Path-based reasoning for biomedical knowledge graphs with BioKGC

Yue Hu[1,2], Svitlana Oleshko[1,5,'], Samuele Firmani[1,'], Zhaocheng Zhu[3,4], Hui Cheng[5], Maria Ulmer[1,2], Matthias Arnold[1,2], Maria Colomé-Tatché[1,2,6], Jian Tang[3,7,8], Sophie Xhonneux[3,4,'], Annalisa Marsico[1,'*]

['equal contribution.

[1]Computational Health Center, Helmholtz Center Munich, Ingolstaedter Landstrasse 1, Neuherberg, 85764, Bavaria, Germany.

[2]School of Life Sciences, Technical University of Munich, Alte Akademie 8, Freising, 85354, Bavaria, Germany.

[3]Department, Mila - Québec AI Institute, 6666 St-Urbain, Montréal, QC H2S 3H1, Quebec, Canada.

[4]Department, Université de Montréal, 2900, boul. Édouard-Montpetit, Montréal, QC H3T 1J4, Quebec, Canada.

[5]School of Computation, Information and Technology, Technical University of Munich, Arcisstrasse 21, Munich, 80333, Bavaria, Germany.

[6]Faculty of Biology, Ludwig-Maximilians University, Grosshaderner Str. 2, Planegg-Martinsried, 82152, Bavaria, Germany.

[7]Department, CIFAR AI Chair, 661 University Ave, Toronto, ON M5G 1M1, Ontario, Canada.

[8]Department, HEC Montréal, 3000 Chem. de la Côte-Sainte-Catherine, Montréal, QC H3T 2A7, Quebec, Canada.

*Corresponding author(s). E-mail(s): annalisa.marsico@helmholtz-munich.de;
Contributing authors: yue.hu@helmholtz-munich.de; svitlana.oleshko@helmholtz-munich.de; samuele.firmani@helmholtz-munich.de; zhaocheng.zhu@umontreal.ca; hui.cheng@helmholtz-munich.de; maria.colometatche@helmholtz-munich.de; jian.tang@hec.ca; sophie.xhonneux@mila.quebec;

Abstract

Understanding complex interactions in biomedical networks is crucial for advancements in biomedicine. Traditional link prediction (LP) methods, using similarity metrics like Personalized PageRank, are limited in capturing the complexity of biological networks. Recently, representation-based learning techniques have emerged, mapping nodes to low-dimensional embeddings to enhance prediction accuracy. However, these methods often face challenges with interpretability and scalability in large, complex networks. Based on a representation of biological systems as knowledge graphs (KGs), which encode entities and their relationships as triplets, we propose here BioKGC, a novel graph neural network framework which builds upon the Neural Bellman-Ford Network (NBFNet). It addresses the limitations of previous methods by utilizing path-based reasoning for LP in biomedical knowledge graphs (KGs). Unlike node-embedding learning frameworks that optimize the embedding space based on single triplets, BioKGC learns representations between nodes by considering all relations along paths. This approach enhances prediction accuracy and interpretability, allowing for the visualization of influential paths and facilitating the validation of biological plausibility. BioKGC leverages a background regulatory graph (BRG) for enhanced message passing and implements a stringent negative sampling strategy to improve learning precision. In evaluations across various LP tasks — gene function annotation, drug-disease interaction prediction, synthetic lethality prediction, and lncRNA-mRNA regulatory relationship inference — BioKGC consistently outperformed state-of-the art methods. BioKGC outperformed knowledge graph embedding and GNN-based methods in gene function prediction, especially with BRG information. We demonstrated that BioKGC effectively predicts drug-disease interactions in zero-shot learning scenarios, surpassing state-of-the-art models like TxGNN. Additionally, BioKGC demonstrated robust performance in synthetic lethality prediction and the capacity for scoring novel lncRNA-mRNA interactions, showcasing its versatility in diverse biomedical applications. One of BioKGC's key advantages is its interpretability, enabling researchers to trace prediction paths and gain insights into molecular mechanisms. Combined with its use of regulatory information for message passing, BioKGC is a powerful tool for predicting complex biological interactions, making it valuable for drug discovery and personalized medicine.

Keywords: biomedical knowledge graph, link prediction, graph neural network

1 Introduction

Biological entities interact in complex ways, crucial for sustaining life in living systems [1, 2]. Understanding these interactions is central to systems biology, with network analysis playing a key role [3]. Biological networks are represented as graphs, where nodes can represent genes, proteins, diseases and more, and edges denote associations between them. Edges in a biological graph between genes can signify co-regulation or causal relationship (regulatory network) [4, 5], physical interactions (in protein-protein

interaction networks (PPI) [6, 7]), as well as the association of genes with diseases (disease gene networks [8, 9]), among many.

Despite increasing high-throughput experiments, our grasp of biological networks is incomplete, leaving many interactions undiscovered. Due to the expense and time involved in wet lab experiments, computational methods such as link prediction (LP) are vital for inferring missing or potential associations within these networks based on the underlying topology [10]. LP is applied across network biology for diverse tasks ranging from predicting protein interactions over inferring gene regulatory networks to the exploring of pathways [11]. By revealing hidden connections, LP facilitates the discovery of biomarkers, drug targets, and insights into biological interactions [12, 13]. To predict potential relationships between unconnected nodes, one prevalent method uses similarity metrics from traditional graph analysis, such as Personalized PageRank, Jaccard or Katz index [14–16]. These metrics have been used for predicting disease-gene associations [17], including ncRNAs-disease relationships, and drug-disease associations [18].

While traditional graph metrics have been successful in biological link prediction, representation-based learning offers greater expressiveness for capturing the nuances and complexity of nodes in a graph. Nodes are mapped to low-dimensional vector representations called embeddings using shallow and deep non-linear transformations. Optimized embeddings position nodes with similar network neighborhoods closely in the embedding space, so that links between nodes can be predicted based on their similarity in this space [19]. Methods include matrix factorization-based (e.g. Mashup [20]) and random walk-based approaches (e.g., DeepWalk [21], Node2Vec [22], struc2Vec). Network embedding techniques have found success in diverse domains, including drug repurposing, adverse drug reaction prediction, gene function prediction, and protein-protein interaction network completion, among others [23–26].

Ruiz et al. introduced the multiscale interactome, a node-embedding method that integrates disease-associated proteins, drug targets, and biological functions to understand disease treatment mechanisms beyond direct drug targets, using biased random walks for predictions [27]. Similarly, GeneWalk predicts gene functions in specific biological contexts through network representation learning, constructing low-dimensional embeddings of gene-gene networks and Gene Ontology terms via random walks [28]. Hu et al. contextualize COVID-19 disease-risk genes by creating a multi-modal network of genes, polygenic risk scores (PRS) for diseases, and their connections [29]. Using DeepWalk, they generate node sequences from random walks and train a neural network to predict target nodes, similar to the skip-gram model [30, 31]. The resulting embeddings, assessed via cosine similarity, uncover associations between COVID-19 genes, other conditions, co-morbidities, and genetic predispositions to COVID-19 severity.

As opposed to the shallow learning approaches, methods such as Graph Convolutional Networks (GCNs) [32], Graph Autoencoders (GAEs) [33] and GraphSAGE [34] learn node embeddings from graph data using deep neural networks, by aggregating node messages from neighbors and learning a representation which reflects the neighborhood. OhmNet uses neural architectures to learn node embeddings in a multi-layer hierarchical network, representing molecular interactions across human tissues. This

multiscale model predicts diverse phenotypes, promoting feature sharing among proteins with similar network neighborhoods and tissue activations [25]. Decagon models polypharmaceutical side effects with GCNs by constructing a multimodal graph of protein-protein, drug-protein, and drug-drug interactions. Using an encoder-decoder approach, it enables multirelational link prediction, capturing various side effect types [26].

Early biological interaction models used basic networks, or uni-relational graphs, which struggled to capture the semantics of various entity associations. For example, protein-protein interaction networks lost distinctions between interaction types like inhibition and activation. Recent efforts turned to heterogeneous multi-relational networks, known as knowledge graphs (KGs), to better represent biological complexities. KGs use subject, predicate and object (SPO) triples to model facts, where subjects and objects are entities and predicates are relations that carry different association semantics. KG research is gaining interest for applications like question answering and information retrieval. Predicting missing links in KGs is a key challenge, involving modeling and inferring connectivity patterns based on observed facts, therefore the problem of knowledge graph completion can be transformed into estimating the missing components of triplets. KG research is gaining interest for applications like question answering and information retrieval. A key challenge is predicting missing links by modeling and inferring connectivity patterns from observed facts, transforming knowledge graph completion into estimating the missing components of triplets.

To address these constraints, Knowledge Graph Embedding (KGE) is effectively used in various KG inference tasks, as it learns low-rank representations of entities and relations of triplet-form graph data, preserving graph structure and specifically encoding the semantics of relations. Randomly initialized node embeddings are updated iteratively by optimizing a training loss that maximizes a score for positive triplets while minimizing those for sampled corrupted (negative) triplets [35].

Representative methods for KGE include TransE [35], DistMult [36], ComplEx [37], and RotatE [38]. TransE interprets relationships as translations in low-dimensional embeddings, suitable for hierarchical relationships like those in the Gene Ontology. DistMult aims to model symmetry patterns between entities in KGs using a neural embedding approach. ComplEx extends DistMult by introducing complex valued embeddings to better model asymmetric relationships. RotatE infers various relation patterns, including symmetry, antisymmetry, inversion, and composition, by defining each relation as a rotation from the source entity to the target entity. The study in [24] shows that KGE methods outperform traditional graph exploration methods in predicting drug-target interactions, polypharmacy side effects, and tissue-specific protein functions.

A recent model with higher expressiveness and the ability to encode indirect semantics using GNNs is the Relational Graph Convolutional Network (R-GCN) for multi-relational KGs [39]. It learns node embeddings by aggregating transformed feature vectors of neighboring nodes via a normalized sum and then uses its decoder, the DistMult factorization model, to perform link prediction. Unlike conventional GCNs, R-GCNs introduces relation-specific transformations based on edge type and direction, making them suitable for multi-relational data in KGs.

With the rapid accumulation of biomedical data, understanding disease biology and molecular factors' roles in phenotypic outcomes is crucial for personalized diagnostics and treatments. KGs have become the dominant knowledge representation in biomedicine, leveraging databases like UniProt [40], Gene Ontology [41, 42], and DrugBank [43]. LP tasks in biomedical KGs, such as Zhang et al.'s COVID-19 drug candidate exploration [44] with RotatE and DistMult, OntoProtein's Gene Ontology-based KG for protein language model pretraining [45], and Biswas et al.'s node embedding algorithms for multimodal biomedical KGs, enhance drug discovery and predict disease co-morbidities [46] via tensor factorization with complex-valued embeddings. Further, task-specific KGs and frameworks like BioCypher [47] further support KG construction, aiding predictive modeling for drug adverse reactions, repurposing, and biological concept associations.

Another prominent biomedical KG, PrimeKG, integrates 20 primary resources including DisGeNet, MONDO Disease Ontology, and DrugBank [48], encompassing 17,080 diseases and four million relationships. It includes diverse disease-associated data such as protein perturbations, biological processes, pathways, phenotypes, and drug therapeutic actions, including 'indications' and 'contradictions' drug-disease edges. TxGNN [49], a KGE-based method, leverages PrimeKG via the R-GCN framework, and maps therapeutic candidates and diseases into a latent space optimized to reflect graph semantics. TxGNN performs zero-shot inference on diseases with no known treatment and minimal molecular understanding, predicting drug-disease relationships with a focus on indication and contraindication tasks, efficiently handling therapeutic tasks through adaptive aggregation and iterative optimization of embeddings.

LP methods using node embeddings have the disadvantage of low interpretability and are not capable of LP in inductive settings, whereby a graph with new nodes are supplied for training. To overcome these challenges, researchers started developing general and flexible representation learning frameworks for LP based on the paths between two nodes. To overcome these challenges, researchers started developing general and flexible representation learning frameworks for LP based on the paths between two nodes. A first application of this concept is the study from [50], who introduce KG4SL, a graph neural network (GNN) model which integrates KG message-passing for synthetic lethality (SL) prediction, leveraging a KG with 11 entity types and 24 relevant relationships associated with SL. Further, the Neural Bellman-Ford Network (NBFNet) introduces a novel framework for LP, inspired by traditional path-based methods [51]. It represents node pairs as the sum of path representations, each derived from edge representations. NBFNet employs a graph neural network with learned operators for efficient path formulation solutions, scalable to large graphs with low time complexity. NBFNet works with both homogeneous and multi-relational graphs, supporting LP across different graph types. Combining traditional path-based methods with GNNs, NBFNet demonstrates superior performance compared to node embedding methods. Additionally, path embedding methods offer better interpretability by visualizing important paths used for prediction, facilitating verification of biological plausibility.

To overcome the challenges of predicting links in noisy Biological knowledge graphs (KGs), which often include irrelevant relations, we introduce BioKGC. This message-passing neural network framework for path representation learning, inspired by NBFNet, specializes in predicting specific node subset relationships like gene-pathway, drug-disease, gene-gene, and lncRNA-target interactions within biomedical KG. As opposed to node-embedding learning frameworks (Figure 1A) that optimize the embedding space based on one-hop relations, BioKGC utilizes path-based reasoning to learn representations between source and target nodes based on relations along the path (Figure 1B). Introducing a background regulatory graph (BRG) like a PPI network enhances BioKGC's effectiveness over prior path representation learning methods when it comes to biomedical KGs. By leveraging additional graph information from the BRG for message passing (Figure 1C-D), BioKGC enriches path representations between node heads and tails, resulting in more precise predictions. In addition, the node type aware negative sampling scheme introduced in BioKGC ensures more difficult and effective contrastive learning (Figure 1C). These two points are especially important to large biomedical KGs that potentially encode noise, derived from errors in experiments, and at the same time are highly structured in how and which biological entities can interact.

We highlight BioKGC's effectiveness across four diverse LP tasks: *gene function prediction task*, *drug repurposing task*, i.e. disease-drug target interaction prediction in a zero-shot scenario, *synthetic lethality prediction task*, i.e. prediction of synthetic lethality gene pairs, *lncRNA-gene target prediction task*, i.e. inference of lncRNA-mRNA regulatory relationships. Despite varying KG requirements for each task, BioKGC always surpasses KGE-based methods, including GNNs in most the tasks. For the *drug repurposing task* and *synthetic lethality prediction task*, it matches or outperforms task-specific models like TxGNN and KR4SL. BioKGC discovers new drug-disease associations, including insights into Alzheimer's disease, and scores potential lncRNA-mRNA interactions, validated against orthogonal datasets. Through examples, we demonstrate how BioKGC enables natural interpretation of predicted links, enhancing understanding of molecular disease mechanisms and regulatory processes.

2 Results

A knowledge graph (KG) is a heterogeneous directed graph comprising various types of entities (nodes) connected by relationships (edges). For instance, a KG might include nodes representing diseases, genes, and potential drug targets, with relationships such as 'indication for' of 'involved in' and model facts such as 'drug A is an indication for disease B' or 'gene C is involved in disease D.' KGs are typically represented as triples consisting of a head node, a tail node, and a relationship. The task of knowledge graph completion involves estimating the missing components of these triples. For example, one might predict the tail entities corresponding to a given head entity linked by a specific relationship, such as predicting diseases for which a particular drug is an indication, based on existing triples (i.e., existing knowledge). Knowledge graph completion methods can be broadly categorized into embedding-based and path-based

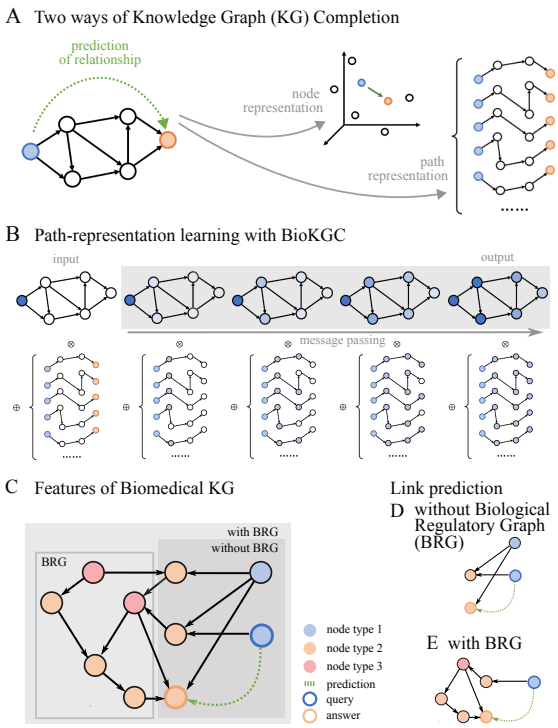


Fig. 1: **Link prediction (LP) in biological knowledge graphs:** A) Inference of links using node-representation (node embedding) vs. path-representation learning. B) Illustration of the idea behind the NBFNet framework, which utilizes the generalized Bellman-Ford algorithm to solve the shortest path problem between a head entity and all tail entities connected by specific relationships. Additionally, it employs neural formulations, specifically message passing GNNs, to learn path representations. Downstream, a Multi-Layer-Perceptron uses the path representations to distinguish between positive and negative relationships in a supervised fashion (not shown). C) Tailored to the requirements of biomedical knowledge graphs (KGs), BioKGC incorporates a background regulatory graph (BRG) that adds additional connections between genes. This supplementary knowledge is leveraged for message passing, enhancing the information flow beyond the edges used for supervised training. Additionally, BioKGC employs an improved negative sampling scheme that considers specific node types to further refine the learning process (not shown). D-E) Examples of paths used for prediction are illustrated between head nodes (blue) and tail nodes (orange) in two scenarios: D) a bipartite sub-graph without BRG, and E) a sub-graph that includes BRG connections leveraged for model prediction. These examples also serve as model explanations, highlighting the paths that lead to the model's predictions.

approaches (Figure 1A). Embedding-based approaches use encoding models, ranging from simple linear models to complex neural networks, to learn feature representations of entities in a knowledge graph. These methods aim to preserve the structure of

the original graph in a lower-dimensional space by minimizing the distance between the head and tail entity embeddings and the relationship embeddings, or by maximizing the similarity between the embeddings of head entities, relations, and tail entities. While embedding-based approaches have shown significant performance in several benchmark tests, they are often limited to one-hop relations. In large biomedical Knowledge Graphs relationships between entities are intricate and may involve multi-hop paths. Encoding a head entity without considering its specific tail entities requires embedding a vast amount of information (considering all possible tail entities). For large graphs, embedding all this information into a lower-dimensional vector is challenging and can lead to imprecise link predictions. Methods such as SEAL and Grail address the problem of predicting links between head and tail entities by embedding the subgraph structure around the link, encoding the two entities as a whole. However, these methods face scalability issues because they generate or materialize a subgraph for every link they try to predict. This process becomes a bottleneck when attempting to perform link prediction for all pairs. To address this challenge, path-based approaches like NBFNet can be leveraged to capture the structural information of knowledge graphs by learning representations for pairs of nodes (instead of single nodes) through paths. BioKGC introduces a novel method for link prediction in biomedical knowledge graphs based on the NBFNet framework. It learns representations for each path from the query entity to potential tail entities, considering the relations along the path. The representation of a pair of nodes is parameterized as the generalized sum of all path representations between the nodes, with each path representation being the generalized product of the edge representations within the path (Figure 1B). This path formulation can be efficiently solved using the generalized Bellman-Ford algorithm based on dynamic programming. Moreover, the efficiency is further enhanced by learning the operators of the generalized Bellman-Ford algorithm with a message passing graph neural network (see Methods).

BioKGC incorporates significant modifications to the NBFNet framework to address the unique characteristics of biomedical knowledge graphs (KGs) and prediction tasks in biomedicine. One key enhancement is the adoption of a more stringent negative sampling strategy, where negative samples are exclusively drawn from the same node type as the positive samples. This approach ensures that negative samples are sufficiently challenging, enabling the model to learn more precise decision boundaries. Additionally, BioKGC leverages an external biological regulatory graph (BRG) to augment the connectivity of entities within the knowledge graph. This enhances prediction accuracy by incorporating supplementary information from external knowledge bases during the message passing step of model training, while ensuring that negative sampling and the associated loss computation remain unaffected. Essentially, predictions can be made using a bipartite graph alone, or additional knowledge can be supplied via a BRG used solely for message passing (Figure 1C). For instance, when predicting the missing link between a head node and a tail node, messages can be passed within the bipartite graph between type 1 and type 2 nodes, resulting in a prediction path as shown in Figure 1D. Alternatively, a BRG can be integrated to further inform the predictions by leveraging additional knowledge bases, such as relations between type 2 and type 3 nodes, as illustrated in Figure 1E. The addition of

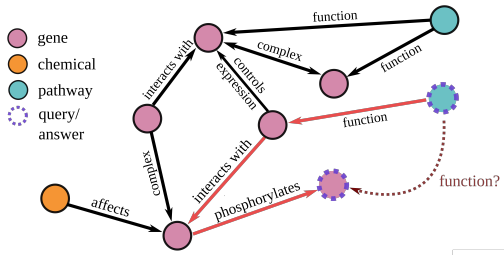
a BRG to the KG not only allows for more expressive path representations between entities, resulting in more effective learning, but also enhances the method's capabilities in interpreting the most important paths leading to a particular prediction. One of the key advantages of NBFNet over node embedding methods is its straightforward interpretability. By incorporating a BRG in BioKGC, it becomes possible to understand the molecular mechanisms behind a prediction by examining the cellular network (interaction partners, regulators) surrounding a specific entity pair. This additional layer of insight leverages the broader biological context provided by the BRG.

2.1 Gene Function Prediction Task

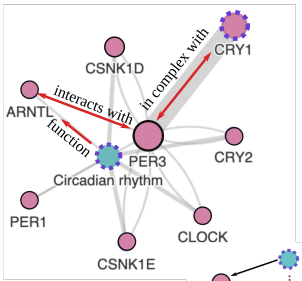
Our first goal was to evaluate the capacity, performance, and robustness of path embedding, and particularly of BioKGC, in biomedical knowledge graph (KG) link prediction, as compared to node embedding techniques, with special emphasis on assessing the use of a BRG for message passing within the BioKGC framework. Therefore, we conducted a proof-of-concept study focusing on gene function prediction. This involves assigning biological information, like cellular pathways, to genes. Our approach involved applying BioKGC to two scenarios: Firstly, utilizing a KG structured as a bipartite graph connecting genes and KEGG pathways through the relation 'function of' sourced from ConsensusPathDB [52]. Secondly, we extended this KG by incorporating a Biomedical Relation Graph (BRG) extracted from Pathway Commons [53–55] interactions encompassing gene-gene, chemical-gene, and chemical-chemical relationships. The objectives of this study were two-fold: firstly, to evaluate BioKGC's performance in link prediction compared to traditional node embedding methods. Secondly, to assess the impact of augmenting the bipartite KG with a BRG on enhancing the accuracy of gene function annotation tasks. Through this investigation, we aimed to validate BioKGC's utility in leveraging complex biomedical data structures for improving predictive modeling in gene function annotation within knowledge graphs.

In direct comparison with Knowledge Graph Embedding (KGE)-based methods such as TransE, DistMult, and RotatE, as well as Graph Convolutional Networks (R-GCN), BioKGC demonstrated consistently superior performance across different metrics (Figure 2B-C). In the setting without utilizing the Biomedical Relation Graph (BRG), BioKGC achieved a Mean Reciprocal Rank (MRR) of 0.464, outperforming the KGE methods which averaged 0.371, and R-GCN which achieved 0.348. For the $Hits@k$ metric, which indicates the percentage of ground truth items captured within the top k predictions, BioKGC obtained 63.5% in the top 10 predictions, compared to RotatE's 56.5% (Figure 2B). Upon leveraging the BRG for biological regulation-enhanced message passing, performance improvements were observed primarily for R-GCN and BioKGC. R-GCN's MRR increased marginally from 0.348 to 0.355, whereas BioKGC's performance rose from 0.464 to 0.549, corresponding to 8.5% gain. In terms of capturing ground truth positives within the top 10 predictions, BioKGC excelled with 72.6%, whereas R-GCN achieved 53.1% (Figure 2C). Interestingly, KGE methods did not leverage the additional BRG information effectively; in fact, the TransE model's performance dropped significantly from 0.376 to 0.272, indicating a disadvantage rather than an enhancement in predictive capability.

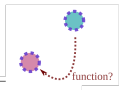
A Schema of Functional Annotation KG



D Visualization of 10 most influential paths for annotation of Circadian rhythm for CRY1

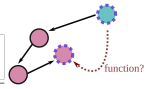


B Functional annotation without BRG



Class	Method	w/o BRG				
		MR	MRR	H@1	H@3	H@10
Embedding	TransE	144.80 ± 0.46	0.373 ± 0.002	0.277 ± 0.002	0.407 ± 0.003	0.565 ± 0.001
	DistMult	147.40 ± 1.58	0.363 ± 0.002	0.269 ± 0.002	0.395 ± 0.003	0.551 ± 0.002
	RotatE	149.30 ± 0.65	0.377 ± 0.001	0.281 ± 0.002	0.410 ± 0.002	0.569 ± 0.001
Graph Convolutions	R-GCN	161.53 ± 3.57	0.338 ± 0.002	0.242 ± 0.004	0.370 ± 0.001	0.533 ± 0.004
Path-based	NBFNet	**112.30** ± 1.29	**0.457** ± 0.008	**0.370** ± 0.011	**0.491** ± 0.007	**0.627** ± 0.006

C with BRG



w/ BRG				
MR	MRR	H@1	H@3	H@10
267.31 ± 2.16	0.309 ± 0.001	0.226 ± 0.002	0.332 ± 0.002	0.479 ± 0.001
336.29 ± 16.53	0.322 ± 0.002	0.238 ± 0.001	0.347 ± 0.003	0.488 ± 0.004
264.09 ± 2.44	0.333 ± 0.001	0.249 ± 0.002	0.358 ± 0.002	0.502 ± 0.002
274.73 ± 4.79	0.344 ± 0.003	0.254 ± 0.003	0.373 ± 0.003	0.526 ± 0.005
123.33 ± 4.06	**0.546** ± 0.003	**0.454** ± 0.002	**0.592** ± 0.005	**0.725** ± 0.005

Fig. 2: **Benchmark of knowledge graph completion algorithms on the gene function annotation task:** A) Illustration of BioKGC leveraging a biological regulatory graph (BRG) encompassing genes, chemicals, and cellular pathways, to predict gene functions, i.e. associations of gene with specific cellular pathways. B-C) Performance comparison of BioKGC on the gene function prediction task against classical KGE-based methods, namely TransE, DistMult and RotatE, as well as convolutional neural network-based methods such as R-GCN for link prediction without the underlying BRG and C) with the BRG. Metrics reported for comparison are mean rank (MR), mean reciprocal rank (MRR) and Hits at 1, 3 and 10. D) The visualization highlights the significant paths employed by the BioKGC model to predict a link between *CRY1* and the *Circadian rhythm*. The top 10 paths are depicted, where the width of each edge corresponds to its weight, and the path with the highest weight is highlighted in red.

By conducting experiments for each method over 5 different model seeds, we observed standard deviation for each method ranging between 0.01 and 0.03, yielding in robust predictions for all methods.

One key advantage of NBFNet is its ability to provide interpretable predictions through paths, which are crucial for understanding the rationale behind specific predictions. Intuitively, these interpretations should highlight paths that significantly influence the prediction. In BioKGC, this follows the NBFNet framework, where the top-k path interpretations for a prediction are formally defined as the first derivative (gradient of the prediction) with respect to each path between a head and a tail node. In this task, we exemplify how BioKGC interprets its predictions and visually presents the most critical paths for predicting the function of the *CRY1* gene, specifically its

association with the *Circadian rhythm* pathway (Figure 2D). The figure illustrates the top 10 most significant paths ranked by gradient, where the width of each edge reflects how frequently it appears among these top paths. Additionally, the most crucial path, ranked highest by gradient, is highlighted in red, encompassing: *CRY1* `in complex with` → *PER3* `interacts with` → *ARNTL* before feeding into the pathway *Circadian rhythm* over the relation `function of` (Figure 2D). The retrieved path makes sense as it recovers the well known mechanisms by which the essential transcription factors controlling the cellular circadian rhythm, ARNTL and CLOCK, upregulate the expression of PER3 and CRY2 [56, 57]. They in turn form heterodimers to repress their own expression, creating a negative feedback loop of regulation [58, 59].

2.2 Drug repurposing task

In the second part of our study, we evaluated BioKGC in a more challenging scenario: predicting new drug candidates for diseases by repurposing existing drugs indicated for other conditions. This drug repurposing task was conducted in a zero-shot scenario, where the target disease has minimal molecular characterization and no available treatments. For this experiment, we followed the data split procedure implemented in TxGNN [49], a state-of-the-art a graph neural network model designed to predict drug-disease relationships in zero-shot scenarios, which build embeddings of nodes and relations from a comprehensive biomedical knowledge graph, the PrimeKG knowledge graph (Figure 3A) [48] (see Methods for more details).

More in detail, TxGNN creates 5 'disease area' splits to simulate zero-shot conditions, ensuring that diseases in the hold-out set used for inference (1) have no approved drugs in the training data, (2) have limited overlap with the training diseases by excluding similar ones, and (3) lack molecular data by removing their biological neighbors from the training set. These splits provide challenging yet realistic evaluation scenarios, mimicking zero-shot drug repurposing (see Methods). Essentially, this setup simulates a new disease with little knowledge, no similar diseases, and no known treatments, removing connections to treatments and most biological neighbors from the training set to prevent their use in message passing. These splits create challenging yet realistic evaluation scenarios for zero-shot drug repurposing by simulating a new disease with minimal knowledge, no similar diseases, and no known treatments. Connections to treatments and most biological neighbors are removed from the training set to prevent their use in message passing. Five distinct zero-shot disease areas were used: adrenal gland, anemia, cardiovascular, cell proliferation, and mental health.

The BioKGC model used approximately 5.7 million directed edges solely for message passing in each prediction setting (Table B3), including non-drug-disease edges like protein-protein and disease-disease relations. In contrast, edges used for both message passing and supervision were limited to drug-disease interactions, such as 'contraindication' and 'indication'. On average, the training set contained around 33,000 edges, and the validation set around 4,000. The number of testing edges varied significantly between disease areas, with 1,047 contraindications and 999 indications in the cell proliferation split, compared to 303 contraindications and 33 indications in the adrenal gland disease area. These differences are due to the number of diseases within each area (Table B3) and likely reflect research biases.

For each disease area split, we evaluated BioKGC against TxGNN by assessing their performance in predicting ground truth drugs for the relations 'contraindication' and 'indication'. Specifically, we ranked all drugs (tail node) based on their likelihood of being an indication or contraindication for a specific disease (head node). This involves computing the probability for each drug to be an indication or contraindication for a disease from both BioKGC and TxGNN, $p(\text{drug} \mid \text{disease}, \text{relation})$ (Figure 3B). In the comparison, BioKGC achieved higher AUPRC than TxGNN in three out of five disease areas for contraindication prediction and in all disease areas for indication prediction (Figure 3C). The difference in performance, Δ, is calculated by subtracting TxGNN's AUPRC from BioKGC's AUPRC; thus, a positive Δ indicates better performance by BioKGC. For contraindications, TxGNN outperformed BioKGC in adrenal gland, cardiovascular, and mental health areas with Δ values of -4.6, -0.2, and -2.0 percentage points, respectively. Conversely, BioKGC outperformed TxGNN in anemia and cell proliferation with Δ values of 4.8 and 9.0. In the indication prediction task, BioKGC consistently had positive Δ values, ranging from 5.9 to 22.6 percentage points (Figure 3D). To summarize the performance in a single metric, the AUPRC is averaged across contraindication and indication for each disease area split (Table B4). For cell proliferation, the difference in performance Δ is $0.119(0.556 - 0.437)$, representing a performance increase for BioKGC over TxGNN of 27.3%. Similarly, the increases are 17.1%, 14.1%, 25.9%, and 16.9% for adrenal gland, anemia, cardiovascular, and mental health, respectively. On average, BioKGC outperforms TxGNN by 20.2% across all disease area splits.

Cell Proliferation Split

A detailed breakdown of BioKGC and TxGNN in terms of Specificity, F1, and Recall@k for the *cell proliferation* disease area split is illustrated in Figure 3E. Both models perform well in identifying true negatives, with BioKGC showing only slightly higher specificity (0.996 vs. 0.981) in the indication setting. The F1 score, representing the balance between precision and sensitivity, was 0.415 for BioKGC vs. 0.330 for TxGNN in the contraindication setting, and 0.393 vs. 0.183 for indication. This trend, observed in the cell proliferation split, holds across all other disease splits. Notably, BioKGC shows a greater improvement in performance for indication than contraindication, though with slightly higher variance compared to TxGNN. In the cell proliferation split, 178 diseases had known indications, with an average of 5.58 indications per disease. For 60 out of 178 diseases, all known treatments are prioritized within the top 10 predictions sampled from a list of 7,000 drug candidates. Recall@k quantifies the proportion of ground truth items found within the top k predictions. For instance, at $k = 20$, the recall for indication across all diseases is 0.619 for BioKGC, meaning 61.9% of ground truth drugs are found within the top 20 predictions. In contrast, Recall@20 for TxGNN is 0.539. BioKGC consistently outperforms at lower k values (Figure 3E) and matches TxGNN at $k = 100$ (not shown).

Case study from Cell Proliferation: Acute Lymphoblastic Leukemia (ALL) After quantitatively evaluating the performances, we further examined individual disease predictions within the Cell Proliferation split. Acute Lymphoblastic Leukemia (ALL) is a complex cancer involving abnormal proliferation of lymphoid

cells in blood and bone marrow, impairing immune function [60]. Contributing factors include genetic syndromes like Down syndrome and viral infections [61–65]. Commonly observed chromosomal aberrations include the t(9;22) translocation, which produces the constitutively active tyrosine kinase BCR-ABL1, associated with Philadelphia chromosome-positive ALL [66].

We used BioKGC for the prediction of the drugs associated with ALL. We were able to retrieve the only known contraindication - drug Aprostadil on rank 1 with a probability score of 0.727, as well as all 21 known indication within the top 34 predictions (Figure 3F). Upon investigating the top indication predictions, we identified the highest-ranked known treatments (in orange): Clofarabine (0.723), Teniposide (0.723), and Methotrexate (0.720). Additionally, the top-ranked unknown treatments (in blue) were Dasatinib (0.724) and Bosutinib (0.721) (Figure 3F). We further set out to interpret our predictions by visualizing the most important paths for the predictions. The visualization plot summarizes the top ten most important paths as ranked by gradient (see Methods), with the edge width reflecting the number of times the edge appear among the top ten paths. The first novel indication prediction with high probability score is Dasatinib 3F), which is not present in the ground truth database PrimeKG. However, Dasatinib, an inhibitor of the constitutively active tyrosine kinase BCR-ABL, is already used for treating Philadelphia chromosome-positive acute lymphoblastic leukemia (Ph+ ALL) in cases of resistance or intolerance to prior therapies. The next novel prediction, Bosutinib, is an unknown drug predicted to treat ALL with a probability score of 0.721 (Figure 3F, left). To gain confidence in this prediction, we visualized the most important paths leading to it, focusing on the local subgraph to explain our results. For Bosutinib as an indication for ALL, the similarity to other (lymphoblastic) leukemia types was revealed, along with significant disease genes AICDA and DUX4 [67–69]. The most crucial path passes through the phenotype *Ph+ ALL*, the disease *chronic myelogenous leukemia, BCR-ABL1 positive*, and the gene *BCR*, before connecting to *Bosutinib* via the *drug protein* relation (Figure 3F, right). Indeed, Bosutinib was originally indicated for chronic myeloid leukemia in 2012 [70, 71] and is currently being investigated for the treatment of ALL [72].

Clofarabine is a known treatment that was included in the ground truth PrimeKG database. For this prediction, the model was able to identify the disease similarity between ALL and the susceptibility to the lymphocytic leukemia and made use of its link to the drug. Curious about which path the model would choose, we excluded links between similar diseases and drugs in the KG, so that the disease similarity cannot be used for the prediction. This time, important nodes that were retrieved were the genes SMC1A, involved in chromosome cohesion during cell division and DNA repair, and POLA1, forming part of the DNA polymerase alpha subunit. Given that Clofarabine is a purine nucleoside that is metabolised intracellularly to inhibit DNA synthesis [71, 73], through alternative paths the model was able to pick up important components on the mode of action of the drug.

Hypothesis generation for treatment of Gastric Cancer

To demonstrate BioKGC's ability to generate hypotheses for lab testing and evaluation, we focus on gastric cancer. Similar to ALL, both contraindications and indications were ranked highly (all 5 contraindications in the top 6, and 5 out of 6 indications in

the top 5) (Figure 3G left). One novel drug predicted for gastric cancer treatment is Acitretin, an oral retinoid similar to Vitamin A, indicated for skin diseases like psoriasis by inhibiting excessive cell growth and keratinization [71]. Although untested for gastric cancer, Acitretin has been considered in combination with Clarithromycin for cutaneous squamous cell carcinoma due to its apoptosis-inducing properties [74, 75]. Interestingly, all paths from gastric cancer to Acitretin in the interpretability plot pass through RBP1 (retinol-binding protein 1), annotated with the *retinoic acid biosynthesis process* (Figure 3G right). Recent studies suggest this pathway's involvement in gastric cancer treatment and provide pre-clinical evidence supporting the use of All-Trans Retinoic-acid (ATRA) [76, 77]. By visualizing important paths between drugs and diseases, researchers can verify predictions' plausibility and generate hypotheses for further laboratory validation.

Predicting drug indications for Alzheimer

For the final experiment in the drug repurposing task, we aimed to investigate a disease not analyzed by TxGNN to evaluate how well BioKGC generalizes to a novel case study. Alzheimer's disease (AD) is a neurodegenerative disorder characterized by extracellular amyloid beta and intracellular tau protein accumulation in the brain. These neuropathological changes occur decades before clinical symptoms, ultimately leading to synapse loss, brain atrophy, and dementia symptoms like memory loss and behavioral changes. While amyloid and tau are central to AD, the exact mechanisms remain unclear. Emerging evidence suggests additional pathways, such as immunoinflammation and bioenergetic dysregulation, may offer promising therapeutic targets [78–80]. Presently, FDA-approved treatments include only two disease-modifying and five symptomatic treatments, none of which provide a cure for AD. To explore the potential of BioKGC for such complex and heterogeneous diseases, we trained BioKGC on a data split tailored for zero-shot prediction on a custom defined Alzheimer's disease area split. Here, we followed the disease evaluation code as provided by TxGNN to exclude all treatments for Alzheimer's, as well as closely related diseases (dementia, Alzheimer disease (without neurofibrillary tangles, Pick disease and Lewy body dementia) (Table B6). We then evaluated the top 20 predictions for indications and contraindications for Alzheimer's disease.

Among the top 14 predictions, seven out of eight drugs classified as known treatments according to PrimeKG, and four out of seven FDA-approved treatments for Alzheimer's disease (AD), were successfully retrieved (see Figure 4A). Additionally, the model identified Epicriptine, a nootropic drug with an unknown mode of action, and Acetylcarnitine, which is functionally involved in β-oxidation of fatty acids [81]. Known drugs with low probability scores included Pramiracetam (ranked 344), used for cognitive impairment in aging and dementia [82], and FDA-approved treatments such as Memantine (ranked 412), an N-methyl-D-aspartate receptor antagonist [83], and the recently approved monoclonal antibodies Lecanemab (ranked 2250) and Aducanumab (ranked 2216) [84, 85]. Interestingly, two drugs currently undergoing clinical trials were among the top 20 predicted novel indications: Nicotine, a nicotinicacetyl choline receptor agonist which is being tested in a Phase II clinical trial (NCT02720445) to improve cognition, and Bupropion, an N-methyl-D-aspartate receptor antagonist that

is being tested as a component of the drug AXS-05 in two Phase III (NCT05557409, NCT04947553) clinical trails [86] to help with agitation associated with AD. Examination of the interpretability graphs shows that both predictions are associated with the brain-derived neurotrophic factor (BDNF), a gene crucial for synaptic maintenance and plasticity in the brain [87] (Figure 4C). Synaptic plasticity plays a pivotal role in AD [88], with research indicating lower levels of BDNF in both blood [89] and brain [90] in AD patients and linking higher levels of brain BDNF with slower cognitive decline [91] in elderly individuals. In addition, analyses of brain gene co-expression networks underscore BDNF's involvement in downregulated pathways ($P_{adj} = 6.93e-17$, adatlas.org [92]). Both predicted drugs, Bupropion and Nicotine, have demonstrated an ability to elevate BDNF levels in serum [93, 94],providing a functional hypothesis for the mechanism of these drugs in the context of AD. Another promising candidate predicted with high probability is Everolismus, an analog of Rapamycin and a selective inhibitor of the mammalian target of Rapamycin (mTOR) kinase signaling pathway. This pathway has been implicated in both normal aging and pathological aging processes, making it a promising target for intervention, particularly in the early stages of disease onset [95]. Currently, Rapamycin is being evaluated as a potential disease modifying therapy in Phase II (NCT04629495) and Phase I (NCT04200911) clinical trials involving older adults with mild cognitive impairment or early AD [86]. Furthermore, Rapamycin has shown beneficial effects on amyloid and tau burden in mouse models of AD [96]. Everolismus, although structurally similar to Rapamycin, has favourable clinical pharmacokinetics that influence, for example, bio-availability and tissue distribution [97]. Therefore, Everolismus may present an additional candidate for targeting the hyperactivated mTOR pathway in AD.

2.3 Synthetic lethality prediction task

Synthetic Lethality (SL) occurs when the simultaneous mutation of two genes leads to cell death, while the mutation of either gene alone is non-lethal [98]. We next examined the prediction of missing SL gene pairs with BioKGC, which is of high interest in anti-cancer drug treatment. In fact, when the first partner of a gene pair is inhibited by mutations in cancer cells, targeting the second partner can induce selective cell death in cancer cells without harming normal cells. This approach is crucial when direct targeting of cancer driver genes is impractical, but their synthetic lethality (SL) partners offer viable treatment alternatives. Given the great potential to design personalized treatments through SL-based therapy, computational methods to predict novel gene interaction partners are of great importance. In our study, we leverage BioKGC for this task and compare it against the state-of-the-art method, named KR4SL [50] which is a path-representation learning GNN-based method to predict and explain synthetic lethality gene pairs (see Methods).

For training and inference of BioKGC, we used the SynLethDB-v2.0 [99] data which was pre-processed by the authors of KR4SL framework and uploaded to their GitHub repository. The resulting training set consists of 9,878 SL gene pairs and 5,444 source genes, validation set - 3,533 SL gene pairs and 2,918 source genes, test set - 7,049 SL gene pairs and 4,474 source genes, and SL part of the heterogeneous graph contains 14,858 known SL gene pairs (Table C8)

Another experiment we conducted was setting a threshold on the SL gene pair confidence provided by SynLethDB. As this database contains genetic interactions from various sources (experiments and predictions), we investigated the effect on performances along thresholds from 0 to 0.8. As an example, removing all SL gene pairs with a confidence score less than 0.3, more than 75 % pairs derived from computational prediction methods were removed. The resulting training set then consisted of 8,770 SL gene pairs, validation set - 3,172 SL gene pairs, test set - 6,254 SL gene pairs, and SL part of the heterogeneous graph contained 13,161 known SL gene pairs (Table C8). We refer to this collection of filtered data sets as 'thresholded data', when the original data will be referred as 'unthresholded data'.

For each seed and each threshold, we evaluated both KR4SL and BioKGC in their ability to predict SL partners for a gene in terms of NDCG@k, Precision@k, and Recall@k, $k \in \{10, 20, 50\}$. Averaged across seeds, we were able to achieve significantly higher performance for all these metrics for unthresholded data at significance level $\alpha = 0.01$. For thresholded data, we significantly outperform KR4SL at level $\alpha = 0.1$ for Recall@10 and Precision@10 and at level $\alpha = 0.01$ for other metrics (Figure 5B). In case of unthresholded data, BioKGC outperformed KR4SL for all metrics with higher NDCG@10 by 2.3%, NDCG@20 by 2.7%, NDCG@50 by 3.1%, Recall@10 by 1.6%, Recall@20 by 3.0%, Recall@50 by 4.6%, Precision@10 by 1.6%, Precision@20 by 3.0%, Precision@50 by 4.6%. Analogously, with thresholded data as input, BioKGC outperformed KR4SL for all metrics with higher NDCG@10 by 1.5%, NDCG@20 by 1.8%, NDCG@50 by 2.1%, Recall@10 by 0.7%, Recall@20 by 1.6%, Recall@50 by 2.9%, Precision@10 by 0.6%, Precision@20 by 1.6%, Precision@50 by 2.9%.

A detailed breakdown for the performance of both methods in terms of MRR, NDCG@k, Precision@k, and Recall@k for $k \in \{10, 20, 50\}$ is reported for each seed in Table C9 for unthresholded data and in Table C10 for thresholded data.

After quantitatively assessing and comparing model performances, we examined BioKGC's predictions for SL gene pairs on thresholded data, focusing on consistently predicted new SL partners across different seeds. We began our analysis with the gene pair EYA4 and MUS81. EYA4 is a protein-coding gene involved in transcriptional activation, eye development, and functioning of the organ of Corti. Mutations in EYA4 are linked to progressive hearing loss and dilated cardiomyopathy, and it acts as an oncogene involved in DNA repair, apoptosis, and innate immunity following DNA damage, cellular damage. MUS81 encodes an endonuclease crucial for resolving recombination intermediates during DNA repair. BioKGC predicts MUS81 as the 7th synthetic lethality partner among the top 20 candidates for EYA4, as shown in Figure 5E. The explanation subgraph in Figure 5C illustrates multiple paths from EYA4 to MUS81, particularly through shared biological processes like DNA repair and cellular response to DNA damage, providing evidence for their SL relationship.

Another example involves POLB and BRCA1. POLB is a repair polymerase that is essential for base-excision repair and has been found to be linked to Werner syndrome and esophageal cancer. BRCA1, a tumor suppressor gene, is crucial for DNA repair and in our model it is predicted as the 9th SL partner among the top 20 candidates for POLB, as shown in Figure 5F. The explanation subgraph in Figure 5D shows 10 most important paths from POLB to BRCA1 involving biological processes like DNA

repair, DNA replication, embryonic development, cellular homeostasis, and intrinsic apoptotic signaling in response to DNA damage. Here, three nodes - DNA repair, in utero embryonic development, and the apoptotic signaling pathway in response to DNA damage - are shared by POLB and MSH2, a known SL partner of TAB2, which is an SL partner of CUL1, which in turn is an SL partner of BRCA1. Additionally, POLB and SUPT16H, a known SL partner of BRCA1, are involved in DNA replication. Finally, the biological process node for cellular homeostasis is shared by POLB and ASXL1, another known SL partner of BRCA1. These shared processes, particularly DNA repair and DNA replication, support the evidence of SL relationship between POLB and BRCA1.

2.4 LncRNA-target prediction task

Long non-coding RNAs (lncRNAs) are a heterogeneous group of transcripts that lack protein-coding potential, usually longer than 200 nt. They encompass a substantial portion of the genomes of complex organisms. The extensive transcription of these non-coding transcripts unveils a significant shift in our understanding of the pivotal role of RNAs in gene regulation. They are implicated in imprinting control, immune response, epigenetic regulation, and gene regulatory networks. Mutations and dysregulation of lncRNAs are linked to various diseases, and pinpointing disease-associated lncRNAs can offer new biomarkers for diagnosis, treatment, and prognosis. Next-generation sequencing data generated by several consortia, such as the ENCODE [100] and FANTOM5 [101], and comprehensive resources of noncoding RNA sequences, such as RNAcentral [102] and the NONCODE [103] lead to an estimate of over 200,000 potential lncRNA transcripts. Research on lncRNAs reveals various functional roles and mechanisms. They may regulate gene expression locally (cis) or distantly (trans) and interact with RNA Binding Proteins (RBPs) and other nucleic acids. lncRNAs can act as signals, scaffolds, guides, and enhancer-like RNA that modulate gene expression via chromatin looping and noncoding competing endogenous RNAs (ceRNAs), which act as microRNA sponges or decoys and de-repress other RNA transcripts by sequestration of shared microRNAs. They can also interact with DNA or chromatin remodelers to regulate transcription and chromatin architecture. Some lncRNAs, like XIST and HOTAIR, coordinate gene silencing by recruiting repressive complexes. In the cytoplasm, lncRNAs regulate translation, metabolism, and signaling. Despite great progress in the last years, the majority of lncRNAs are functionally uncharacterized, and their implication for disease biogenesis and progression is unknown. The imperative task of elucidating the functions and mechanisms of numerous lncRNAs underscores the urgency of identifying their targets using both experimental and computational approaches, which is a crucial initial step in functional analysis. Identifying the targets of lncRNAs, whether proteins, RNA sequences, or chromatin, is crucial in lncRNA research. The identification of such targets usually relies on the genetic context and sequences of lncRNAs, enabling the construction of comprehensive interaction networks. Experimental methods like RNA pull-down, ChIRP, RIP, and CLIP systematically screen and identify lncRNA targets.

On the KG derived from the lncRNA regulatory graph in LncTarD 2.0, BioKGC significantly outperformed both node embedding-based methods and the basic

NBFNet across all metrics (Table D12). The lower performance of embedding-based methods highlights the importance of considering node and gene types in gene regulatory knowledge graph completion, an aspect neglected by the basic versions of TransE and DistMult used in this study. Additionally, our experiments demonstrate the effectiveness of BioKGC's negative sampling strategy and the successful integration of the external Biological Regulatory Graph (BRG) with the NBFNet framework. To demonstrate BioKGC's ability to uncover novel lncRNA regulations, we focused on PVT1, a prominent oncogenic lncRNA. Using a trained model, we performed KGC with PVT1 as the viewpoint. PVT1, a Myc regulator often overexpressed in cancers, plays a crucial role in tumor initiation, proliferation, invasion, epithelial-mesenchymal transition, and apoptosis and is associated with poor prognosis and resistance to therapies. We set PVT1 as the head node and computed all conditional probabilities $p(t|\text{PVT1}, r)$ for all nodes across all relationship types. The top 5 novel predictions with the highest probabilities are shown in Table D13, where "novel" indicates the absence of a direct connection between these genes in the knowledge graph. Additionally, the top 10 most crucial paths for these predictions, ranked by gradient, are illustrated in Figure 6. For all five predictions involving PVT1, the model identified significant edges that do not form a direct path from PVT1 to the target gene. Instead, the predictions are inferred through bipartite graphs, where PVT1 and other lncRNAs are on one side, and the PVT1 target gene, along with other co-regulated genes, are on the other. This aligns with the biological understanding that genes in the same cluster are often regulated by the same factors. For example, Figures 6C and 6D show that PVT1 and GIHCG interact with the EZH2 protein, while GIHCG epigenetically regulates MIR200A, MIR200B, and MIR429. Thus, the model infers that MIR200A and MIR429 are also regulated by PVT1. This prediction is meaningful because it is known that the MIR200 family, which includes MIR200A, MIR200B, MIR429, MIR141, and MIR200C, is crucial in cancer initiation and metastasis. Evidence in the literature also indicates that PVT1 promotes cervical cancer progression by silencing MIR200B through EZH2 interaction, leading to histone H3K27 trimethylation and MIR200B inhibition [104]. PVT1 may also influence melanoma by regulating MIR200C via EZH2 [105]. Notably, PVT1-EZH2 regulation appears in all novel predictions, with four highlighted in the plots, underscoring EZH2's role in PVT1 regulation. This aligns with experimental evidence of PVT1-EZH2 interactions in various cancers, including gastric, thyroid, glioma, and hepatocellular carcinoma. Furthermore, BioKGC predicts an interaction between PVT1 and SUZ12, a member of the Polycomb Repressive Complex 2 (PRC2), along with EZH2. The model identifies a path through the physical interaction of the oncogenic lncRNA APTR, which represses the CDKN1A/p21 gene promoter via PRC2, involving both EZH2 and SUZ12.

Finally, in the past year, novel datasets of potential lncRNA-target interactions have been generated, shedding light on the regulatory mechanisms of lncRNAs. To evaluate the generalization capacity of BioKGC and its underlying KG on the lncRNA-target prediction task, we assessed the method's recall on two datasets, treating them as independent test sets. The first dataset comprises lncRNA target genes showing significant perturbation in a study by Liu et al. [106], which involved the interference of thousands of lncRNA loci using CRISPRi. In this study, the authors reported

target gene-lncRNA regulation pairs. The second smaller dataset encompasses a set of enhancer-like lncRNAs and their potential cis targets determined via chromatin interactions, as defined in Ntini et al. [107]. In a transductive setting, we evaluated the method by determining how many novel interactions from these datasets (i.e., those not included in the KG) could be identified at different probability thresholds, thereby constructing a recall curve for each dataset. By comparing these results to a random scenario, where the conditional probabilities of the potential novel pairs were randomly sampled from a background distribution (see Methods), we observed that the recall of true lncRNA-target gene pairs in both datasets exceeded that of the recall curve derived from random pairs. This indicates that BioKGC, trained on the lncRNA-target gene prediction task, can score new datasets containing potential new regulatory interactions significantly better than random.

3 Discussion

With our novel method, BioKGC, we harness the power of path representation learning for link prediction on biomedical knowledge graphs (KGs) for the first time. In the functional annotation task of reconstructing the KEGG pathway database, BioKGC significantly outperformed classical KGE methods (TransE, DistMult, RotatE) and the GCN-based R-GCN. Without BRG, BioKGC achieved over 20% improvement compared to KGE models and over 30% compared to R-GCN. With BRG, the performance gains exceeded 50% for both. While all other methods could not effectively use the additional information supplied over the underlying biological regulation for the prediction of the pathways and genes, BioKGC achieved the highest performance. We hypothesize that BioKGC outperformed embedding methods because it prioritizes relational paths between entity groups of interest, rather than treating each triplet equally. This approach is crucial for noisy KGs, which may contain non-essential triplets like those in the BRG. Notably, utilizing the BRG in BioKGC enhanced the prediction of gene functional annotations.

For a more challenging application of predicting drugs for disease treatment, we applied BioKGC to a zero-shot prediction scenario defined by TxGNN, a state-of-the-art method. TxGNN's training involves pre-training and fine-tuning. In pre-training, all edges are used to learn node embeddings with equal importance. During fine-tuning, only triplets with specific relations (`indication`, `contraindications`, and `off-label use`) are used, focusing on drug and disease nodes. Additionally, disease nodes with limited molecular characterization are enhanced with a disease similarity auxiliary embedding in a gated manner, based on the nodes' degree. In BioKGC, non-drug-disease triplets were used for message passing as a BRG but not for supervision, allowing a direct focus on relevant relations. Unlike TxGNN, BioKGC did not need auxiliary embeddings; meaningful predictions could be made as long as there was some path connecting the disease entities and target drugs. BioKGC outperformed TxGNN in all five zero-shot disease splits (adrenal gland, anemia, cardiovascular, cell proliferation, and mental health). The average AUPRC increase was 20.2%, highlighting the effectiveness of path-based reasoning in predicting indications. BioKGC also achieved higher Recall@k values, indicating better prioritization of known treatments in the

top predictions. However, BioKGC showed slightly higher performance variance, likely due to its reliance on available edges for learning path representations, compared to TxGNN's use of stable auxiliary node embeddings for disease similarity.

The interpretability tool based on path gradients provides insights into drug-disease predictions like Bosutinib for ALL, validating hypotheses such as Acitretin for gastric cancer and Nicotin and Bupropion for AD. By highlighting essential paths, such as those involving SMC1A and POLA1 for Clofarabine's mode of action, the model captures meaningful associations. Supplying a custom graph during inference, where similar diseases to ALL are removed, demonstrates path-based methods' flexibility compared to embedding-based approaches. This semi-inductive experiment hints at the potential for fully inductive KG reasoning, promising in scenarios where new nodes appear during inference. Future studies will explore fully inductive reasoning tasks further.

BioKGC's limitations include potential biases encoded in training data. In the case of AD, nearly all known treatments were among the top predictions, yet FDA-approved drugs absent from PrimeKG, like Pramiracetam, Memantine, and Lecanemab, weren't retrieved due to lacking disease indications. As these drugs weren't connected in the database, the model couldn't learn from their associations, thus failing to identify them as potential treatments. Future enhancements might involve excluding dominant relations like indication from message passing and prioritizing molecular interactions to uncover disease-modifying treatments more effectively. Furthermore, while the learned paths suffice for treatment prediction, they provide limited insight into the specific molecular mechanisms post-drug intake, such as subsequent signaling cascades. Enhancements in biomedical KG for drug prediction could involve: 1) sparsifying and removing dominant and uninformative edges, and 2) augmenting training data with detailed molecular interaction information at the protein level, replacing generic protein-protein interactions with specific relations like "phosphorylation", "ubiquitination", "up-regulation", and "down-regulation" sourced from databases such as Biogrid.

Of note, under the open-world assumption, evaluating model performance metrics on incomplete knowledge graphs (KGs) may not fully reflect the model's capabilities, as performance metrics often rank known positive ground truth items against unknown triplets. This can lead to inconsistencies where positive items absent in the KG rank higher than known test answers, thus affecting performance rankings. Yang et al. [108] discuss these issues, noting that metrics like Mean Reciprocal Rank (MRR) degrade in high incompleteness scenarios, particularly when missing links correlate with specific entities. In biomedical KGs, gaps in knowledge about biological processes are unevenly distributed, with some entities or entity types receiving more attention than others due to factors such as prevalence or complexity.

Finally, BioKGC allows us to visualize predictions by analyzing gradients along the paths, pinpointing the most influential nodes and relations. This interpretability feature enables verification and explanation of results, which is lacking in traditional embedding methods. Interpretability also enables hypothesis generation and verification of their plausibility, such as proposing new drug-disease relationships, discovering new synthetic lethality pairs, and unraveling novel lncRNA-mRNA interactions.

Conclusion

In conclusion, this study serves as a compelling demonstration of the efficacy of biological knowledge graph reasoning for functional prediction, validating the concept. We then showcased BioKGC's ability to prioritize drugs to treat diseases in a zero-shot prediction scenario, outperforming the state-of-the-art method. Importantly, we illustrate the strength of BioKGC's interpretability tool to visualize important paths for a prediction, allowing for better understanding, uncovering training or data biases, and evaluating it for biological plausibility. Future work could focus on evaluating BioKGC performance in inductive settings, tailoring the KG to include more informative sources of information, and using fine-grained relations. It might also provide helpful using condition-specific KGs for reasoning, with detailed knowledge on tissue, patient, pathway and disease such that is possible to query from BioCypher [109]. Other directions include integrating node features in learning path representations, for instance, to allow adding experimental sequencing data to inform predictions further. An extension might further investigate the reasoning on hypergraphs with edges connecting more than two nodes, such as the case with polypharmaceutical effects when two drugs are used together, causing an unwanted effect.

4 Methods

4.1 Knowledge Graph Completion

A knowledge graph $KG = \{(u, r, v)\}_{u,v \in E, r \in R}$ is a heterogeneous directed graph with entities E as nodes, relations R, and a list of triplets (u, r, v) that represent the edges in the graph. Here, u (head) and v (tail) are entities, and r (relation) is an edge or link. The graph is considered heterogeneous because different entities may have different types, e.g. a node representing a gene versus a node representing a disease. The graph is directed because (u, r, v) being in the KG does not imply (v, r, u) is contained as well. Knowledge graph completion involves predicting missing the missing links, i.e. triplets, categorized into three tasks: 1) Tail prediction $(u, r, ?)$ - predicting the tail entity given the head entity and the relationship; 2) Head prediction $(?, r, v)$ - predicting the head entity given the relationship and the tail entity; 3) Relation prediction $(u, ?, v)$ - predicting the relationship given the head and tail entities [110].

4.2 Neural Bellman-Ford Network (NBFNet)

Our newly developed BioKGC is a path-representation learning-based method for graph completion built upon the NBFNet framework. Below, we describe the foundations of NBFNet to provide an understanding of how BioKGC operates.

Unlike node embedding methods or node GNN encoders that infer links between entities in a KG by learning node representations in an embedding space, NBFNet is a general graph neural network framework that performs link prediction by learning representations for each path from the query entity to potential tail entities. This is done by considering the relations along each possible path between the query entity and the potential tail entity. Given a KG, NBFNet learns to predict the tail node for a query $(u, q, ?)$. It does this by learning a node pair representation $h_q(u, v)$ for

nodes u and v, which captures all the possible paths between u and v conditioned on q. In NBFNet, the path formulation is represented by a generalized sum of path representations between u and v, denoted by $h_q(P)$, with a commutative summation operator $\bigoplus$:

$$h_q(u,v) = h_q(P_1) \oplus h_q(P_2) \oplus ..... \oplus h_q(P_M) \tag{1}$$

where M denotes the possible number of paths between u and v. Each path representation $h_q(P)$ is defined as the generalized product of the edges (or edges representations) belonging to that path, with the multiplication operator $\bigotimes$.

$$h_q(P) = w_q(e_1) \otimes w_q(e_1) \otimes ..... \otimes w_q(e_p) \tag{2}$$

where p is number of edges $e_1, ...e_p$ belonging to path P. In compact form, the path formulation can be written as:

$$h_q(u,v) = \bigoplus_{p \in P_{u,v}} \bigotimes_{e \in P} w_q(e) \tag{3}$$

This, in practice, means that we compute the representation between nodes u and v under query relation q by considering every path between the source node u and the target node v. Each path is represented by the product of edge representations w_q along that path, and we sum these products to obtain the final representation.

Two key factors contribute to NBFNet's scalability for large graphs and its effectiveness in learning tasks: the use of the generalized Bellman-Ford dynamic programming framework for path representation and the abstraction of this process into a neural formulation.

Generalized Bellmann-Ford path representation

The approach above, which predicts a link between a head and tail node by enumerating all possible paths and summing their contributions, is not scalable. As the network grows, the number of possible paths between the head and tail nodes increases exponentially, making this method impractical for larger networks. To achieve a scalable path formulation, NBFNet utilizes the Bellman-Ford dynamic programming algorithm, which efficiently finds the shortest path from a single source node to all other nodes in a weighted graph through a recursive process [111]:

$$d[v] = min(d[u] + w(u,v)) \tag{4}$$

By extending this equation to generalize the addition operator to any summation operator $\bigoplus$ and the minimum operator to any multiplication operator $\bigotimes$, we derive the generalized Bellman-Ford algorithm [112]. This generalization transforms the original Bellman-Ford algorithm for shortest path calculation into a versatile framework that simultaneously computes pair representations $h_q(u,v)$ for a given entity u, query relation q, and all vertices v in a graph. This approach reduces the computational cost to polynomial time relative to the number of nodes and edges in the graph. The formulation of the generalized Bellman-Ford algorithm is as follows:

$$h_q^{(0)}(u,v) \leftarrow \mathbb{1}_q(u=v) \tag{5}$$

$$h_q^{(t)}(u,v) \leftarrow (\bigoplus_{(x,v)\in\mathcal{E}} h_q^{(t-1)}(u,x) \otimes w_q(x,r,v)) \oplus h_q^{(0)}(u,v) \tag{6}$$

It first initializes the boundary condition on the source node (equation 5), representing the shortest path between u and v at the start. If the head and tail nodes coincide ($u = v$), the boundary condition is set to the generalized $\mathbb{1}$, which corresponds to 0 in the shortest path context (i.e., the shortest distance between a node and itself is zero) and to ∞ in the case $u \neq v$. Equation 6 describes the Bellman-Ford iteration, updating the shortest path distance between u and v. In each iteration, the representation from the previous layer $(t-1)$ is multiplied by the transition edge representation w_q to obtain the new representation $h_q(u,v)$. Here, W_q is a function of the relation between u and v. The boundary condition is added at each step to ensure accurate path formulation. The algorithm starts by propagating the boundary condition from the source node to its neighbors. Thanks to the distributive properties of the multiplication operator, all prefixes sharing this propagation are computed simultaneously. This iterative process continues, assessing potential target nodes, until all paths from the source node to the tail node are covered after t iterations, where t represents the path length.

Neural formulation

The generalized summation and multiplication operators are handcrafted. By abstracting the boundary condition in equation 5 to an indicator function, the multiplication operator in equation 6 to a message passing formulation, and the summation operator to a general aggregation function, NBFNet extends the generalized path formulation of the Bellman-Ford algorithm into a neural network framework. In this formulation, neural network functions parametrize the boundary condition, multiplication, and summation functions, allowing a generalized GNN framework to learn these three neural functions effectively.

$$h_v^{(0)} \leftarrow INDICATOR(u,v,q) \tag{7}$$

$$h_v^{(t)} \leftarrow AGGREGATE(\{MESSAGE(h_x^{(t-1)}, w_q(x,r,v)) | (x,r,v) \in \mathcal{E}(v)\} \cup \{h_v^{(0)}\}) \tag{8}$$

For the indicator function, NBFNet learns the embedding of the query relation q and assigns q to node v if v equals the source node u. For message passing, NBFNet borrows relational operators from KG embeddings, such as translation from TransE, multiplication from DistMult, and rotation from RotatE. Aggregation functions are implemented using permutation-invariant functions from GNN literature, including sum, mean, max, and principal neighborhood aggregation (PNA). Traditionally, edge representations are defined as transition probabilities or lengths. However, since an edge's contribution varies with the query relation, NBFNet parameterizes edge representations as a linear function of the query relation [51]. NBFNet can be interpreted

as a novel GNN framework for learning pair representations. Unlike common GNN frameworks, which compute pair representations as two independent node representations $h_q(u)$ and $h_q(u)$, in NBFNet, each node learns a representation conditioned on the source node. The learned pair representation $h_q(u, v)$ is then used to solve the link prediction problem, i.e., predicting the tail entity v given the head entity u and relation q. This is formulated as the conditional likelihood of the tail entity v as:

$$p(v|u, q) = \sigma(f(h_q(u, v))) \tag{9}$$

where where $\sigma(_)$ is the sigmoid function and $f(_)$ is a feed-forward neural network.

4.3 Biomedical Knowledge Graph Completion (BioKGC)

Considering the unique characteristics of biological knowledge graphs (KGs), we introduce BioKGC (Code availability at https://github.com/emyyue/BioKGC), a graph neural network framework based on NBFNet [51]. BioKGC is designed to predict missing links in biomedical KGs and is applied to four biological tasks:

1. *Gene Function Prediction*: Identifying potential novel functions for genes via gene-pathway associations.
2. *Drug Repurposing*: Discovering new indications for existing drugs by analyzing drug-disease associations for established drugs.
3. *Synthetic Lethality Prediction*: Identifying novel synthetic lethality pairs.
4. *lncRNA-Gene Target Prediction*: Prediction of new biological roles for lncRNAs by identifying novel gene-lncRNA relationships.

Path representation in BioKGC

As path-based reasoning method, path representations $h_q(u, v)$ in BioKGC are learned starting at a source node u to all potential target nodes v based on the relations r along the path, following the NBFNet parametrization (equations 7 and 8) but with an important enhancement which makes BioKGC more suited for biological KGs. Firstly, we make use of entity type information, which is not used in NBFNet originally. Secondly, we pool additional data sources beyond the target KG to augment the knowledge available during reasoning for the target link prediction task. Specifically, given a KG G_1 for which we wish to predict the missing links we add an additional graph G_2 such as biological regulatory graph (BRG) into the path representation computation. These augmentation to the original NBFNet method constitute our BioKGC.

The incorporation of an external biological regulatory graph (BRG), e.g. protein-protein interaction or gene regulatory network, provides additional edges (knowledge) that are used solely for message passing, enhancing the prediction of links of interest (Figure 1C). For example, in predicting the missing link (u, r, v), where (u, r, v) always comes from the target KG G_1 , messages can be passed only along paths in G_1 yielding a prediction path such as in Figure 1D. Alternatively, a BRG G_2 can be supplied to further information about the predictions, leveraging other knowledge bases (e.g., including relations between type 2 and 3 nodes), as illustrated in Figure 1E).

Consequently, in BioKGC, equation 8 is modified to always take $(u, r, ?)$ from G_1 but do aggregate and send messages across all edges $G_1 \cup G_2$. In equation 8 the edges $(x, r, z) \in \mathcal{E}(v)$ come from $G_1 \cup G_2$ rather than just G_1.

After performing link prediction with the modified message passing scheme, BioKGC ranks all candidate tail entities according to their likelihood $p(v|u, q)$, to form a true triplet with a given head entity and relation q as the query. During training of the model in a supervised setting, the negative log likelihood is minimized between positive samples $\langle u, r, v\rangle$ (i.e. known triplets composed of a head node tail node and relationship between them) and negative samples $\langle u, r, v'\rangle$, which are generated by corrupting v (i.e., substituting the true v with another node v').

$$L_{KG} = -logp(u, q, v) - \sum_{i=1}^{n} \frac{1}{n} log(1 - p(u, q, v_i')) \tag{10}$$

where n is the number of negative samples per positive sample and (u, q, v_i') is the i-th negative sample for knowledge graphs. The same approach is used for the prediction of v given u and r^{-1} for the reverse relation of r. Unlike the original NBFNet, BioKGC implements an entity-type aware negative sampling scheme. This means that when sampling negative v' for training, we consider the node type and sample v' only from the same type as v. This approach dramatically reduces the sample space and allows the model to learn better decision boundaries by focusing on sufficiently difficult negative samples. Importantly, the additional edges from the BRG are not used for sampling positive and negative triplets, ensuring that the computation of the loss in equation 10 remains unchanged.

4.4 Interpretation of prediction - Visualization of most important paths

Leveraging the NBFNet framework, BioKGC predictions can be directly interpreted through paths. This feature is crucial for biomedical tasks, where understanding the mechanisms behind each prediction is essential. These interpretations highlight the paths that most significantly contribute to the prediction $p(v|u, q)$. Using local interpretation methods, we approximate the local landscape of NBFNet with a linear model over the set of all paths, akin to a 1st-order Taylor polynomial. The importance of a path is defined by its weight in the linear model, which can be computed as the partial derivative of the prediction with respect to the path [51]. Formally, the top-k path interpretations for $p(u, q, v)$ are defined as:

$$P_1, P_2, ..., P_k = top - k_{P \in P_{uv}} \frac{\partial p(u, q, v)}{\partial P} \tag{11}$$

While directly computing the importance of all paths is intractable, NBFNet approximate them with edge importance. Specifically, the importance of each path is approximated by the sum of the importance of edges in that path, and therefore intuitively the top-k path interpretations are equivalent to the top-k longest paths on the edge importance graph.

For the visualization plot, we consider the top 10 most important paths ranked by gradient, with the edge width reflecting the number of times the edge appears in paths. Furthermore, the most important path is highlighted in red. In summary, our interpretability allows predictions to be assessed by their biological plausibility for hypothesis generation or validation in the laboratory.

4.5 KG construction, Data pre-processing and BioKGC training

Gene function prediction task

For this task, we used the knowledge graph (KG) from the KEGG database, extracted from ConsensuPathDB to train the model on gene (G) - pathway (P) interactions. Additionally, we utilized a biological regulatory graph (BRG) containing regulatory relationships between gene-gene (G-G), gene-chemical (G-C), and chemical-chemical (C-C) obtained from Pathway Commons [113, 114]. These interactions were represented as triplets in the format `(node1, relationType, node2)`, such as `(BABAM1, interacts-with, PSMD14)`. The BRG consisted of 30,885 nodes, of which 19,554 were genes and 11,331 were small molecules. It encompassed 1,884,146 relations of 13 unique types (Table A1), including interactions such as *controls phosphorylation of*, *interacts with*, and *chemical affects* (Table A2). During data pre-processing, we removed KEGG pathways with fewer than 10 annotations to genes. Next, we loaded both the BRG and KEGG bipartite graph (of pathways and genes [P-G]) as a multigraph in the network, maintaining only the biggest connected component, thereby removing 11 nodes present in components of only 2–3 nodes. We trained our BioKGC model using 70% of the P-G triplets, which were randomly split, with and without incorporating the underlying BRG as a message-passing graph. For validation, we used 10% of the P-G triplets, and the remaining 20% were reserved for testing. When the BRG was not utilized, we took an additional step to exclude triplets containing genes that appeared in the validation or test sets but were absent from the training set. The following hyperparameters of BioKGC were optimized for: adversarial temperature (corresponding to the temperature in self-adversarial negative sampling) $\{0.5, 1.0, 2.0\}$; negative samples: $\{32, 64\}$; aggregation function: sum, PNA (Principal Neighborhood Aggregator [115]; number of hidden layers: $\{4, 5, 6\}$; hidden layer dimension: $\{32, 64\}$; learning rate $\{5e-3, 1e-3, 5e-4\}$. Optimal hyperparameter set of BioKGC {'layers': 6, 'hidden_dim': 32, 'num_negatives': 32, 'lr': 0.005, 'adversarial_temperature': 0.5, 'batch_size': 8, 'epochs': 8}. Hyperparameters were optimized on the validation MRR.

Drug repurposing task

For this task, we used The PrimeKG database, an extensive, multimodal knowledge graph designed to integrate and unify diverse types of resources and biomedical and clinical data, such as gene-gene interactions, gene-disease associations and drug-disease information [48] (Figure 3A). In detail, we considered the PrimeKG database spanning 129,375 nodes from 10 different node classes and 8,100,498 edges from 27 different relationship types. A summary of the node and edge relations can be found in [49].

For training and evaluating BioKGC we used the same data and data splits, as defined by TxGNN, a geometric deep learning model for zero-shot drug repurposing predictions also based on PrimeKG. Five distinct zero-shot disease areas were used: adrenal gland disorders, anemia, cardiovascular diseases, cell proliferation issues, and mental health conditions. A disease area refers to a specific list of related diseases. We utilized TxGNN's data split to create training, validation, and testing datasets reflecting a zero-shot prediction scenario, in a proportion 0.83:0.12:0.05. By using the TxGNN code (https://github.com/mims-harvard/TxGNN from Apr 13, 2023, commit "1000aac"), the different splits were created by removing all triplets with the relation types `indication`, `contraindications`, and `off-label use` for a disease area from the training dataset, along with 95% of connections to biomedical entities such as proteins and phenotypes [49]. This split simulates minimal molecular characterization of a disease area combined with no knowledge of therapeutic opportunities. Next, TxGNN constructs the reverse edges. However, we removed the reverse relations since BioKGC inherently adds reverse triplets (and the reverse relations) during reasoning.

Only edges between drugs and diseases—such as `indication`, `contraindication`, and `off-label use`—were used to train BioKGC in a supervised manner. The remainder of the PrimeKG graph served as the BRG for message passing, similar to the gene function prediction task, as using a BRG graph was shown to enhance model performance. After removing reverse relations, approximately 5.7 million directed edges were used exclusively for message passing in the BioKGC model for each prediction setting (Table B3). Unlike the triplets used for supervised training, these edges were not drug-disease connections but rather protein-protein or disease-disease relations. On average, the training and validation sets contained about 33,000 and 4,000 edges, respectively, across the five examined disease areas (Table B3).

We further created our custom data split with TxGNN's "disease_eval" code to evaluate the performance in predicting drugs for the neurodegenerative disorder Alzheimer's disease (AD). Drugs that were associated to various AD diseases were moved to the test set (Table B6). All models were trained for 10 epochs, employing an early stopping mechanism that retained the best model based on validation set performance (MRR, see below). The final hyperparameters for all five disease area splits are reported in Table B5, where the last line reports the best hyperparameters set for the Alzheimer's disease area split.

Experiments were repeated for five different data split seeds, using the exact seeds employed by TxGNN to ensure a fair comparison. Each data split seed resulted in slightly different training, validation sets for each disease area, due to the random removal of edges to simulate the zero-shot scenario for the disease under study every time. Performance metrics were reported on the test set as the mean ± standard deviation across the five seeds.

Synthetic lethality (SL) prediction task

For training and inference of BioKGC, we used the SynLethDB-v2.0 [99] data as knowledge graph, an updated database compiling SL relationships derived from screening experiments, as well as computational predictions, providing a comprehensive resource for exploring gene interactions in cancer. BioKGC was trained on

preprocessed data, i.e. the SL gene pairs extracted from SynLethDB-v2.0, as provided by KR4SL, a knowledge graph-based model designed to predict SL interactions in cancer [50] and downloaded from their GitHub repository (https://github.com/JieZheng-ShanghaiTech/KR4SL from Dec 8, 2023 - commit "61b5c84"). In detail, the SL gene pairs from SynLethDB-v2.0 for human were randomly split into train, validation and test triples in a ratio 7:1:2, following the data split of KR4SL (see Table C8 for more details). The pre-procesed data was modified to fit the BioKGC format by removing the reverse edges introduced by KR4SL for SL gene pairs (following the same pre-processing scheme in the drug repurposing task), as reverse edges are implicitly added by BioKGC by default.

On top of the SL pairs, SynLethDB-v2.0 constructs a KG including relations between gene entities, pathways, and three types of Gene Ontology (GO) terms: biological processes (BP), molecular functions (MF), and cellular components (CC) augmented using OntoProtein [45] with the specific relation types and corresponding statistics summarized in Table C7 and C11. While SL pairs were used for supervised training, the rest of the KG derived from SynLethDB-v2.0 was used as BRG for message passing only, following the same scheme from previous tasks.

We train BioKGC for 15 epochs and for five random seeds and tuned hyperparameters such as the number of hidden layers in $\{4, 5\}$, the number of sampled negatives per positive example in $\{32, 64\}$, and the adversarial temperature in $\{0.5, 1.0, 2.0, 5.0\}$. Optimal hyperparameter set of BioKGC {'layers': 5, 'hidden_dim': 32, 'num_negatives': 64, 'lr': 0.005, 'adversarial_temperature': 0.5, 'batch_size': 64 , 'epochs': 15}.

Another set of experiments were based on setting a confidence threshold (called the "r.statistic_score" in SynLethDB-v2.0) on the SL pairs, varying from 0.1 to 0.8. As the SL pairs in the database come from various sources, we compared the performance of both methods across thresholds. For instance, by removing all pairs below a threshold of 0.3, we excluded most computational predictions, allowing the algorithms to learn from high confidence pairs derived from low- and high-throughput experiments. This data scheme is referred to as thresholded data, as opposed to unthresholded data, i.e. SL pairs directly from SynLethDB-v2.0 which were not filtered by their corresponding confidence score. For fair comparison, we run BioKGC for the same number of epochs using the same seeds and tuned hyperparameters on unthresholded data which later were used to train the model on thresholded data.

Since SL relationships are symmetric between genes, the final score for a gene v to be an SL partner of gene u is computed by considering both the transformed representation of tail node v given head node u and the SL relation, $f(h_q(u, v))$, as well as the transformed symmetric representation, $f(h_{q^{-1}}(u, v))$. Thus, the final SL score for a gene v to be an SL partner of gene u from BioKGC is reported as:

$$p(v|u) = \sigma\left(\frac{f(h_q(u, v)) + f(h_{q^{-1}}(u, v))}{2}\right) \tag{12}$$

LncRNA-gene target prediction task

For this task, we used the LncTarD 2.0 [116], a manually curated database of 8360 experimentally supported functional lncRNA-target regulatory interactions in human diseases, categorized into seven mechanisms of lncRNA-target regulation: *ceRNA or sponge*, *chromatin looping*, *epigenetic regulation*, *expression association*, *interact with mRNA*, *interact with protein* and *transcriptional regulation*. First, incomplete gene information, such as missing Ensembl IDs or gene names, was resolved via Gencode and HGNC mapping. Second, as only 12 pairs of regulations belonged to the chromatin looping category, we re-labeled them as transcriptional regulation after manually inspecting every regulatory interaction in the scientific literature. This condensed the interaction relationships into six distinct types.

A KG was constructed from LncTarD 2.0, where entities are the genes involved, and relations are the regulatory mechanisms. For a triplet (u, r, v), the head u corresponds to the regulator (e.g. lncRNA), the relation r to the regulation mechanisms, and the tail v to the target gene. The KG constructed by LncTarD 2.0 consists mainly of a densely connected subgraph of 2646 genes and 6084 involved regulations, surrounded by many isolated subgraphs containing only two or three genes. Only the largest connected component was kept and split into train, validation, and test set using the python package PyKeen [117], at a ratio of 0.8, 0.1, and 0.1, while ensuring that all nodes were included in the training set with at least one triplet. The specific numbers of nodes and edges for each subgraph are detailed in Table 3.1 (supplementary material).

On top of the LncTarD 2.0 KG, we added the BRG derived from PathwayCommons, the same used for the gene function prediction task. As there is no direct link between small molecules and lncRNA, we only used PPI, discarding other types of relations. This enriches the original KG with additional connectivity. In the end, BioKGC was trained on the lncRNA interactions from LncTarD 2.0 KG, while the BRG graph contained $15,806$ nodes and $1,035,133$ edges. To enable node type-aware negative sampling, node entities were labeled with six different categories: *lncRNA*, *mRNA*, *microRNA*, *transcription factor*, *protein* and *protein_ppi* (this last one to specifically identify nodes from ppi interactions from the BRG, whose edges are only used for message passing and not supervised training). Optimal hyperparameter set of BioKGC: {'layers': 6, 'hidden_dim': 32, 'num_negatives': 32, 'lr': 0.005, 'adversarial_temperature': 0.5, 'batch_size': 64, 'epochs': 10}.

Further, we ranked the interaction partners of 42 lncRNAs, including PVT1, sourced from the studies of [107] and [106]. We considered a regulatory relationship between the lncRNA of interest and its target genes, when there existed any relation r for which the conditional probability exceeded a threshold t, $p(v|u, q) \geq t$. Conversely, if no such relation exists, it is considered that there is no regulatory relationship. The threshold we used here was 0.45, which is the average probability outputted by BioKGC. These probabilities approximate an exponential normal distribution and were also used for random sampling constituting a random baseline.

4.6 Comparison with baselines

Knowledge Graph Embedding (KGE) methods

We compared BioKGC against the following methods: Among the KG Embedding methods, we benchmarked TransE, DistMult and RotatE, belonging to shallow models learning embeddings with an encoder for each relation and node. The latent embedding space is restricted by the semantic relationship r between u and v nodes.

TransE

The key idea of TransE is to model relationships as translations in the embedding space. For each triplet (u, r, v), the relationship r is modeled as a translation vector such that the embedding of the tail entity v should be close to the embedding of the head entity u plus the embedding of the relationship r: $\mathbf{u} + \mathbf{r} \approx \mathbf{v}$. The plausibility of a triplet (u, r, v) is determined by a scoring function based on the distance between $\mathbf{u} + \mathbf{r}$ and v: $f(u, r, v) = \|\mathbf{u} + \mathbf{r} - \mathbf{v}\|_2$. TransE is trained to minimize the distance for correct triplets and maximize the distance for incorrect triplets using a margin-based ranking loss [76]. The following hyperparameters were used: {'embedding_dim': 512, 'num_negatives': 512, 'lr': 0.0005, 'epochs': 100}.

RotatE

RotatE is a KGE model that represents entities as complex-valued vectors and relationships as rotations in the complex plane. This allows the model to capture various relational patterns, including symmetry, antisymmetry, inversion, and composition. Each entity u and v (head and tail) is represented as a complex vector, and each relationship r is represented as a complex vector with a modulus of 1, representing a rotation in the complex plane. For a triplet (u, r, v), the relationship r rotated u to match v $v \approx u$, where $\circ$ denotes the element-wise (Hadamard) product. The plausibility of a triplet (u, r, v) is determined by a scoring function based on the distance between u and v: $f(u, r, v) = \|\mathbf{u} \circ \mathbf{r} - \mathbf{v}\|_2$. RotatE is trained to minimize the distance for correct triplets and maximize the distance for incorrect triplets using a margin-based ranking loss [38]. For the comparison with BioKGC, the following parameters were used: dim = 512, max score = 6, exam = 512, and lr = 5e-4. The following hyperparameters were used: {'embedding_dim': 512, 'num_negatives': 512, 'lr': 0.0005, 'epochs': 100}.

DistMult

DistMult (Multiplicative Distance) is a knowledge graph embedding model that represents entities and relationships as vectors and captures interactions through a multiplicative mechanism. Each entity u and v (head and tail) and each relationship r is represented as a vector in the embedding space. For a triplet (u, r, v), the relationship r interacts multiplicatively with the head entity u to predict the tail entity v: $v \approx u \circ, where \circ$ denotes the element-wise (Hadamard) product. The plausibility of a triplet (u, r, v) is determined by a scoring function that measures the similarity between $u \circ and$v:f(u,r,v) = -$\|\mathbf{u} \circ \mathbf{r} - \mathbf{v}\|_2^2$. DistMult is trained to maximize the score for correct triplets and minimize the score for incorrect triplets using, similarly to

TransE, a margin-based ranking loss [35]. The following hyperparameters were used: {'embedding_dim': 512, 'num_negatives': 512, 'lr': 0.001, 'L_3 regularization' = 1e-4, 'epochs': 100}.

4.6.1 Graph neural network-based methods

Relational Graph Convolutional Networks (R-GCNs)

R-GCN is a variant of the Graph Convolutional Network (GCN) specifically designed to handle graphs with multiple types of relations. R-GCN extends GCN to incorporate relational data, making it suitable for KGs. R-GCN employs a neural encoder to learn the representation for each node, while the decoder remained the scoring functions from the KGE models. Initially, the embeddings for each node are initialized [118]. During each layer of the GNN's message passing, messages are passed, aggregated, and used to update the node's embedding. The messages are obtained by applying a $W_{r,M}^{(l)}$ matrix on the embeddings of the previous layer $h_i^{(l-1)}$ separately for each relation r, allowing R-GCN to handle graphs with multiple types of edges by learning separate weight matrices for each relation type: $m_{r,i}^{(l)} = W_{r,M}^{(l)} h_i^{(l-1)}$. Next, the incoming messages of each node v_i from the neighboring nodes $N_r(i)$ are aggregated by the taking the average: $\widetilde{m_{r,i}^{(l)}} = \frac{1}{|N_r(l)|} \sum_{j \in N_r(i)} m_{r,i}^{(l)}$. Finally, the node's embedding from the previous layer is updated by $h_i^{(l)} = h_i^{(l-1)} + \sum_{r \in T_R} \widetilde{m_{r,i}^{(l)}}$. This process ensures that each node's embedding is updated with information from its neighbors, considering the type of relation connecting them. The final embeddings of the nodes are derived after L layers of propagation. The hyperparameters for R-GCN were optimized using grid search on the parameters *num_negatives*, *lr* and *hidden_dim*. Two separate searches were conducted: one without the BRG, yielding the best set as {'num_negatives': 64, 'lr': 10^{-4}, 'layers': 3, 'hidden_dim' = 256}, and another with the BRG, resulting in the best set as {'num_negatives': 32, 'lr': 10^{-5}, 'layers': 5, 'hidden_dim' = 128}.

Transductive Graph Neural Network (TxGNN)

TxGNN is a graph neural network-based model designed to predict drug-disease relationships in zero-shot scenarios where there is minimal prior information or treatment history. Utilizing PrimeKG, a comprehensive biomedical knowledge graph, TxGNN employs R-GCNs to learn embeddings of drugs and diseases, capturing complex interactions by mapping them into a shared latent space. This allows TxGNN to predict drug-disease interactions, focusing on indications and contraindications, even for diseases with limited molecular characterization. It achieves this through the use of disease signature vectors, adaptive embedding aggregation, and iterative optimization of non-linear transformations. TxGNN improves drug-disease prediction by splitting training into pre-training and fine-tuning phases. In pre-training, all triplets are used, while fine-tuning focuses on drug-disease relations like `indication`, `contraindications`, and `off-label use`. This is in contrast to BioKGC, where we do not divide our approach into these two phases or perform pre-training. Instead, we use all triplets containing non-drug-disease relations in the message passing BRG

without considering them for supervision. TxGNN was trained using the hyperparameters provided in the original publication: pretraining: {'lr': $1e-3$, 'batch_size': 1024, 'pretrain epochs': 2}; finetuning: {'lr': $5e-4$, 'hidden_dim': 512, 'batch_size': 1024, 'pretrain epochs': 500}. The hyperparameters used in fine-tuning were the same for all disease splits.

After quantitatively evaluating BioKGC, we examined the top drug predictions for acute lymphoblastic leukemia (ALL) and gastric cancer within the cell proliferation split and investigated known and unknown predictions. We discovered that the data split provided by PrimeKG only mirrored a near zero-shot prediction scenario since some treatments for diseases remained available. Thus, in a small experiment investigating known treatments, we further excluded triplets of similar diseases from the inference graph (same as the training graph) by excluding any disease whose name contained the string *leukemia, neoplasm, cancer, teratoma, lymphoma, lymphosarcoma, cytoma, tumor*, or *carcinoma*. This approach allowed us to study which path the model takes when disease similarity cannot be leveraged for prediction, hoping to uncover biological insights. AD, as a devastating disease with little knowledge of its cause and treatment options, we evaluated the predictions with experts to further understand the advantages and limitations of BioKGC.

4.6.2 Path representation learning-based methods

Knowledge Representation for Synthetic Lethality (KR4SL)

KR4SL is a path-representation learning GNN-based method for the prediction of SL gene pairs. The framework begins with the construction of a heterogeneous KG that integrates genes, gene interactions, pathways, GO terms, and an SL graph augmented with reverse and identity edges. This graph is processed through an encoder that identifies relevant relational paths for each primary gene, combining structural graph information with textual semantics of entities. KR4SL utilizes a gated recurrent unit (GRU) to enhance the sequential semantics of the relational paths. As messages propagate through the network, they are enriched by fusing structural and textual information. An attentive aggregation mechanism evaluates the importance of different edges, focusing on the most essential paths for model explanations. Finally, a scoring decoder processes the semantic representations of candidate SL partners, generating scores to rank and select the top-k candidates as the most probable SL partners. The SL relation is symmetric, meaning that if (u, v) is an SL pair, then (v, u) is also an SL pair. The model computes the probability of gene v being lethal when gene u is already lethal: $p(v|u, r) = sigmoid(\mathbf{h}_u \cdot \mathbf{r} \cdot \mathbf{t}_v)$, where $\mathbf{h}_u$, $\mathbf{t}_v$ are the embeddings of gene u and gene v, and r is the embedding of the SL relation. Since SL is symmetric, the model also computes the reverse probability $p(u|v, r) = sigmoid(\mathbf{h}_v \cdot \mathbf{r} \cdot \mathbf{t}_u)$. The final score for a gene pair (u, v) is an aggregate of both probabilities: $SL(u, v) = p(u|v, r) + p(v|u, r)/2$.

We use the code from GitHub to train KR4SL in the transductive setting. We trained KR4SL on the original (unthresholded) data using the same hyperparameters specified by the authors on GitHub, 'weight decay rate': 0.000089, 'lr': 0.0011, 'batch_size': 50, 'epochs': 15, 'hidden_dim': 48, 'drop_out': 0}. The experiments were run five times with different seeds. The same set of parameters was used for thresholded data.

4.7 Model evaluation

Models were evaluated by ranking positive triplets $\langle h, r, t \rangle$ of the test set against all negative triplets $\langle h, r, t' \rangle$ that are not present in the KG, following the filtered ranking protocol [35]. Equally, for the reverse relation direction r^{-1}, positive triplets of the form $\langle t, r, h \rangle$ are ranked against all negatives $\langle t, r^{-1}, h \rangle$. This approach yields a very stringent evaluation that forces the model to rank positive samples highly. The metrics used are mean rank (MR), mean reciprocal rank (MRR), and hits at k (Hits@k). MR is calculated by averaging the ranks q of the positive samples among negatives. Lower values are ideal, with the best value at 1.

$$MR = \frac{1}{|Q|} \sum_{q \in Q} q \tag{13}$$

MRR is the average of the reciprocal rank $\frac{1}{q}$, placing less emphasis on lowly ranked triplets. The values range in $[0, 1]$, and the larger the value, the better the model.

$$MRR = \frac{1}{|Q|} \sum_{q \in Q} \frac{1}{q} \tag{14}$$

Finally, the metric Hits@k provides the probability of correct predictions in the top k predicted triplets made by the model. It can also be considered the percentage of positive triplets in the top k and is, therefore, equivalent to Recall@k. Like MRR, values range in $[0, 1]$, and the larger the value, the better the model.

$$H@k = \frac{\big|\{q \in Q : q < k\}\big|}{\big|Q\big|} \tag{15}$$

The best model is selected based on the highest MRR in a validation set.

While the conditional probability is modeled in KGC by predicting the tail entity t given the head entity h and the relation r, it might be sensible to evaluate under the joint probability of h, t, and r. Therefore, to remain consistent with the TxGNN we compared against in the drug prediction task, we also used AUPRC, which summarizes the precision and recall at different probability thresholds, as well as specificity and F1 score at the probability threshold of 0.5 (eq 16) using code inherent to TxGNN. They evaluate the performance per disease node.

$$Precision = \frac{TP}{TP + FP}, \; Recall = \frac{TP}{TP + FN}, \; F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{16}$$

TxGNN calculates AUPRC using two strategies. The first compares each positive ground truth item vs. one negative from the list of drugs of a disease area, referred to as `AUPRC 1:1`. The second considers all positive and negative ground truth items, referred to as `AUPRC`. The first metric conveys information on the model's ability to distinguish a positive from a randomly sampled negative in direct comparison. In contrast, the latter focuses on distinguishing positives and negatives across the whole dataset. Furthermore, the `AUPRC` accounts for the positive/negative class imbalance and provides a

balanced evaluation of precision and recall. We chose to predominately use the latter, which reflects a real-world scenario of identifying therapeutical opportunities within a given set of drugs.

For SL prediction task, we compared seed-wise performance of our model with performance of KR4SL using metrics inherent to KR4SL framework's code: NDCG@k, Recall@k, and Precision@k. Moreover, we computed MRR from equation 14 for both BioKGC and KR4SL by first calculating MRR for each query gene and then averaging gene-wise MRR over all query genes.

Recall@k here is utilized to evaluate how effectively a model predicts SL relationships within its top k predictions and is defined as follows:

$$Recall@k = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{i=1}^{k}\mathbb{1}\{G_n^{KR}(i) \in G_n^{KP}\}}{|G_n^{KP}|} \tag{17}$$

where N is the total number of query genes, G_n^{KP} is the set of known SL partner candidates of gene n and G_n^{KR} is the set of top-k SL gene partners for gene n ranked by prediction scores meaning that $G_n^{KR}(i)$ is i-th predicted SL partner for gene n.

Respectively, Precision@k is employed to assess how effectively a model predicts SL relationships within its top-k predictions, normalized by either the size of the top-k predictions or the size of the known SL relationships, whichever is smaller. Precision@k is defined as follows:

$$Precision@k = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{i=1}^{k}\mathbb{1}\{G_n^{KR}(i) \in G_n^{KP}\}}{min\{|G_n^{KP}|, k\}} \tag{18}$$

NDCG@k is used to evaluate a model's capability to prioritize the top-k candidate genes associated with SL relationships and is defined as following:

$$NDCG@k = \frac{1}{N}\sum_{n=1}^{N}\frac{DCG@k(n)}{IDCG@k(n)} \tag{19}$$

where $DCG@k(n)$ and $IDCG@k(n)$ are defined as follows:

$$DCG@k(n) = \sum_{i=1}^{k}\frac{2^{\mathbb{1}\{G_n^{KR}(i) \in G_n^{KP}\}} - 1}{log_2(i+1)} \tag{20}$$

$$IDCG@k(n) = \sum_{i=1}^{min\{|G_n^{KP}|,k\}}\frac{1}{log_2(i+1)} \tag{21}$$

References

[1] Tavassoly, I., Goldfarb, J., Iyengar, R.: Systems biology primer: the basic methods and approaches. Essays in Biochemistry **62**(4),

487–500 (2018) https://doi.org/10.1042/EBC20180003 . _eprint: https://portlandpress.com/essaysbiochem/article-pdf/62/4/487/887339/ebc-2018-0003c.pdf

[2] Urry, L.A., Cain, M.L., Wasserman, S.A., Minorsky, P.V., Reece, J.B.: Campbell Biologie. Pearson Deutschland, Munich, Germany (2019). https://elibrary.pearson.de/book/99.150005/9783863268671

[3] Barabasi, A.-L., Oltvai, Z.: Network Biology: Understanding The Cell's Functional Organization. Nature reviews. Genetics **5**, 101–13 (2004) https://doi.org/10.1038/nrg1272

[4] Smith, J., Theodoris, C., Davidson, E.H.: A gene regulatory network subcircuit drives a dynamic pattern of gene expression. Science **318**(5851), 794–797 (2007). Publisher: American Association for the Advancement of Science

[5] Karlebach, G., Shamir, R.: Modelling and analysis of gene regulatory networks. Nature reviews Molecular cell biology **9**(10), 770–780 (2008). Publisher: Nature Publishing Group UK London

[6] Walhout, A.J.M., Vidal, M.: High-Throughput Yeast Two-Hybrid Assays for Large-Scale Protein Interaction Mapping. Methods **24**(3), 297–306 (2001) https://doi.org/10.1006/meth.2001.1190

[7] Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M., Goliaei, B., Peyvandi, A.A.: Protein-protein interaction networks (PPI) and complex diseases. Gastroenterology and Hepatology From Bed to Bench **7**(1), 17–31 (2014). Accessed 2024-05-09

[8] Piñero, J., Ramírez-Anguita, J.M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., Furlong, L.I.: The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic acids research **48**(D1), 845–855 (2020). Publisher: Oxford University Press

[9] Goh, K.-I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabási, A.-L.: The human disease network. Proceedings of the National Academy of Sciences **104**(21), 8685–8690 (2007). Publisher: National Acad Sciences

[10] Musawi, A.F.A., Roy, S., Ghosh, P.: A Review of Link Prediction Applications in Network Biology. arXiv. arXiv:2312.01275 [cs, q-bio] (2023). http://arxiv.org/abs/2312.01275 Accessed 2024-01-26

[11] Wang, Y., You, Z., Li, L., Chen, Z.: A survey of current trends in computational predictions of protein-protein interactions. Frontiers of Computer Science **14**, 1–12 (2020). Publisher: Springer

[12] Abbas, K., Abbasi, A., Dong, S., Niu, L., Yu, L., Chen, B., Cai, S.-M., Hasan, Q.:

Application of network link prediction in drug discovery. BMC Bioinformatics **22**(1), 187 (2021) https://doi.org/10.1186/s12859-021-04082-y . Accessed 2024-05-09

[13] Kumar, A., Singh, S.S., Singh, K., Biswas, B.: Link prediction techniques, applications, and performance: A survey. Physica A: Statistical Mechanics and its Applications **553**, 124289 (2020). Publisher: Elsevier

[14] Page, L., Brin, S., Rajeev, M., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab (1999)

[15] Jaccard, P.: Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull Soc Vaudoise Sci Nat **37**, 547–579 (1901)

[16] Katz, L.: A new status index derived from sociometric analysis. Psychometrika **18**(1), 39–43 (1953). Publisher: Springer

[17] Iván, G., Grolmusz, V.: When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks. Bioinformatics **27**(3), 405–407 (2011) https://doi.org/10.1093/bioinformatics/btq680 . Accessed 2024-05-04

[18] Lu, Y., Guo, Y., Korhonen, A.: Link prediction in drug-target interactions network using similarity indices. BMC Bioinformatics **18**, 39 (2017) https://doi.org/10.1186/s12859-017-1460-z . Accessed 2023-08-02

[19] Nelson, W., Zitnik, M., Wang, B., Leskovec, J., Goldenberg, A., Sharan, R.: To Embed or Not: Network Embedding as a Paradigm in Computational Biology. Frontiers in Genetics **10**, 381 (2019) https://doi.org/10.3389/fgene.2019.00381 . Accessed 2021-06-17

[20] Cho, H., Berger, B., Peng, J.: Compact Integration of Multi-Network Topology for Functional Analysis of Genes. Cell Systems **3**(6), 540–5485 (2016) https://doi.org/10.1016/j.cels.2016.10.017 . Accessed 2022-10-12

[21] Perozzi, B., Al-Rfou, R., Skiena, S.: Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, ??? (2014). https://doi.org/10.1145/2623330.2623732 . http://dx.doi.org/10.1145/2623330.2623732

[22] Grover, A., Leskovec, J.: node2vec: Scalable Feature Learning for Networks. KDD : proceedings. International Conference on Knowledge Discovery & Data Mining **2016**, 855–864 (2016) https://doi.org/10.1145/2939672.2939754 . Accessed 2021-04-13

[23] Su, C., Tong, J., Zhu, Y., Cui, P., Wang, F.: Network embedding in biomedical data science. Brief Bioinform **21**(1), 182–197 (2020)

[24] Mohamed, S.K., Nounu, A., Nováček, V.: Biological applications of knowledge graph embedding models. Briefings in Bioinformatics **22**(2), 1679–1693 (2021) https://doi.org/10.1093/bib/bbaa012 . Accessed 2022-11-21

[25] Zitnik, M., Leskovec, J.: Predicting multicellular function through multi-layer tissue networks. Bioinformatics **33**(14), 190–198 (2017) https://doi.org/10.1093/bioinformatics/btx252 . Accessed 2022-08-18

[26] Zitnik, M., Agrawal, M., Leskovec, J.: Modeling Polypharmacy Side Effects with Graph Convolutional Networks, 9 (2018)

[27] Ruiz, C., Zitnik, M., Leskovec, J.: Identification of disease treatment mechanisms through the multiscale interactome. Nat Commun **12**(1), 1796 (2021)

[28] Ietswaart, R., Gyori, B.M., Bachman, J.A., Sorger, P.K., Churchman, L.S.: Genewalk identifies relevant gene functions for a biological context using network representation learning. Genome Biology **22** (2019)

[29] Hu, Y., Rehawi, G., Moyon, L., Gerstner, N., Ogris, C., Knauer-Arloth, J., Bittner, F., Marsico, A., Mueller, N.: Network embedding across multiple tissues and data modalities elucidates the context of host factors important for covid-19 infection. Front Genet **13**, 909714 (2022)

[30] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs] (2013). arXiv: 1301.3781. Accessed 2022-03-31

[31] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. arXiv:1310.4546 [cs, stat] (2013). arXiv: 1310.4546. Accessed 2021-10-13

[32] Kipf, T.N., Welling, M.: Variational Graph Auto-Encoders. arXiv. arXiv:1611.07308 [cs, stat] (2016). http://arxiv.org/abs/1611.07308 Accessed 2023-10-23

[33] Kipf, T.N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. arXiv:1609.02907 [cs, stat] (2017). arXiv: 1609.02907. Accessed 2021-04-12

[34] Inductive Representation Learning on Large Graphs

[35] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating Embeddings for Modeling Multi-relational Data. In: Advances in Neural Information Processing Systems, vol. 26. Curran Associates, Inc., ??? (2013). https://proceedings.neurips.cc/paper/2013/hash/1cecc7a77928ca8133fa24680a88d2f9-Abstract.html Accessed 2023-05-12

[36] Yang, B., Yih, W.-t., He, X., Gao, J., Deng, L.: Embedding Entities and Relations for Learning and Inference in Knowledge Bases. arXiv. arXiv:1412.6575 [cs] (2015). http://arxiv.org/abs/1412.6575 Accessed 2023-05-12

[37] Trouillon, T., Welbl, J., Riedel, S., Gaussier, E., Bouchard, G.: Complex Embeddings for Simple Link Prediction. arXiv. arXiv:1606.06357 [cs, stat] (2016). http://arxiv.org/abs/1606.06357 Accessed 2024-01-28

[38] Sun, Z., Deng, Z.-H., Nie, J.-Y., Tang, J.: RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. arXiv:1902.10197 [cs, stat] (2019). arXiv: 1902.10197. Accessed 2022-04-13

[39] Schlichtkrull, M., Kipf, T.N., Bloem, P., Berg, R.v.d., Titov, I., Welling, M.: Modeling Relational Data with Graph Convolutional Networks. arXiv. arXiv:1703.06103 [cs, stat] (2017). http://arxiv.org/abs/1703.06103 Accessed 2023-06-07

[40] Consortium, T.U.: UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research **51**(D1), 523–531 (2022) https://doi.org/10.1093/nar/gkac1052 . _eprint: https://academic.oup.com/nar/article-pdf/51/D1/D523/48441158/gkac1052.pdf

[41] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.*: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000). Publisher: Nature Publishing Group

[42] Aleksander, S.A., Balhoff, J., Carbon, S., Cherry, J.M., Drabkin, H.J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N.L., *et al.*: The gene ontology knowledgebase in 2023. Genetics **224**(1), 031 (2023). Publisher: Oxford University Press US

[43] Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A.C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., *et al.*: DrugBank 4.0: shedding new light on drug metabolism. Nucleic acids research **42**(D1), 1091–1097 (2014). Publisher: Oxford University Press

[44] Zhang, R., Hristovski, D., Schutte, D., Kastrin, A., Fiszman, M., Kilicoglu, H.: Drug repurposing for covid-19 via knowledge graph completion. J Biomed Inform **115**, 103696 (2021)

[45] Zhang, N., Bi, Z., Liang, X., Cheng, S., Hong, H., Deng, S., Lian, J., Zhang, Q., Chen, H.: Ontoprotein: Protein pretraining with gene ontology embedding. arXiv preprint arXiv:2201.11147 (2022)

[46] Biswas, S., Mitra, P., Rao, K.: Relation prediction of co-morbid diseases using knowledge graph completion. IEEE/ACM Trans Comput Biol Bioinform **12**(2),

708–717 (2021)

[47] Lobentanzer, S., Aloy, P., Baumbach, J., Bohar, B., Carey, V., Charoentong, P., Danhauser, K., Dogan, T., Dreo, J., Dunham, I., Farr, E., Fernandez-Torras, A., Gyori, B., Hartung, M., Hoyt, C., Klein, C., Korcsmaros, T., Maier, A., Mann, M., Ochoa, D., Pareja-Lorente, E., Popp, F., Preusse, M., Probul, N., Schwikowski, B., Sen, B., Strauss, M., Turei, D., Ulusoy, E., Waltemath, D., Wodke, J., J., S.-R.: Democratizing knowledge representation with biocypher. Nat Biotechnol **41**(8), 1056–1059 (2023)

[48] Chandak, P., Huang, K., Zitnik, M.: Building a knowledge graph to enable precision medicine. Scientific Data **10**(1), 67 (2023) https://doi.org/10.1038/s41597-023-01960-3 . Accessed 2023-07-02

[49] Huang, K., Chandak, P., Wang, Q., Havaldar, S., Vaid, A., Leskovec, J., Nadkarni, G., Glicksberg, B.S., Gehlenborg, N., Zitnik, M.: Zero-shot prediction of therapeutic use with geometric deep learning and clinician centered design. preprint, Health Informatics (March 2023). https://doi.org/10.1101/2023.03.19.23287458 . http://medrxiv.org/lookup/doi/10.1101/2023.03.19.23287458 Accessed 2023-09-16

[50] Zhang, K., Wu, M., Liu, Y., Feng, Y., Zheng, J.: KR4SL: knowledge graph reasoning for explainable prediction of synthetic lethality. Bioinformatics **39**(Supplement_1), 158–167 (2023) https://doi.org/10.1093/bioinformatics/btad261 . Accessed 2024-04-01

[51] Zhu, Z., Zhang, Z., Xhonneux, L.-P., Tang, J.: Neural Bellman-Ford Networks: A General Graph Neural Network Framework for Link Prediction. In: Advances in Neural Information Processing Systems, vol. 34, pp. 29476–29490. Curran Associates, Inc., ??? (2021). https://proceedings.neurips.cc//paper/2021/hash/f6a673f09493afcd8b129a0bcf1cd5bc-Abstract.html Accessed 2022-04-13

[52] Kamburov, A., Wierling, C., Lehrach, H., Herwig, R.: ConsensusPathDB—a database for integrating human functional interaction networks. Nucleic Acids Research **37**(Database issue), 623–628 (2009) https://doi.org/10.1093/nar/gkn698 . Accessed 2022-09-29

[53] Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G.D., Sander, C.: Pathway commons, a web resource for biological pathway data. Nucleic Acids Research $\mathbf{39}(\text{suppl}_1), 685 - -690(2010)$

[54] Goldberg, R., Cary, M., Demir, E.: BioPAX A Community Standard for Pathway Data Sharing. Nature Biotechnology (28) (2010)

[55] Rodchenkov, I., Babur, O., Luna, A., Aksoy, B.A., Wong, J.V., Fong, D., Franz, M., Siper, M.C., Cheung, M., Wrana, M., Mistry, H., Mosier, L., Dlin, J., Wen, Q., O'Callaghan, C., Li, W., Elder, G., Smith, P.T., Dallago, C., Cerami, E.,

Gross, B., Dogrusoz, U., Demir, E., Bader, G.D., Sander, C.: Pathway commons 2019 update: integration, analysis and exploration of pathway data. Nucleic Acids Research **48**(D1), 489–497 (2019) https://doi.org/10.1093/nar/gkz946 https://academic.oup.com/nar/article-pdf/48/D1/D489/31697764/gkz946.pdf

[56] Gekakis, N., Staknis, D., Nguyen, H.B., Davis, F.C., Wilsbacher, L.D., King, D.P., Takahashi, J.S., Weitz, C.J.: Role of the clock protein in the mammalian circadian mechanism. Science **280**(5369), 1564–1569 (1998)

[57] Olkkonen, J., Kouri, V.-P., Kuusela, E., Ainola, M., Nordström, D., Eklund, K.K., Mandelin, J.: Dec2 blocks the effect of the arntl2/npas2 dimer on the expression of per3 and dbp. Journal of Circadian Rhythms **15** (2017)

[58] Kume, K., Zylka, M., Sriram, S., Shearman, L., Weaver, D., Xu, J.-J., Maywood, E., Hastings, M., Reppert, S.: mcry1 and mcry2 are essential components of the negative limb of the circadian clock feedback loop. Cell **98**, 193–205 (1999) https://doi.org/10.1016/S0092-8674(00)81014-4

[59] Griffin, E.A., Staknis, D., Weitz, C.J.: Light-independent role of cry1 and cry2 in the mammalian circadian clock. Science **286**(5440), 768–771 (1999) https://doi.org/10.1126/science.286.5440.768 https://www.science.org/doi/pdf/10.1126/science.286.5440.768

[60] Pui, C.-H., Evans, W.E.: Acute lymphoblastic leukemia. New England Journal of Medicine **339**(9), 605–615 (1998). Publisher: Mass Medical Soc

[61] Chessells, J.M., Harrison, G., Richards, S.M., Bailey, C.C., Hill, F.G., Gibson, B.E., Hann, I.M.: Down's syndrome and acute lymphoblastic leukaemia: clinical features and response to treatment. Archives of Disease in Childhood **85**(4), 321–325 (2001) https://doi.org/10.1136/adc.85.4.321

[62] Bielorai, B., Fisher, T., Waldman, D., Lerenthal, Y., Nissenkorn, A., Tohami, T., Marek, D., Amariglio, N., Toren, A.: Acute lymphoblastic leukemia in early childhood as the presenting sign of ataxia-telangiectasia variant. Pediatric Hematology and Oncology **30**(6), 574–582 (2013) https://doi.org/10.3109/08880018.2013.777949

[63] German, J.: Bloom's syndrome. XX. The first 100 cancers. Cancer Genetics and Cytogenetics **93**(1), 100–106 (1997) https://doi.org/10.1016/s0165-4608(96)00336-6

[64] Shah, A., John, B.M., Sondhi, V.: Acute lymphoblastic leukemia with treatment–naive Fanconi anemia. Indian Pediatrics **50**(5), 508–510 (2013)

[65] Sehgal, S., Mujtaba, S., Gupta, D., Aggarwal, R., Marwaha, R.K.: High incidence of Epstein Barr virus infection in childhood acute lymphocytic leukemia: a preliminary study. Indian Journal of Pathology & Microbiology **53**(1), 63–67

(2010) https://doi.org/10.4103/0377-4929.59186

[66] Terwilliger, T., Abdul-Hay, M.: Acute lymphoblastic leukemia: a comprehensive review and 2017 update. Blood Cancer Journal **7**(6), 577–577 (2017) https://doi.org/10.1038/bcj.2017.53 . Number: 6 Publisher: Nature Publishing Group. Accessed 2024-01-06

[67] Messina, M., Chiaretti, S., Iacobucci, I., Tavolaro, S., Lonetti, A., Santangelo, S., Elia, L., Papayannidis, C., Paoloni, F., Vitale, A., Guarini, A., Martinelli, G., Foà, R.: AICDA expression in BCR/ABL1-positive acute lymphoblastic leukaemia is associated with a peculiar gene expression profile. British Journal of Haematology **152**(6), 727–732 (2011) https://doi.org/10.1111/j.1365-2141.2010.08449.x

[68] Tanaka, Y., Kawazu, M., Yasuda, T., Tamura, M., Hayakawa, F., Kojima, S., Ueno, T., Kiyoi, H., Naoe, T., Mano, H.: Transcriptional activities of DUX4 fusions in B-cell acute lymphoblastic leukemia. Haematologica **103**(11), 522–526 (2018) https://doi.org/10.3324/haematol.2017.183152 . Accessed 2024-02-04

[69] Zhang, J., McCastlain, K., Yoshihara, H., Xu, B., Chang, Y., Churchman, M.L., Wu, G., Li, Y., Wei, L., Iacobucci, I., Liu, Y., Qu, C., Wen, J., Edmonson, M., Payne-Turner, D., Kaufmann, K.B., Takayanagi, S.-i., Wienholds, E., Waanders, E., Ntziachristos, P., Bakogianni, S., Wang, J., Aifantis, I., Roberts, K.G., Ma, J., Song, G., Easton, J., Mulder, H.L., Chen, X., Newman, S., Ma, X., Rusch, M., Gupta, P., Boggs, K., Vadodaria, B., Dalton, J., Liu, Y., Valentine, M.L., Ding, L., Lu, C., Fulton, R.S., Fulton, L., Tabib, Y., Ochoa, K., Devidas, M., Pei, D., Cheng, C., Yang, J., Evans, W.E., Relling, M.V., Pui, C.-H., Jeha, S., Harvey, R.C., Chen, I.-M.L., Willman, C.L., Marcucci, G., Bloomfield, C.D., Kohlschmidt, J., Mrózek, K., Paietta, E., Tallman, M.S., Stock, W., Foster, M.C., Racevskis, J., Rowe, J.M., Luger, S., Kornblau, S.M., Shurtleff, S.A., Raimondi, S.C., Mardis, E.R., Wilson, R.K., Dick, J.E., Hunger, S.P., Loh, M.L., Downing, J.R., Mullighan, C.G.: DEREGULATION OF DUX4 AND ERG IN ACUTE LYMPHOBLASTIC LEUKEMIA. Nature genetics **48**(12), 1481–1489 (2016) https://doi.org/10.1038/ng.3691 . Accessed 2024-02-04

[70] Mortlock, A.A., Wilson, D.M., Kettle, J.G., Goldberg, F.W., Foote, K.M.: 5.02 - Selective Kinase Inhibitors in Cancer. In: Chackalamannil, S., Rotella, D., Ward, S.E. (eds.) Comprehensive Medicinal Chemistry III, pp. 39–75. Elsevier, Oxford (2017). https://doi.org/10.1016/B978-0-12-409547-2.12391-1 . https://www.sciencedirect.com/science/article/pii/B9780124095472123911

[71] Knox, C., Wilson, M., Klinger, C.M., Franklin, M., Oler, E., Wilson, A., Pon, A., Cox, J., Chin, N.E.L., Strawbridge, S.A., Garcia-Patino, M., Kruger, R., Sivakumaran, A., Sanford, S., Doshi, R., Khetarpal, N., Fatokun, O., Doucet, D., Zubkowski, A., Rayat, D.Y., Jackson, H., Harford, K., Anjum, A., Zakir, M., Wang, F., Tian, S., Lee, B., Liigand, J., Peters, H., Wang, R.Q.R., Nguyen, T.,

So, D., Sharp, M., Silva, R., Gabriel, C., Scantlebury, J., Jasinski, M., Ackerman, D., Jewison, T., Sajed, T., Gautam, V., Wishart, D.S.: DrugBank 6.0: the DrugBank Knowledgebase for 2024. Nucleic Acids Research **52**(D1), 1265–1275 (2024) https://doi.org/10.1093/nar/gkad976

[72] Varallo-Rodriguez, C., Freyer, C.W., Ontiveros, E.P., Griffiths, E.A., Wang, E.S., Wetzler, M.: Bosutinib for the Treatment of Philadelphia Chromosome-Positive Leukemias. Expert opinion on orphan drugs **3**(5), 599–608 (2015) https://doi.org/10.1517/21678707.2015.1036027 . Accessed 2024-02-04

[73] Huguet, F., Leguay, T., Raffoux, E., Rousselot, P., Vey, N., Pigneux, A., Ifrah, N., Dombret, H.: Clofarabine for the treatment of adult acute lymphoid leukemia: the Group for Research on Adult Acute Lymphoblastic Leukemia intergroup. Leukemia & Lymphoma **56**(4), 847–857 (2015) https://doi.org/10.3109/10428194.2014.887708 . Publisher: Taylor & Francis

[74] Zhao, Y., Zhu, Y., Wang, H., Ji, C.: Case Report: Successful Treatment of Cutaneous Squamous Cell Carcinoma in Three Patients With a Combination of Acitretin and Clarithromycin. Frontiers in Oncology **11** (2021). Accessed 2024-02-04

[75] Lin, X.-Y., He, C.-D., Xiao, T., Jin, X., Chen, J., Wang, Y.-K., Liu, M., Wang, K.-B., Jiang, Y., Wei, H.-C., Chen, H.-D.: Acitretin induces apoptosis through CD95 signalling pathway in human cutaneous squamous cell carcinoma cell line SCL-1. Journal of Cellular and Molecular Medicine **13**(9a), 2888–2898 (2009) https://doi.org/10.1111/j.1582-4934.2008.00397.x . Accessed 2024-02-04

[76] Nguyen, P.H., Giraud, J., Staedel, C., Chambonnier, L., Dubus, P., Chevret, E., Bœuf, H., Gauthereau, X., Rousseau, B., Fevre, M., Soubeyran, I., Belleannée, G., Evrard, S., Collet, D., Mégraud, F., Varon, C.: All-trans retinoic acid targets gastric cancer stem cells and inhibits patient-derived gastric carcinoma tumor growth. Oncogene **35**(43), 5619–5628 (2016) https://doi.org/10.1038/onc.2016.87

[77] Guarrera, L., Kurosaki, M., Garattini, S.-K., Gianni', M., Fasola, G., Rossit, L., Prisciandaro, M., Di Bartolomeo, M., Bolis, M., Rizzo, P., Nastasi, C., Foglia, M., Zanetti, A., Paroni, G., Terao, M., Garattini, E.: Anti-tumor activity of all-trans retinoic acid in gastric-cancer: gene-networks and molecular mechanisms. Journal of Experimental & Clinical Cancer Research **42**(1), 298 (2023) https://doi.org/10.1186/s13046-023-02869-w

[78] Heppner, F.L., Ransohoff, R.M., Becher, B.: Immune attack: the role of inflammation in alzheimer disease. Nature Reviews Neuroscience **16**(6), 358–372 (2015)

[79] Batra, R., Arnold, M., Wörheide, M.A., Allen, M., Wang, X., Blach, C., Levey, A.I., Seyfried, N.T., Ertekin-Taner, N., Bennett, D.A., *et al.*: The landscape

of metabolic brain alterations in alzheimer's disease. Alzheimer's & Dementia **19**(3), 980–998 (2023)

[80] Bellenguez, C., Küçükali, F., Jansen, I.E., Kleineidam, L., Moreno-Grau, S., Amin, N., Naj, A.C., Campos-Martin, R., Grenier-Boley, B., Andrade, V., *et al.*: New insights into the genetic etiology of alzheimer's disease and related dementias. Nature genetics **54**(4), 412–436 (2022)

[81] Jones, L.L., McDonald, D.A., Borum, P.R.: Acylcarnitines: role in brain. Progress in lipid research **49**(1), 61–75 (2010)

[82] Malykh, A.G., Sadaie, M.R.: Piracetam and piracetam-like drugs: from basic science to novel clinical applications to CNS disorders. Drugs **70**(3), 287–312 (2010)

[83] Deardorff, W.J., Feen, E., Grossberg, G.T.: The use of cholinesterase inhibitors across all stages of alzheimer's disease. Drugs & aging **32**, 537–547 (2015)

[84] Van Dyck, C.H., Swanson, C.J., Aisen, P., Bateman, R.J., Chen, C., Gee, M., Kanekiyo, M., Li, D., Reyderman, L., Cohen, S., *et al.*: Lecanemab in early alzheimer's disease. New England Journal of Medicine **388**(1), 9–21 (2023)

[85] Budd Haeberlein, S., Aisen, P., Barkhof, F., Chalkias, S., Chen, T., Cohen, S., Dent, G., Hansson, O., Harrison, K., Von Hehn, C., *et al.*: Two randomized phase 3 studies of aducanumab in early alzheimer's disease. The journal of prevention of Alzheimer's disease **9**(2), 197–210 (2022)

[86] Cummings, J., Zhou, Y., Lee, G., Zhong, K., Fonseca, J., Cheng, F.: Alzheimer's disease drug development pipeline: 2023. Alzheimer's & Dementia: Translational Research & Clinical Interventions **9**(2), 12385 (2023) https://doi.org/10.1002/trc2.12385 . _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/trc2.12385. Accessed 2024-01-16

[87] Kowiański, P., Lietzau, G., Czuba, E., Waśkow, M., Steliga, A., Moryś, J.: Bdnf: a key factor with multipotent impact on brain signaling and synaptic plasticity. Cellular and molecular neurobiology **38**, 579–593 (2018)

[88] Styr, B., Slutsky, I.: Imbalance between firing homeostasis and synaptic plasticity drives early-phase alzheimer's disease. Nature neuroscience **21**(4), 463–473 (2018)

[89] Ng, T.K.S., Ho, C.S.H., Tam, W.W.S., Kua, E.H., Ho, R.C.-M.: Decreased serum brain-derived neurotrophic factor (bdnf) levels in patients with alzheimer's disease (ad): a systematic review and meta-analysis. International journal of molecular sciences **20**(2), 257 (2019)

[90] Wan, Y.-W., Al-Ouran, R., Mangleburg, C.G., Perumal, T.M., Lee, T.V., Allison, K., Swarup, V., Funk, C.C., Gaiteri, C., Allen, M., et al.: Meta-analysis of the alzheimer's disease human brain transcriptome and functional dissection in mouse models. Cell reports **32**(2) (2020)

[91] Buchman, A.S., Yu, L., Boyle, P.A., Schneider, J.A., De Jager, P.L., Bennett, D.A.: Higher brain bdnf gene expression is associated with slower cognitive decline in older adults. Neurology **86**(8), 735–741 (2016)

[92] Wörheide, M.A., Krumsiek, J., Nataf, S., Nho, K., Greenwood, A.K., Wu, T., Huynh, K., Weinisch, P., Römisch-Margl, W., Lehner, N., et al.: An integrated molecular atlas of alzheimer's disease. medRxiv, 2021–09 (2021)

[93] Tafseer, S., Gupta, R., Ahmad, R., Jain, S., Bhatia, M., Gupta, L.K.: Bupropion monotherapy alters neurotrophic and inflammatory markers in patients of major depressive disorder. Pharmacology Biochemistry and Behavior **200**, 173073 (2021)

[94] Jamal, M., Does, W., Elzinga, B.M., Molendijk, M.L., Penninx, B.W.: Association between smoking, nicotine dependence, and bdnf val66met polymorphism with bdnf concentrations in serum. Nicotine & Tobacco Research **17**(3), 323–329 (2015)

[95] Mannick, J.B., Lamming, D.W.: Targeting the biology of aging with mtor inhibitors. Nature Aging, 1–19 (2023)

[96] Kaeberlein, M., Galvan, V.: Rapamycin and alzheimer's disease: time for a clinical trial? Science translational medicine **11**(476), 4289 (2019)

[97] MacKeigan, J.P., Krueger, D.A.: Differentiating the mtor inhibitors everolimus and sirolimus in the treatment of tuberous sclerosis complex. Neuro-oncology **17**(12), 1550–1559 (2015)

[98] Kaelin Jr, W.G.: The concept of synthetic lethality in the context of anticancer therapy. Nature reviews cancer **5**(9), 689–698 (2005). Publisher: Nature Publishing Group UK London

[99] Wang, J., Wu, M., Huang, X., Wang, L., Zhang, S., Liu, H., Zheng, J.: SynLethDB 2.0: a web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery. Database: The Journal of Biological Databases and Curation **2022**, 030 (2022) https://doi.org/10.1093/database/baac030 . Accessed 2024-05-16

[100] Luo, Y., Hitz, B.C., Gabdank, I., Hilton, J.A., Kagda, M.S., Lam, B., Myers, Z., Sud, P., Jou, J., Lin, K., Baymuradov, U.K., Graham, K., Litton, C., Miyasato, S.R., Strattan, J.S., Jolanki, O., Lee, J.-W., Tanaka, F.Y., Adenekan, P., O'Neill, E., Cherry, J.M.: New developments on the Encyclopedia of DNA Elements

(ENCODE) data portal **48**(D1), 882–889 https://doi.org/10.1093/nar/gkz106231713622

[101] Noguchi, S., Arakawa, T., Fukuda, S., Furuno, M., Hasegawa, A., Hori, F., Ishikawa-Kato, S., Kaida, K., Kaiho, A., Kanamori-Katayama, M., Kawashima, T., Kojima, M., Kubosaki, A., Manabe, R.-i., Murata, M., Nagao-Sato, S., Nakazato, K., Ninomiya, N., Nishiyori-Sueki, H., Noma, S., Saijyo, E., Saka, A., Sakai, M., Simon, C., Suzuki, N., Tagami, M., Watanabe, S., Yoshida, S., Arner, P., Axton, R.A., Babina, M., Baillie, J.K., Barnett, T.C., Beckhouse, A.G., Blumenthal, A., Bodega, B., Bonetti, A., Briggs, J., Brombacher, F., Carlisle, A.J., Clevers, H.C., Davis, C.A., Detmar, M., Dohi, T., Edge, A.S.B., Edinger, M., Ehrlund, A., Ekwall, K., Endoh, M., Enomoto, H., Eslami, A., Fagiolini, M., Fairbairn, L., Farach-Carson, M.C., Faulkner, G.J., Ferrai, C., Fisher, M.E., Forrester, L.M., Fujita, R., Furusawa, J.-i., Geijtenbeek, T.B., Gingeras, T., Goldowitz, D., Guhl, S., Guler, R., Gustincich, S., Ha, T.J., Hamaguchi, M., Hara, M., Hasegawa, Y., Herlyn, M., Heutink, P., Hitchens, K.J., Hume, D.A., Ikawa, T., Ishizu, Y., Kai, C., Kawamoto, H., Kawamura, Y.I., Kempfle, J.S., Kenna, T.J., Kere, J., Khachigian, L.M., Kitamura, T., Klein, S., Klinken, S.P., Knox, A.J., Kojima, S., Koseki, H., Koyasu, S., Lee, W., Lennartsson, A., Mackay-sim, A., Mejhert, N., Mizuno, Y., Morikawa, H., Morimoto, M., Moro, K., Morris, K.J., Motohashi, H., Mummery, C.L., Nakachi, Y., Nakahara, F., Nakamura, T., Nakamura, Y., Nozaki, T., Ogishima, S., Ohkura, N., Ohno, H., Ohshima, M., Okada-Hatakeyama, M., Okazaki, Y., Orlando, V., Ovchinnikov, D.A., Passier, R., Patrikakis, M., Pombo, A., Pradhan-Bhatt, S., Qin, X.-Y., Rehli, M., Rizzu, P., Roy, S., Sajantila, A., Sakaguchi, S., Sato, H., Satoh, H., Savvi, S., Saxena, A., Schmidl, C., Schneider, C., Schulze-Tanzil, G.G., Schwegmann, A., Sheng, G., Shin, J.W., Sugiyama, D., Sugiyama, T., Summers, K.M., Takahashi, N., Takai, J., Tanaka, H., Tatsukawa, H., Tomoiu, A., Toyoda, H., family=Wetering, p.d.u. given=Marc, family=Berg, p.d.u. given=Linda M., Verardo, R., Vijayan, D., Wells, C.A., Winteringham, L.N., Wolvetang, E., Yamaguchi, Y., Yamamoto, M., Yanagi-Mizuochi, C., Yoneda, M., Yonekura, Y., Zhang, P.G., Zucchelli, S., Abugessaisa, I., Arner, E., Harshbarger, J., Kondo, A., Lassmann, T., Lizio, M., Sahin, S., Sengstag, T., Severin, J., Shimoji, H., Suzuki, M., Suzuki, H., Kawai, J., Kondo, N., Itoh, M., Daub, C.O., Kasukawa, T., Kawaji, H., Carninci, P., Forrest, A.R.R., Hayashizaki, Y.: FANTOM5 CAGE profiles of human and mouse samples **4**(1), 170112 https://doi.org/10.1038/sdata.2017.112 . Accessed 2024-06-15

[102] The RNAcentral Consortium: RNAcentral: A hub of information for non-coding RNA sequences **47**(D1), 221–229 https://doi.org/10.1093/nar/gky1034 . Accessed 2024-06-15

[103] Zhao, L., Wang, J., Li, Y., Song, T., Wu, Y., Fang, S., Bu, D., Li, H., Sun, L., Pei, D., Zheng, Y., Huang, J., Xu, M., Chen, R., Zhao, Y., He, S.: NONCODEV6: An updated database dedicated to long non-coding RNA annotation in both animals and plants **49**(D1), 165–171 https://doi.org/10.1093/nar/gkaa1046 . Accessed

2024-06-15

[104] Zhang, Y., He, Q., Hu, Z., Feng, Y., Fan, L., Tang, Z., Yuan, J., Shan, W., Li, C., Hu, X., Tanyi, J.L., Fan, Y., Huang, Q., Montone, K., Dang, C.V., Zhang, L.: Long noncoding rna linp1 regulates repair of dna double-strand breaks in triple-negative breast cancer. Nature Structural & Molecular Biology **23**(6), 522–530 (2016) https://doi.org/10.1038/nsmb.3211

[105] Chen, L., Ma, D., Li, Y., Li, X., Zhao, L., Zhang, J., Song, Y.: Effect of long non-coding RNA PVT1 on cell proliferation and migration in melanoma **41**(3), 1275–1282 https://doi.org/10.3892/ijmm.2017.3335 29286144

[106] Liu, S.J., Horlbeck, M.A., Cho, S.W., Birk, H.S., Malatesta, M., He, D., Attenello, F.J., Villalta, J.E., Cho, M.Y., Chen, Y., Mandegar, M.A., Olvera, M.P., Gilbert, L.A., Conklin, B.R., Chang, H.Y., Weissman, J.S., Lim, D.A.: CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells **355**(6320), 7111 https://doi.org/10.1126/science.aah7111 27980086

[107] Ntini, E., Budach, S., Vang Ørom, U.A., Marsico, A.: Genome-wide measurement of RNA dissociation from chromatin classifies transcripts by their dynamics and reveals rapid dissociation of enhancer lncRNAs **14**(10), 906–9226 https://doi.org/10.1016/j.cels.2023.09.005 37857083

[108] Yang, H., Lin, Z., Zhang, M.: Rethinking Knowledge Graph Evaluation Under the Open-World Assumption (2022)

[109] Lobentanzer, S., Aloy, P., Baumbach, J., Bohar, B., Carey, V.J., Charoentong, P., Danhauser, K., Doğan, T., Dreo, J., Dunham, I., Farr, E., Fernandez-Torras, A., Gyori, B.M., Hartung, M., Hoyt, C.T., Klein, C., Korcsmaros, T., Maier, A., Mann, M., Ochoa, D., Pareja-Lorente, E., Popp, F., Preusse, M., Probul, N., Schwikowski, B., Sen, B., Strauss, M.T., Turei, D., Ulusoy, E., Waltemath, D., Wodke, J.A.H., Saez-Rodriguez, J.: Democratizing knowledge representation with biocypher. Nature Biotechnology **41**, 1056–1059 (2023) https://doi.org/10.1038/s41587-023-01848-y

[110] Chen, Z., Wang, Y., Zhao, B., Cheng, J., Zhao, X., Duan, Z.: Knowledge Graph Completion: A Review. IEEE Access **8**, 192435–192456 (2020)

[111] Bang-Jensen, J., Gutin, G.: Section 2.3.4: The Bellman-Ford-Moore algorithm. Digraphs: Theory, Algorithms and Applications. Springer, ??? (2000)

[112] Baras, J.S., Theodorakopoulos, G.: Path problems in networks. Synthesis Lectures on Communication Networks **3**(1), 1–77 (2010)

[113] Cerami, E.G., Gross, B.E., Demir, E., Rodchenkov, I., Babur, Ö., Anwar, N., Schultz, N., Bader, G.D., Sander, C.: Pathway commons, a web resource for

biological pathway data. Nucleic acids research **39**(suppl_1), 685–690 (2010)

[114] Goldberg, R., Cary, M., Demir, E.: BioPAX a community standard for pathway data sharing. Nature Biotechnology (28) (2010)

[115] Corso, G., Cavalleri, L., Beaini, D., Liò, P., Veličković, P.: Principal neighbourhood aggregation for graph nets. arXiv preprint arXiv:2004.05718 (2020) https://arxiv.org/pdf/2004.05718.pdf

[116] Zhao, H., Yin, X., Xu, H., Liu, K., Liu, W., Wang, L., Zhang, C., Bo, L., Lan, X., Lin, S., Feng, K., Ning, S., Zhang, Y., Wang, L.: LncTarD 2.0: An updated comprehensive database for experimentally-supported functional lncRNA-target regulations in human diseases **51**(D1), 199–207 https://doi.org/10.1093/nar/gkac984 36321659

[117] Pykeen/Pykeen. PyKEEN. https://github.com/pykeen/pykeen Accessed 2024-04-01

[118] Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Teh, Y.W., Titterington, M. (eds.) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 9, pp. 249–256. PMLR, Chia Laguna Resort, Sardinia, Italy (2010). https://proceedings.mlr.press/v9/glorot10a.html

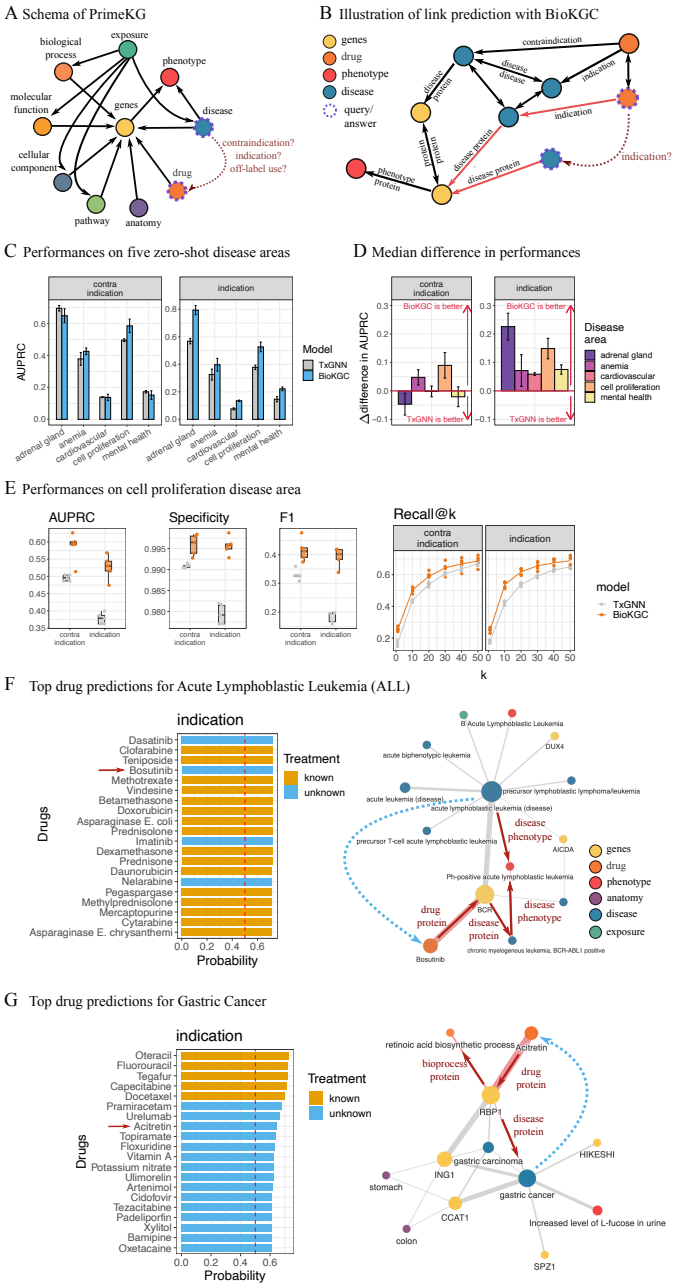


Fig. 3: Comparison of BioKGC and the state-of-the-art TxGNN model for predicting drug-disease relations on PrimeKG: A) Schema of PrimeKG, a multimodal knowledge graph with 10 biological node types (e.g., protein, disease, drug) and over 5 million relations, used for predicting drug-disease interactions. B) Illustration of BioKGC making use of paths between drug and disease for indication prediction. C) Comparison of AUPRC performance in zero-shot prediction across five disease area splits (adrenal gland, anemia, cardiovascular, cell proliferation, mental health). AUPRC is calculated by comparing all ground truth positive and negative drugs for each disease, then averaging metrics across all diseases within each area D) Mean differences in AUPRC for each disease area between BioKGC and TxGNN, with a positive delta reflecting higher AUPRC for BioKGC. E) Performance metrics specifically for the disease area split of cell proliferation. Recall@k reflects the proportion of ground truth edges retrieved in the top k predictions. Top predictions are reported for contraindication and indication for the disease F) Acute Lymphoblastic Leukemia and G) Gastric Cancer within the disease area cell proliferation. Visualization of gradient importance on paths predicting (H) Bosutinib for Acute Lymphoblastic Leukemia and (I) Acitretin for Gastric Cancer.

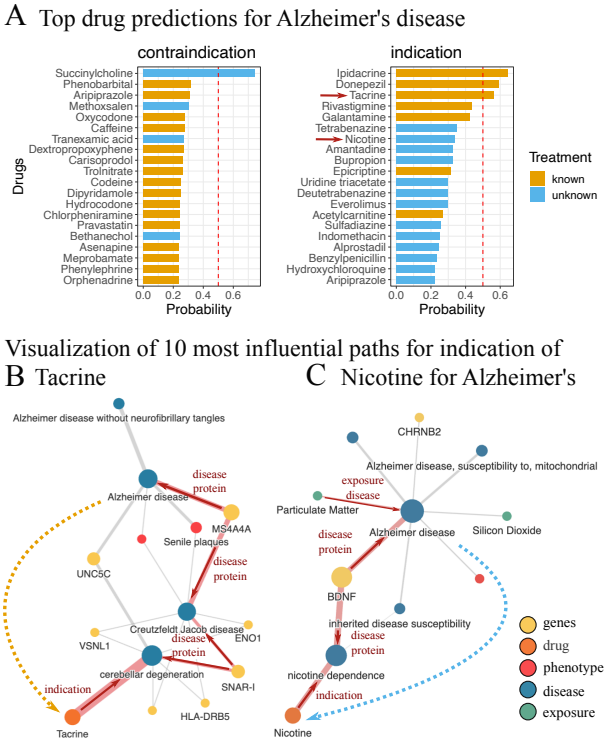


Fig. 4: **Predictions of BioKGC on custom data split of Alzheimer's disease:** Top predictions of contraindication and indication for custom Alzheimer's disease. Data split obtained over TxGNN's disease evaluation code. Known treatments are included in the ground truth of PrimeKG. Visualization of gradients on paths important for the prediction B) known Tacrine and C) unknown Nicotine for Alzheimer's. Top 10 paths are displayed, with edge width reflecting edge weight, path of the highest weight highlighted in red.

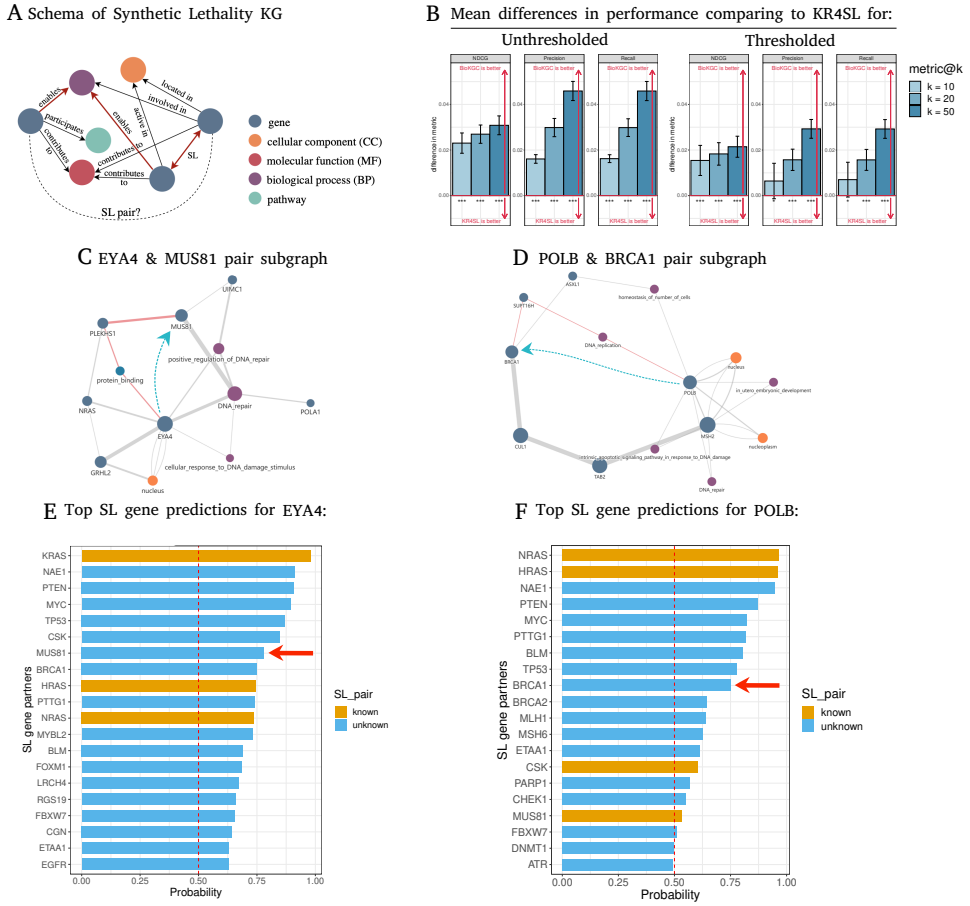


Fig. 5: **Comparison of BioKGC with state-of-the-art SL gene pair prediction algorithm KR4SL:** A) Illustration of SynLeth KG for the prediction of SL gene pairs, consisting of genes and their SL interactions, cellular components (CC), molecular functions (MF), biological pathway (BP) and pathway. B) Mean difference in performances between BioKGC and KR4SL given as NDCG, Precision and Recall for both unthresholded and thresholded data. C) Visualization of gradients on paths important for the prediction of the EYA4 - MUS81 pair. D) Visualization of gradients on paths important for the prediction of the POLB - BRCA1 pair. E) Top predicted SL gene partners for EYA4. F) Top predicted SL gene partners for POLB.

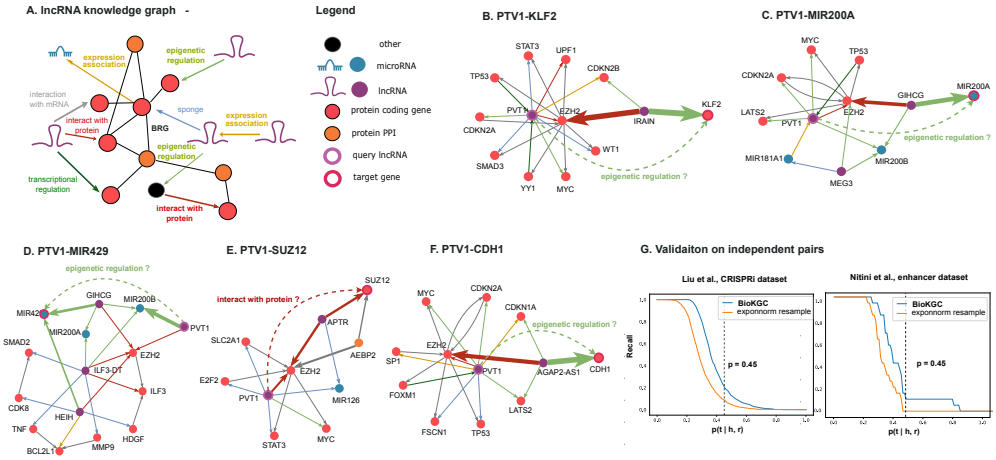


Fig. 6: **Prediction of novel lncRNA-target regulatory interactions A.** Depiction of the lncRNA-mediated regulation knowledge graph (KG) constructed from the LncTarD 2.0 KG, augmented by incorporating a protein-protein interaction (PPI) network as a biological regulatory graph (BRG) for message passing. The graph features six types of node entities: lncRNAs, microRNAs, mRNAs, pseudogenes, transcription factors, proteins, and protein_ppi, the latter representing genes from the external PPI network not originally included in LncTarD 2.0. Various types of regulatory relationships are indicated by directed edges of different colors, while protein-protein interactions from the BRG are shown with black undirected edges. **B-F** Examples of predicted novel interactions for the cancer lncRNA PTV1 using BioKGC. The top 10 most crucial paths for prediction, ranked by gradient, are shown for all five examples. The edge width represents the frequency of appearance in paths, therefore the importance of that connection for the prediction. Edge colors indicate the different regulatory mechanisms, following the color code of Figure 6A. **G-H** Independent evaluation of predictions based on external datasets from CRISPR and enhancer-based experiments.