ORIGINAL RESEARCH **Open access**

Self-Supervised Graph Representation Learning for Predicting Drug Repurposing Candidates for Rare Pediatric Cancers Using Protein-Protein Interaction Networks and Gene Expression Data

Luis Fernandez[1*], Marta Ruiz[1], Sergio Navarro[2]

Abstract

Rare pediatric cancers are difficult to treat due to their very low incidence, which limits drug development and makes experimental screening of therapies slow, costly, and dependent on scarce tumor samples. Traditional supervised machine learning approaches are also constrained by the lack of labeled drug–response data, while rich but unlabeled protein–protein interaction networks remain underutilized. We propose a self-supervised graph representation learning framework that integrates protein interaction networks with patient gene expression data to support drug repurposing. The model builds a heterogeneous graph of drugs, genes, diseases, and proteins, and uses a graph neural network trained with self-supervised objectives such as contrastive learning and masked prediction to learn molecular representations without labeled data. It is then fine-tuned on small pediatric cancer datasets. The framework enables prediction of candidate drug therapies by combining learned biological network representations with disease-specific expression profiles. This approach reduces reliance on large labeled datasets and allows adaptation to rare cancer contexts, offering a scalable strategy for computational drug repurposing in pediatric oncology.

Keywords Self-supervised learning, Graph neural networks, Drug repurposing, Rare pediatric cancers, Protein-protein interaction networks, Gene expression

*Correspondence:
Luis Fernandez
luis.fernandez@gmail.com
[1] Department of Intelligent Healthcare Analytics, University of Salamanca, Salamanca, Spain
[2] Department of AI Clinical Systems, University of Valladolid, Valladolid, Spain

Introduction

Rare pediatric cancers, defined as malignancies affecting fewer than 15 cases per 100,000 children annually, encompass devastating diseases such as neuroblastoma, medulloblastoma, osteosarcoma, Ewing sarcoma, and rhabdomyosarcoma that collectively represent a substantial burden of childhood mortality [1, 2]. Despite advances in understanding the genomic landscape of these tumors, the standard therapeutic armamentarium remains limited to surgery, chemotherapy, and radiation, modalities that carry significant long-term morbidity for developing children [3, 4].

The drug development pipeline for pediatric oncology faces fundamental economic barriers, as the small patient populations cannot support the decade-long, multi-billion-dollar investment required to bring a novel therapeutic agent to market [5]. Consequently, children with rare cancers frequently receive treatments originally developed for adult malignancies, often without the benefit of pediatric-specific clinical trials that would establish optimal dosing and efficacy [6].

Drug repurposing, the identification of new therapeutic indications for already-approved drugs with established

safety profiles, represents a pragmatic alternative to de novo drug development for rare pediatric cancers [7, 8]. Repurposed drugs can potentially reach patients more rapidly since pharmacokinetic properties, dosing parameters, and toxicity profiles have already been characterized, significantly reducing the time and cost associated with clinical translation [9]. Computational approaches to drug repurposing have traditionally employed signature matching methods that compare gene expression changes induced by drugs to those characterizing disease states, network-based algorithms that propagate information across biological interaction maps, and supervised machine learning classifiers trained on known drug-disease associations [10, 11]. However, these methods uniformly struggle when applied to rare cancers because labeled drug response data, which serves as the training signal for supervised models, is extremely sparse or entirely absent for pediatric malignancies [12, 13].

Protein-protein interaction networks capture the complex web of physical and functional relationships among proteins that orchestrate cellular behavior, providing a rich substrate for understanding how genetic alterations in cancer cells disrupt normal signaling pathways [14, 15]. These networks, compiled from decades of biochemical experiments and literature curation in resources such as STRING and BioGRID, contain tens of thousands of nodes and hundreds of thousands of edges, representing an expansive repository of biological knowledge [16]. Simultaneously, gene expression data derived from patient tumor samples offers a direct window into the transcriptional consequences of cancer-specific genomic alterations, revealing which pathways are aberrantly activated or suppressed in malignant cells [17, 18]. The integration of these two complementary data modalities, the static wiring diagram of protein interactions and the dynamic readout of transcriptional activity, holds particular promise for identifying therapeutic vulnerabilities in rare cancers where comprehensive drug screening data is unavailable [19].

This manuscript proposes a conceptual framework that leverages self-supervised graph representation learning to predict drug repurposing candidates for rare pediatric cancers by integrating protein-protein interaction networks with patient gene expression data. The central thesis is that a graph neural network pre-trained on unlabeled protein interaction data through self-supervised objectives can learn transferable molecular representations that generalize effectively to rare cancer types when fine-tuned on small gene expression datasets [20, 21]. The framework builds upon recent advances in graph contrastive learning, which have demonstrated that meaningful representations can emerge from simply maximizing agreement between differently augmented views of the same graph, without requiring any labels [22]. Following this introduction, we provide background context on rare pediatric cancers, drug repurposing approaches, protein-protein interaction networks, and self-supervised graph learning, then systematically present the framework architecture, graph construction methodology, pre-training strategies, prediction mechanisms, cancer-specific adaptations, evaluation strategy, and limitations.

Background

Rare pediatric cancers

Rare pediatric cancers include malignancies such as neuroblastoma arising from neural crest cells, medulloblastoma originating in the cerebellum, osteosarcoma affecting bone, Ewing sarcoma characterized by specific chromosomal translocations, and rhabdomyosarcoma derived from skeletal muscle progenitors, each presenting distinct molecular pathologies and clinical challenges [23, 24]. Standard treatment protocols for these diseases combine surgical resection where feasible, multi-agent chemotherapy regimens, and radiation therapy, approaches that have improved survival rates substantially over the past four decades but have now reached a plateau in effectiveness [25]. Relapse rates remain unacceptably high for metastatic and refractory disease, and survivors of childhood cancer frequently experience severe long-term sequelae including secondary malignancies, cardiotoxicity, endocrinopathies, and neurocognitive deficits resulting from the genotoxic effects of conventional therapies on developing tissues [26]. Recent large-scale genomic profiling initiatives have revealed that pediatric cancers harbor fewer somatic mutations than adult malignancies but are enriched in epigenetic alterations and developmental pathway dysregulation, suggesting that targeting these unique vulnerabilities could yield more effective and less toxic treatments [27].

Drug repurposing approaches

Computational drug repurposing encompasses a diverse methodological landscape including signature matching

approaches that identify drugs whose transcriptional effects oppose disease-associated expression changes, network-based methods that traverse biological interaction graphs to connect drugs to disease gene modules, and machine learning classifiers that predict novel drug-disease associations from heterogeneous data sources [10, 11]. Experimental repurposing strategies complement these computational approaches through high-throughput screening of approved drug libraries against cancer cell lines, genome-wide CRISPR knockout screens that identify genetic dependencies, and patient-derived organoid models that recapitulate tumor biology [7, 8]. Despite their demonstrated utility, both computational and experimental approaches share a fundamental limitation when applied to rare cancers: supervised methods require labeled training data comprising known effective and ineffective drug-cancer pairs, information that is extremely sparse for diseases affecting few patients [13, 28]. Transfer learning strategies that leverage knowledge from more common cancers have been explored, yet the distinct biology of pediatric malignancies compared to adult tumors complicates straightforward knowledge transfer [29].

Protein-protein interaction networks

Protein-protein interaction networks represent the complete set of physical contacts and functional associations among proteins within a cell, forming a structural scaffold that underlies virtually all biological processes from signal transduction to gene regulation [14]. These networks are constructed from diverse experimental evidence sources including yeast two-hybrid assays, affinity purification coupled with mass spectrometry, and literature-curated interactions documented in structured databases, with leading resources such as STRING integrating multiple evidence types into comprehensive interaction scores [15, 16]. The topology of protein interaction networks exhibits characteristic properties including scale-free degree distributions where a small number of hub proteins mediate numerous interactions, modular community structure reflecting functional complexes and pathways, and small-world architecture that enables efficient information flow [17]. In cancer biology, proteins encoded by mutated genes often occupy central positions within interaction networks, and drugs targeting these network hubs can disrupt multiple oncogenic pathways simultaneously, an observation that motivates network-based approaches to therapeutic target identification [18].

Self-supervised graph learning

Self-supervised graph learning has emerged as a transformative paradigm for representation learning on graph-structured data, enabling models to extract informative features from unlabeled graphs by solving carefully designed pretext tasks that derive supervisory signals from the graph structure itself [19, 20]. Contrastive methods such as Deep Graph Infomax maximize mutual information between local node representations and global graph summaries, while GraphCL and its variants generate multiple augmented views through stochastic edge perturbation, node feature masking, and subgraph sampling, then pull together representations of related views while pushing apart unrelated ones [21, 22]. Generative self-supervised approaches including GraphMAE follow the masked autoencoding principle, where portions of node features or graph structure are masked and the model learns to reconstruct the missing information from the visible context [15]. Predictive pretext tasks encompass edge prediction, which trains encoders to distinguish true connections from randomly sampled negative edges, and node property prediction, where artificially masked attributes must be recovered from neighborhood information, collectively forcing the GNN to capture meaningful topological and semantic patterns without human annotation [16].

Framework Overview

High-level architecture

The proposed framework follows a two-stage architecture wherein a heterogeneous biological graph encompassing proteins, drugs, and diseases is first subjected to self-supervised pre-training using a graph neural network encoder, after which the pre-trained model undergoes fine-tuning on rare pediatric cancer gene expression data to generate drug repurposing predictions [17]. During the pre-training phase, the GNN learns to produce dense vector representations for all nodes in the graph by solving multiple self-supervised pretext tasks that require the model to understand both local neighborhood structure and global graph topology [1]. This pre-training leverages the vast quantity of unlabeled protein interaction data available from public databases, completely circumventing the need for drug response labels that are unavailable for rare cancers [2]. In the fine-tuning phase, gene expression signatures from a small cohort of patients with a specific rare pediatric cancer are incorporated as node features, and the model

weights are updated to align the learned representations with the transcriptional state characteristic of that malignancy, ultimately enabling the ranking of candidate repurposing drugs based on predicted therapeutic efficacy [3].

Figure 1 presents the proposed self-supervised graph representation learning framework, showing how heterogeneous biomedical graph construction, unlabeled pre-training, rare-cancer fine-tuning, candidate drug ranking, and validation are organized into a directional translational architecture.

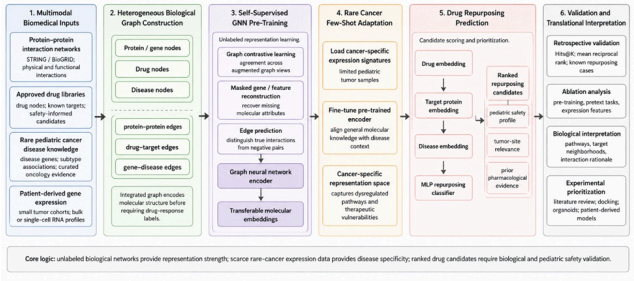

Figure 1. Self-Supervised Graph Representation Learning Framework for Rare Pediatric Cancer Drug Repurposing.

Core assumptions

The framework rests upon several foundational assumptions regarding data availability and biological plausibility that must be satisfied for successful implementation [4]. We assume the availability of a large, reasonably comprehensive protein-protein interaction network covering a substantial fraction of the human proteome, such as those provided by STRING or BioGRID, which serves as the primary source of structural information for self-supervised pre-training [5]. Additionally, we assume access to gene expression profiles from tumor samples of the target rare pediatric cancer, even if only from a small patient cohort, which provide the disease-specific transcriptional signal necessary for fine-tuning the pre-trained model [6]. The framework further assumes the existence of known drug-target interactions for a library of approved drugs, enabling the connection of drug nodes to the protein interaction network through their established molecular targets, and serving as a validation resource for evaluating prediction quality [7].

Design principles

Three core design principles guide the architecture and implementation decisions throughout this framework [8]. The principle of self-supervision dictates that all pre-training objectives must derive their supervisory signal from the graph data itself, without requiring any externally provided labels, thereby enabling the model to exploit the full scale of available protein interaction data without being limited by annotation sparsity [9]. Transferability constitutes the second principle, ensuring that representations learned during pre-training on the general protein interaction network remain useful when the model is subsequently adapted to the specific context of a rare pediatric cancer through fine-tuning on limited expression data [19]. The third principle emphasizes interpretability, requiring that the model's predictions be attributable to specific biological pathways and molecular interactions, enabling researchers to understand the mechanistic rationale underlying each repurposing recommendation and facilitating experimental validation [14,15].

Table 1 clarifies how each architectural component responds to a specific data-scarcity barrier while contributing a distinct function to rare pediatric cancer drug repurposing.

Table 1. Conceptual Alignment between Framework Components, Data-Scarcity Barriers, and Repurposing Functions

Framework component	Data-scarcity barrier addressed	Conceptual function in the model	Added for r pedia canc
Heterogeneous biological graph	Lack of labeled drug–cancer response pairs	Converts fragmented biomedical knowledge into a shared representational structure	Enal learnin protein and di relation when c specific are sp
Self-supervised graph pre-training	Insufficient supervised training labels	Learns molecular embeddings from graph structure itself	Us abun unlab biolo netw instead clini respons

Contrastive learning	Weak generalization from small cohorts	Encourages robust representations across perturbed graph views	Impr stability patient- datase sm
Edge prediction	Limited knowledge of indirect drug–disease mechanisms	Forces the encoder to learn interaction topology	Supp infere thro netw neighbo and pat
Gene masking	Missing or noisy molecular features	Learns functional dependence among genes from context	Helps pathwa signals incom expre da
Few-shot fine-tuning	Small rare-cancer expression cohorts	Adapts general embeddings to cancer-specific transcriptional states	Allo disea spe predi withou drug sc data
Repurposing classifier	Need for candidate prioritization rather than raw embeddings	Scores drug–target–disease compatibility	Prod rank candida downs valid
Interpretability layer	Translational uncertainty of computational predictions	Links predictions to pathways, targets, and interaction neighborhoods	Supp biolo plaus asses bef experi

Graph Construction

Node types

The heterogeneous biological graph comprises three primary node categories that collectively represent the molecular entities relevant to drug repurposing [20]. Protein or gene nodes constitute the largest node population, with each node corresponding to a human protein-coding gene and its associated protein product, collectively forming the backbone of the network through which drug effects and disease perturbations propagate [16]. Drug nodes represent approved pharmaceutical compounds drawn from curated databases of medications with established safety profiles, each annotated with known protein targets that connect them to specific protein nodes within the interaction network [10, 11]. Disease nodes correspond to rare pediatric cancer types and subtypes, linked to protein nodes through curated gene-disease associations derived from genomic studies, literature mining, and clinical databases that document genes recurrently altered in specific malignancies [23, 24].

Edge types

Three distinct edge types establish relationships among the heterogeneous node categories, each encoding a specific biological or pharmacological association [21]. Protein-protein edges capture both physical binding interactions identified through biochemical experiments and functional associations inferred from co-expression patterns, phylogenetic profiles, and literature co-occurrence, weighted by confidence scores that reflect the strength of supporting evidence [14, 15]. Drug-protein edges represent known drug-target relationships wherein a pharmaceutical compound binds to and modulates the activity of a specific protein, with edge weights potentially encoding binding affinity, mechanism of action, or therapeutic relevance [7, 8]. Gene-disease edges connect cancer types to the genes known to be somatically mutated, differentially expressed, or epigenetically altered in those malignancies, derived from resources such as the TARGET database, the Gabriella Miller Kids First initiative, and published pediatric cancer genomic studies [25-27].

Node features

Initial node features encode diverse molecular properties that provide the GNN encoder with rich input representations from which to learn higher-level abstractions [22]. Gene expression values measured from patient tumor samples, obtained through bulk RNA sequencing or single-cell transcriptomics platforms, serve as continuous feature vectors for protein nodes in the fine-tuning phase, capturing the transcriptional consequences of oncogenic alterations [17, 18]. Where expression data is unavailable during pre-training, protein sequence embeddings generated by pre-trained language models

such as ProtBERT provide alternative feature vectors that encode evolutionary and structural properties in a format amenable to neural network processing [22]. Drug nodes are initialized with molecular fingerprint representations including Morgan fingerprints that encode circular substructure information and MACCS keys that capture predefined structural features, enabling the model to relate chemical structure to biological activity [9, 10].

Self-Supervised Pre-Training

Graph contrastive learning

The primary self-supervised objective employs graph contrastive learning, wherein augmented views of the protein-protein interaction network are generated through stochastic transformations, and the GNN encoder is trained to maximize agreement between representations of the same node across different views while minimizing agreement between representations of distinct nodes [23]. Augmentation strategies include node dropout, which randomly removes a fraction of nodes and their incident edges to create a perturbed view that tests the encoder's robustness to missing information, and edge perturbation, which randomly adds or removes edges to simulate noise and incompleteness in the interaction network [21, 22]. A contrastive loss function, implemented as normalized temperature-scaled cross-entropy, pulls together positive pairs derived from the same node while pushing apart negative pairs sampled from different nodes, a training signal that encourages the encoder to learn representations invariant to the specific perturbations applied [15]. This approach is motivated by the observation that proteins participating in similar biological processes or residing in the same network neighborhood should maintain coherent representations despite arbitrary structural variations introduced during augmentation [19].

Edge prediction pretext task

A complementary edge prediction task directly trains the GNN to capture the topological organization of the protein interaction network by requiring it to distinguish true protein-protein interactions from randomly sampled non-interacting protein pairs [24]. During pre-training, a subset of existing edges is masked from the graph, and the encoder must predict the presence or absence of these masked connections based on the learned node representations, a binary classification task that forces the model to internalize the structural patterns governing protein interactions [14, 16]. This pretext task is particularly valuable for drug repurposing because therapeutic effects often propagate through networks via paths of interacting proteins, and an encoder that accurately understands interaction topology is better positioned to identify connections between drug targets and disease-associated proteins [2, 3]. The edge prediction objective is implemented as a sigmoid-activated scoring function applied to concatenated node representations, optimized through binary cross-entropy loss against the true edge labels [20].

Gene masking pretext task

The gene masking pretext task operates on node features rather than graph structure, randomly masking a fraction of gene expression or sequence embedding dimensions for selected protein nodes and training the GNN to reconstruct the original feature values from the visible context [25]. This objective is inspired by the masked language modeling paradigm that has proven highly effective in natural language processing and has been adapted to graph domains through methods that reconstruct masked node attributes using neighborhood information [15, 19]. By learning to infer missing functional annotations from the surrounding network context, the encoder develops an understanding of the functional relationships among proteins that transcends simple topological proximity, capturing more nuanced biological principles such as pathway co-membership and compensation mechanisms [12, 13]. The gene masking task synergizes with graph contrastive learning and edge prediction, as each pretext objective encourages the learning of complementary aspects of biological organization, and their combination during pre-training yields representations that comprehensively encode both structural and functional properties of the protein interaction landscape [21].

Drug Repurposing Prediction

Fine-tuning on rare cancer data

Following self-supervised pre-training on the comprehensive protein-protein interaction network, the framework enters a fine-tuning phase where the GNN encoder is adapted to the specific biological context of a rare pediatric cancer using gene expression data derived from patient tumor samples [26]. The fine-tuning dataset, which may comprise as few as ten to one hundred expression profiles from a specific malignancy such as

Ewing sarcoma or atypical teratoid rhabdoid tumor, provides a transcriptional snapshot of pathway dysregulation that characterizes that cancer type, and these expression values are loaded as node features for the protein nodes within the heterogeneous graph [27, 28]. During fine-tuning, all GNN parameters are updated through backpropagation using a supervised objective that trains the model to associate the expression signatures of known disease-associated genes with the corresponding disease nodes, effectively aligning the pre-trained molecular representations with the specific transcriptional state of the rare cancer [1, 2]. The limited size of rare cancer expression datasets necessitates careful regularization strategies including dropout, weight decay, and early stopping to prevent overfitting, while the rich representations inherited from self-supervised pre-training provide a robust initialization that enables effective learning even from few examples, embodying principles of few-shot transfer learning applied to the biomedical domain [3, 29].

Repurposing scoring

The drug repurposing prediction mechanism computes a therapeutic efficacy score for each candidate drug by aggregating information across multiple graph-based signals that connect drug nodes to disease nodes through the protein interaction network [27]. The scoring function integrates the learned embeddings of the drug node, the target protein nodes, and the disease node through a multi-layer perceptron classifier that outputs a probability estimate representing the likelihood that the drug would demonstrate therapeutic activity against the specified rare pediatric cancer [10, 11]. The model ranks all candidate drugs in descending order of their predicted efficacy scores, producing a prioritized list from which the highest-scoring compounds can be selected for further investigation through literature review, in silico binding analysis, or ultimately experimental validation in appropriate preclinical models [7, 8]. Importantly, the ranking mechanism can incorporate additional filtering criteria such as pediatric safety profiles, blood-brain barrier penetration for central nervous system tumors, and existing evidence from adult cancer indications, enabling domain experts to apply clinical judgment alongside computational predictions when selecting candidates for further evaluation [23, 24].

Rare Pediatric Cancer Focus

Few-shot fine-tuning

The framework explicitly addresses the data scarcity challenge inherent to rare pediatric cancers through a few-shot fine-tuning strategy that leverages the representations learned during self-supervised pre-training to enable generalization from extremely limited labeled examples [28]. Few-shot learning, a paradigm where models are expected to perform well on new tasks given only a handful of training instances, is naturally aligned with the rare cancer setting where comprehensive drug screening data does not exist and may never be feasible to generate due to limited patient material [3, 4]. The pre-trained GNN encoder functions as a feature extractor that has already internalized general principles of protein interaction network organization, pathway structure, and drug-target pharmacology, meaning that fine-tuning requires the model only to adapt these general representations to the specific molecular context of the target rare cancer rather than learning biological principles from scratch [5, 6]. This approach draws upon the success of pre-training strategies in other domains, where models trained on large general corpora through self-supervision consistently outperform those trained from random initialization when labeled data for downstream tasks is scarce, a pattern we expect to hold for biological network analysis [19, 21].

Cross-cancer transfer

The framework incorporates mechanisms for cross-cancer transfer learning, wherein knowledge gained from cancers with more abundant data, such as common adult malignancies with extensive drug screening resources, can be transferred to inform predictions for rare pediatric cancers [29]. Domain adaptation techniques bridge the gap between source cancer types where labeled drug response data is available and target rare pediatric cancers where only gene expression profiles exist, by aligning the representation spaces learned for these different biological contexts through adversarial training or distribution matching objectives [9, 13]. The framework further enables the construction of cancer-specific subgraphs within the larger heterogeneous graph, where disease nodes for different malignancies share the same underlying protein interaction infrastructure but connect to distinct sets of disease-associated genes, allowing the model to recognize both shared and cancer-specific therapeutic mechanisms [15, 16]. This architecture is particularly relevant for pediatric cancers that may share molecular features with adult malignancies, such as pediatric gliomas harboring histone mutations that affect pathways also dysregulated in adult glioblastoma, while simultaneously respecting the

unique developmental biology that distinguishes childhood tumors from their adult counterparts [25, 26].

Evaluation Strategy

Table 2 provides a translational validation matrix that distinguishes computational ranking performance from biological plausibility, pediatric feasibility, and experimental evidence.

Table 2. Validation Matrix for Translating Graph-Based Repurposing Predictions into Pediatric Oncology Evidence

Validation layer	Primary question answered	Suitable evidence source	Recomme metric outpu
Retrospective ranking validation	Can the model recover known or plausible repurposing successes?	Clinical trials, case reports, pediatric oncology literature	Hits@K, I Recipro Rank enrichme know candida
Baseline comparison	Does self-supervised graph learning outperform simpler methods?	Signature matching, random walk, supervised classifiers, network proximity models	Compara rankir performa and calibı
Ablation of pre-training	Is unlabeled graph pre-training essential?	Model variants with random initialization	Performa drop af removing trainir
Ablation of pretext tasks	Which self-supervised objectives contribute most?	Variants removing contrastive learning, edge prediction, or masking	Compon level performa chang
Expression-data ablation	Does rare-cancer gene expression improve disease specificity?	Models with and without patient-derived expression features	Change disease-sp ranking q
Biological plausibility review	Are predicted drugs connected to credible mechanisms?	Pathway enrichment, target neighborhoods, literature mining	Mechan explana map
Pediatric feasibility screening	Are top-ranked candidates suitable for children?	Pediatric dosing data, toxicity profiles, prior pediatric use	Safety-info candida shortli
Experimental validation	Do candidates demonstrate biological activity in relevant models?	Patient-derived cell lines, organoids, xenografts, CRISPR-informed assays	Viabili reducti pathwa modulat dose-resp eviden

Retrospective validation

Retrospective validation assesses the framework's ability to recover known drug repurposing successes for rare pediatric cancers documented in clinical trials or case reports, such as retinoic acid for acute promyelocytic leukemia or imatinib for Philadelphia chromosome-positive acute lymphoblastic leukemia [1, 3, 7, 8]. Standard metrics including Hits@K and Mean Reciprocal Rank quantify whether known effective drugs rank among the highest-scoring predictions, providing performance indicators comparable against baseline methods [10, 11]. This approach acknowledges the absence of prospective data while offering a meaningful utility signal, though known successes may be biased toward heavily-studied drugs [23].

Ablation studies

A comprehensive ablation strategy measures performance degradation when framework components are removed, attributing predictive power to specific architectural choices [14, 15]. Key conditions include replacing self-supervised pre-training with random initialization, removing individual pretext tasks to assess complementarity, and eliminating gene expression features to evaluate their contribution beyond network topology [19, 21, 22]. These controlled comparisons identify essential elements and guide practical implementation decisions [16].

Biological validation

Computational predictions require biological validation through a tiered approach: in silico methods employing molecular docking and literature mining, followed by in vitro testing using patient-derived cell lines or organoid models where resources permit [7, 9, 24-27]. The framework produces interpretable predictions with attention mechanisms identifying specific protein interactions driving each recommendation, enabling biologists to evaluate mechanistic rationale before committing to experimental follow-up [17, 18].

Limitations

Technical limitations

The framework relies on protein-protein interaction networks that remain incomplete and biased toward well-studied proteins, while drug-target databases show similar skew toward characterized target classes like kinases and G protein-coupled receptors [10, 12, 14, 16]. Bulk tumor gene expression conflates signals from malignant cells, stromal fibroblasts, and immune infiltrates, potentially reflecting microenvironment rather than cancer cell-intrinsic vulnerabilities [17, 18]. Graph construction hyperparameters substantially impact performance yet are difficult to optimize in low-data regimes [20, 21].

Clinical limitations

Pediatric cancers exhibit unique developmental biology with distinct mutational spectra featuring fewer point mutations and more copy number alterations, plus epigenetic dysregulation not recapitulated in adult cancers from which most drug response data derives [23-27]. The framework cannot anticipate off-target toxicities of particular concern in children, where drugs well-tolerated in adults may cause developmental toxicity, growth impairment, or long-term organ damage [28]. Computational predictions require extensive preclinical validation, and the pathway from ranked candidates to pediatric clinical trials remains lengthy, expensive, and uncertain [29].

Conclusion

This manuscript presented a self-supervised graph representation learning framework for predicting drug repurposing candidates in rare pediatric cancers, integrating protein-protein interaction networks with patient gene expression data through a heterogeneous graph neural network pre-trained via contrastive learning, edge prediction, and gene masking. The framework addresses the fundamental data scarcity barrier by learning molecular representations from unlabeled protein interactions, then adapting through few-shot fine-tuning on small cancer-specific expression datasets. Key advantages include principled alignment with available data resources, efficient use of limited labeled examples, and transferability across cancer types through shared network infrastructure.

Important limitations include incomplete protein interaction maps, biases toward well-studied targets, cellular heterogeneity in bulk expression data, and the distinct developmental biology of pediatric malignancies that may resist knowledge transfer from adult contexts. Computational predictions require extensive experimental validation, and the path to clinical trials remains lengthy. Implementation on public resources such as STRING, DepMap, TARGET, and the Gabriella Miller Kids First initiative would enable initial assessment of practical utility. This framework establishes a foundation connecting self-supervised graph learning with the pressing need for new pediatric cancer treatments, respecting both available data richness and disease-specific resource constraints.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 19 Dec 2024 Revised: 19 Feb 2025 Accepted: 16 Apr 2025
Published online: 20 July 2025

References

Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F, et al. deepDR: a network based deep learning approach to in silico drug repositioning. Bioinformatics. 2019;35(24):5191-8.

Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife. 2017;6:e26726.

Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics. 2018;34(13):i457-i466.

Cai L, Lu C, Xu J, Meng Y, Wang P, Fu X, et al. Drug repositioning based on the heterogeneous information fusion graph convolutional network. Brief Bioinform. 2021;22(6):bbab319.

Cui C, Ding X, Wang D, Chen L, Xiao F, Xu T, et al. Drug repurposing against breast cancer by integrating drug exposure expression profiles and drug-drug links based on graph neural network. Bioinformatics. 2021;37(18):2930-7.

Hsieh K, Wang Y, Chen L, Zhao Z, Savitz S, Jiang X, et al. Drug repurposing for COVID-19 using graph neural network and harmonizing multiple evidence. Sci Rep. 2021;11(1):23179.

Doshi S, Chepuri SP. A computational approach to drug repurposing using graph neural networks. Comput Biol Med. 2022;150:105992.

Yao K, Wang X, Li W, Zhu H, Jiang Y, Li Y, et al. Semi supervised heterogeneous graph contrastive learning for drug-target interaction prediction. Comput Biol Med. 2023;163:107199.

Jin S, Zhang Y, Yu H, Lu M. SADR: self supervised graph learning with adaptive denoising for drug repositioning. IEEE/ACM Trans Comput Biol Bioinform. 2024;21(2):265-77.

Gharizadeh A, Abbasi K, Ghareyazi A, Mofrad MR, Rabiee HR. HGTDR: advancing drug repurposing with heterogeneous graph transformers. Bioinformatics. 2024;40(7):btae349.

Li W, Ma W, Yang M, Tang X. Drug repurposing based on the DTD-GNN graph neural network: revealing the relationships among drugs, targets and diseases. BMC Genomics. 2024;25(1):584.

Ceskoutsé RF, Bomgni AB, Zanfack DR, Agany DD, Thomas BB, Zohim EG, et al. Heterokgrep: heterogeneous knowledge graph based drug repositioning. Knowl Based Syst. 2024;305:112638.

Cui H, Duan M, Bi H, Li X, Hou X, Zhang Y, et al. Heterogeneous graph contrastive learning with gradient balance for drug repositioning. Brief Bioinform. 2025;26(1):bbae650.

Hu W, Liu B, Gomes J, Zitnik M, Liang P, Pande V, et al. Strategies for pre training graph neural networks. arXiv preprint arXiv:1905.12265. 2019.

Veličković P, Fedus W, Hamilton WL, Liò P, Bengio Y, Hjelm RD, et al. Deep graph infomax. arXiv preprint arXiv:1809.10341. 2018.

You Y, Chen T, Sui Y, Chen T, Wang Z, Shen Y, et al. Graph contrastive learning with augmentations. Adv Neural Inf Process Syst. 2020;33:5812-23.

Zhu Y, Xu Y, Yu F, Liu Q, Wu S, Wang L, et al. Graph contrastive learning with adaptive augmentation. In: Proc Web Conf. 2021. p. 2069-80.

Liu Y, Jin M, Pan S, Zhou C, Zheng Y, Xia F, et al. Graph self supervised learning: a survey. IEEE Trans Knowl Data Eng. 2022;35(6):5879-900.

Xie Y, Xu Z, Zhang J, Wang Z, Ji S. Self supervised learning of graph neural networks: a unified review. IEEE Trans Pattern Anal Mach Intell. 2022;45(2):2412-29.

Zhu Y, Xu Y, Yu F, Liu Q, Wu S, Wang L, et al. Deep graph contrastive representation learning. arXiv preprint arXiv:2006.04131. 2020.

Jing M, Zhu Y, Zang T, Yu J, Tang F. Graph contrastive learning with adaptive augmentation for recommendation. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham: Springer; 2022. p. 590-605.

Wang Y, Wang J, Cao Z, Barati Farimani A. Molecular contrastive learning of representations via graph neural networks. Nat Mach Intell. 2022;4(3):279-87.

Wong M, Mayoh C, Lau LM, Khuong Quang DA, Pinese M, Kumar A, et al. Whole genome, transcriptome and methylome profiling enhances actionable target discovery in high risk pediatric cancer. Nat Med. 2020;26(11):1742-53.

Gröbner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood cancers. Nature. 2018;555(7696):321-7.

Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature. 2018;555(7696):371-6.

Newman S, Nakitandwe J, Kesserwan CA, Azzato EM, Wheeler DA, Rusch M, et al. Genomes for kids: the scope of pathogenic mutations in pediatric cancer revealed by comprehensive DNA and RNA sequencing. Cancer Discov. 2021;11(12):3008-27.

Van Tilburg CM, Pfaff E, Pajtler KW, Langenberg KP, Fiesel P, Jones BC, et al. The pediatric precision oncology INFORM registry: clinical outcome and benefit for patients with very high evidence targets. Cancer Discov. 2021;11(11):2764-79.

Berlanga P, Pierron G, Lacroix L, Chicard M, Adam de Beaumais T, Marchais A, et al. The European MAPPYACTS trial: precision medicine program in pediatric and adolescent patients with recurrent malignancies. Cancer Discov. 2022;12(5):1266-81.

Lau LM, Khuong Quang DA, Mayoh C, Wong M, Barahona P, Ajuyah P, et al. Precision guided treatment in high risk pediatric cancers. Nat Med. 2024;30(7):1913-22.