The human disease network. *Proceedings of the National Academy of Sciences* **104**, 8685–8690 (2007).

14. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12**, 56–68 (2011).

15. Li, M. M., Huang, K. & Zitnik, M. Graph representation learning in biomedicine and healthcare. *Nature Biomedical Engineering* 1–17 (2022).

16. Gysi, D. M. *et al.* Network medicine framework for identifying drug-repurposing opportunities for covid-19. *Proceedings of the National Academy of Sciences* **118** (2021).

17. Cao, M. *et al.* Going the distance for protein function prediction: A new distance metric for protein interaction networks. *PLoS ONE* **8**, e76339 (2013).

18. Cheng, F. *et al.* Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nature Communications* **9** (2018).

19. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34**, i457–i466 (2018).

20. Guney, E., Menche, J., Vidal, M. & Barábasi, A.-L. Network-based in silico drug efficacy screening. *Nature Communications* **7**, 1–13 (2016).

21. Cheng, F., Kovács, I. A. & Barabási, A.-L. Network-based prediction of drug combinations. *Nature Communications* **10**, 1–11 (2019).

22. Fermaglich, L. J. & Miller, K. L. A comprehensive study of the rare diseases and conditions targeted by orphan drug designations and approvals over the forty years of the orphan drug act. *Orphanet Journal of Rare Diseases* **18**, 1–8 (2023).

23. Guney, E. Reproducible drug repurposing: When similarity does not suffice. In *Pacific Symposium on Biocomputing 2017*, 132–143 (World Scientific, 2017).

24. Avram, S. *et al.* DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Research* **49**, D1160–D1169 (2021).

25. Schlichtkrull, M. S., De Cao, N. & Titov, I. Interpreting graph neural networks for NLP with differentiable edge masking. *International Conference on Learning Representations* (2021).

26. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *NeurIPS* **30** (2017).

27. Wang, Q., Huang, K., Chandak, P., Zitnik, M. & Gehlenborg, N. Extending the nested model for user-centric xai: A design study on gnn-based drug repurposing. *IEEE Transactions on Visualization and Computer Graphics* **29**, 1266–1276 (2023).

28. Cao, M. *et al.* Going the distance for protein function prediction: a new distance metric for protein interaction networks. *PloS one* **8**, e76339 (2013).

29. Schlichtkrull, M. *et al.* Modeling relational data with graph convolutional networks. In *ESWC*, 593–607 (Springer, 2018).

30. Hu, Z., Dong, Y., Wang, K. & Sun, Y. Heterogeneous graph transformer (2020).

31. Wang, X. *et al.* Heterogeneous graph attention network (2019).

32. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* btz682 (2019).

33. Duran-Frigola, M. *et al.* Extending the small-molecule similarity principle to all levels of biology with the chemical checker. *Nature Biotechnology* **38**, 1087–1096 (2020).

34. Bickel, S., Brückner, M. & Scheffer, T. Discriminative learning under covariate shift. *Journal of Machine Learning Research* **10** (2009).

35. Schölkopf, B. *et al.* On causal and anticausal learning. *ICML* 1255–1262 (2012).

36. Niven, T. & Kao, H.-Y. Probing neural network comprehension of natural language arguments. *Proc. 57th Annual Meeting of the Association of Computational Linguistics* 4658–4664 (2019).

37. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine* **15**, e1002683 (2018).

38. Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nature Machine Intelligence* **2**, 665–673 (2020).

39. Agarwal, C., Queen, O., Lakkaraju, H. & Zitnik, M. Evaluating explainability for graph neural networks. *Scientific Data* **10** (2023).

40. Agarwal, C., Zitnik, M. & Lakkaraju, H. Probing GNN explainers: A rigorous theoretical and empirical analysis of gnn explanation methods. In *International Conference on Artificial Intelligence and Statistics*, 8969–8996 (2022).

41. Ying, Z., Bourgeois, D., You, J., Zitnik, M. & Leskovec, J. Gnnexplainer: Generating explanations for graph neural networks. *NeurIPS* **32** (2019).

42. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *ICML*, 3319–3328 (PMLR, 2017).

43. Wang, J. *et al.* Empower post-hoc graph explanations with information bottleneck: A pre-training and fine-tuning perspective. In *KDD*, 2349–2360 (2023).

44. Tukey, J. W. Comparing individual means in the analysis of variance. *Biometrics* 99–114 (1949).

45. Bomalaski, M. N., Claflin, E. S., Townsend, W. & Peterson, M. D. Zolpidem for the treatment of neurologic disorders: a systematic review. *Jama Neurology* **74**, 1130–1139 (2017).

46. Boisgontier, J. *et al.* Case report: Zolpidem's paradoxical restorative action: A case report of functional brain imaging. *Frontiers in Neuroscience* **17**, 1127542 (2023).

47. Sripad, P. *et al.* Effect of zolpidem in the aftermath of traumatic brain injury: an meg study. *Case reports in neurological medicine* **2020** (2020).

48. Landrum, M. J. *et al.* Clinvar: improvements to accessing data. *Nucleic acids research* **48**, D835–D844 (2020).

49. Javed, S. *et al.* Aldh1 & cd133 in invasive cervical carcinoma & their association with the outcome of chemoradiation therapy. *The Indian journal of medical research* **154**, 367 (2021).

50. Ghoussaini, M. *et al.* Open targets genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic acids research* **49**, D1311–D1320 (2021).

51. Goltsman, I. *et al.* Rosiglitazone treatment restores renal responsiveness to atrial natriuretic peptide in rats with congestive heart failure. *Journal of Cellular and Molecular Medicine* **23**, 4779–4794 (2019).

52. Bryan, P. M., Xu, X., Dickey, D. M., Chen, Y. & Potter, L. R. Renal hyporesponsiveness to atrial natriuretic peptide in congestive heart failure results from reduced atrial natriuretic peptide receptor concentrations. *American Journal of Physiology-Renal Physiology* **292**, F1636–F1644 (2007).

53. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2018).

54. Seetharaman, J. & Sarma, M. S. Chelation therapy in liver diseases of childhood: Current status and response. *World Journal of Hepatology* **13**, 1552 (2021).

55. Alsentzer, E. *et al.* Deep learning for diagnosing patients with rare genetic diseases. *medRxiv* 2022–12 (2022).

56. O'Connell, D. Neglected diseases. *Nature* **449**, 157–157 (2007).

57. Tambuyzer, E. *et al.* Therapies for rare diseases: therapeutic modalities, progress and challenges ahead. *Nature Reviews Drug Discovery* **19**, 93–111 (2020).

58. Zhang, A., Xing, L., Zou, J. & Wu, J. C. Shifting machine learning for healthcare from development to deployment and from models to data. *Nature Biomedical Engineering* 1–16 (2022).

59. Duffy, Á. *et al.* Development of a human genetics-guided priority score for 19,365 genes and 399 drug indications. *Nature Genetics* 1–9 (2024).

60. Cheng, J., Dasoulas, G., He, H., Agarwal, C. & Zitnik, M. GNNDelete: a general strategy for unlearning in graph neural networks. *International Conference on Learing Representations* (2023).

61. Huang, K., Jin, Y., Candes, E. & Leskovec, J. Uncertainty quantification over graph with conformalized graph neural networks. *Advances in Neural Information Processing Systems* **36** (2024).

62. Cai, C. J. *et al.* Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*, 1–14 (2019).

63. Macefield, R. How to specify the participant group size for usability studies: a practitioner's guide. *Journal of usability studies* **5**, 34–45 (2009).

64. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *ICML*, 1263–1272 (PMLR, 2017).

65. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 249–256 (2010).

66. Yang, B., Yih, W.-t., He, X., Gao, J. & Deng, L. Embedding entities and relations for learning and inference in knowledge bases. *ICLR* (2015).

67. Griggs, R. C. *et al.* Clinical research for rare disease: opportunities, challenges, and solutions. *Molecular Genetics and Metabolism* **96**, 20–26 (2009).

68. Boycott, K. M., Vanstone, M. R., Bulman, D. E. & MacKenzie, A. E. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics* **14**, 681–691 (2013).

69. Thomas, S. & Caplan, A. The orphan drug act revisited. *Jama* **321**, 833–834 (2019).

70. Lin, Y., Liu, Z., Sun, M., Liu, Y. & Zhu, X. Learning entity and relation embeddings for knowledge graph completion. In *AAAI* (2015).

71. Stang, P. E. *et al.* Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Annals of Internal Medicine* **153**, 600–606 (2010).

72. Klann, J. G., Joss, M. A., Embree, K. & Murphy, S. N. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. *PloS ONE* **14**, e0212463 (2019).

73. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods* **17**, 261–272 (2020).

74. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*, vol. 57 (2010).